VRIJE UNIVERSITEIT

RESEARCH MASTER THESIS

Topic modelling *De Gids*: An explorative study into the use of topic modelling on a cultural periodical

Author:
Leon VAN WISSEN

Supervisors:
Dr. Jacqueline BEL
Dr. Roser MORANTE

Third reader:
Dr. Serge TER BRAAKE

A thesis submitted in fulfilment of the requirements
for the degree of Master of Arts

in the

Faculty of Humanities
Department Language, Literature & Communication
Modern Dutch Literature

February, 2019

VU VRIJE UNIVERSITEIT AMSTERDAM

Declaration of Authorship

I, Leon VAN WISSEN, declare that this thesis titled, "Topic modelling *De Gids*: An explorative study into the use of topic modelling on a cultural periodical" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date

VRIJE UNIVERSITEIT

Abstract

Faculty of Humanities

Department Language, Literature & Communication

Master of Arts

Topic modelling *De Gids*: An explorative study into the use of topic modelling on a cultural periodical

by Leon VAN WISSEN

This thesis investigates how the technique of Topic Modelling can give a Digital Humanities scholar an entry point to a body of literary and historical texts. In three exemplary case studies, it shows how the technique can complement a traditional close reading approach, and how a corpus can be analysed diachronically. For this, it applies the technique to the Dutch general cultural magazine *De Gids* (1837-), which in its early days has proven to be highly influential in the field of literature, culture and science, and nowadays still functions as an archive for the evaluation of culture and science in the recent history of The Netherlands.

In the first case study, the technique is applied to a single volume (72 articles) of *De Gids* to find information on the Dutch constitutional reform of 1848. The corpus size is scaled up in the second case study (1,642 articles), in which writings on the Dreyfus affair (1894-1906) are found and analysed with help of the topic models outcome. The third and last case study analyses all articles from the first one hundred volumes (10,624 articles) of *De Gids* on a macro scale to investigate the propagation of disciplinary trends in the magazine by identifying and interpreting them, and by connecting changes in proportion to the magazine's internal and external factors.

This study shows that a topic model is able to show *De Gids*'s involvement in politics and liberalism in 1848 and that it is able to highlight other themes and text types in articles from the period of the Dreyfus affair. It thereby shows the applicability and validity of a topic models outcome for literary historical research. The real benefit of the technique exists in applying it to largescale corpora as is shown in the third case study in which fluctuations in the distribution of (academic) disciplines and fields in the magazine over time are visualised.

Acknowledgements

This research master thesis concludes the master programme I followed for the past years at the *Vrije Universiteit Amsterdam*, at which I have strived to combine the two fields of literature and (computational) linguistics. Although not literary in the strict sense of the term, does this thesis touch upon literary themes and similar methods that are faced in (computational) historical research.

I am very grateful for the freedom I got during my master's programme to combine courses that were sometimes coming from the opposite ends of the fields. I learned a lot from the courses I followed in the Linguistic Engineering track of the Linguistics master and I really enjoyed putting them into practise during my internship at the *Huygens Instituut* in October 2016 - January 2017, supervised by Peter Boot. These courses and this experience definitely lead to where I am now, working at CREATE on the Golden Agents project at the *Universiteit van Amsterdam*.[1]

There is still a lot of work to do to bring the field of the Digital Humanities forwards and to bridge the gap between the computational and the non-computational. Every step is one in the right direction, and this thesis, which I would rather have written in Dutch, hopefully contributes to this area.

I would like to thank my university teachers, who have become colleagues, for working with them and for helping me get the most out of my studies and my degree. Foremost, I thank my supervisors Jacqueline Bel and Roser Morante for their suggestions, critique and their patience on the forthcoming of this thesis. I sincerely hope that we can join forces someday, either at the *Vrije Universiteit* or the *Universiteit van Amsterdam*. Finally, I am especially grateful to my mentor, former supervisor and colleague Ben Peperkamp (†2017). Without his inspiring ideas, his motivation and his confidence, this thesis nor my master's degree would have been the same.

Amsterdam, November 2018

[1] But my internship at the *Meertens Instituut* supervised by René van Stipriaan in summer 2014 during my bachelor Dutch Language and Culture already instigated this.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For scholars dealing with textual corpora, more and more data become available each day through digitisation programmes of libraries,[1] increased (online) publication, and better scraping techniques to capture online content such as tweets, blog posts or other publications. This means that the amount of stored and accessible data that can be used as an object of research has rapidly increased, which, in turn, has led to a radical change in the way scholars from the humanities and social sciences can do research. Not only does it get easier to obtain (primary) literature, but the (massive) collections of (textual) data allow for a different style of humanities research.

distant reading

This different "style" can be described by the term "distant reading". The term was introduced by literary scholar Franco Moretti (2000) (and more extensively outlined in Moretti (2013)) to (now famously) characterise the practise of moving research away from the primary literature, towards secondary sources or meta information, to analysing traditionally (by close reading) incomprehensible corpora sizes with computational methods.[2] Instead of focussing on a single text or a few texts, the literary scholar is - with the help of new digital methods - able to zoom out and work on a different scale, analysing not a single book or text, but thousands at once.

Much has been said about this "distant reading," among others by Matthew Jockers (who calls this "Macroanalysis") in his *Macroanalysis: Digital methods and literary history* (Jockers, 2013), and more recently by Fleming (2017), who signals a move back from "computational reading" to "exemplary reading," or a combination of both. Fleming (ibid., pp. 441-442) criticises Moretti (2000) and points out how he fails to interpret the results of his computational models in the light of the texts they were trained on. Instead of using close reading to validate and critically evaluate the models, close reading is merely used to provide a context for the results. Computational

[1] The Koninklijke Bibliotheek [Dutch Royal Library] (https://www.kb.nl/organisatie/onderzoek-expertise/digitaliseringsprojecten-in-de-kb) and the British Library (http://www.bl.uk/aboutus/stratpolprog/collectioncare/digitalpreservation/strategy/dpstrategy.html) clearly state their digitisation strategies on their websites.

[2] Also on: https://litlab.stanford.edu/.

methods show real potential when it comes to analysing larger corpora in research, but these have to be verified themselves, e.g. by means of close reading:

> This dimension of computational humanities or cultural analytics [...] produces new examples, exemplary passages that one needs to engage and interpret, that then have implications both for the analysis of other exemplary passages and the (quantitative) model that produced them in the first place. (Fleming (2017, p. 452) and Long and So (2016))

In other words, the close and distant reading approaches rule each other out; it is not one approach over the other, and instead, distant reading can and should always be complemented by close reading. The method of distant reading can serve as a method to get to know a corpus or to find interesting phenomena in the data that are worth investigating on a closer level, and which otherwise would have stayed unnoticed or would have been very difficult to spot. Data presented by any (computational) method should always be interpreted in the light of the text's contents, and this requires close reading or detailed analysis of the source.

Topic Modelling

Several computational tools and techniques have been developed in recent years to "make sense" of a bulk of data and facilitate and assist the Digital Humanities' researcher in analysing texts on this macro level. One of these tools is the technique called Topic Modelling, a method to analyse a large corpus of texts and to discover the "themes" that are embedded in this corpus. Nowadays, topic modelling is seen as one of the many techniques that can be used in text mining, but, in my view, its possibilities for literary and historical research are underestimated. This is, for instance, shown in the lack of clear and straightforward applications in the Humanities and the field of the Literary Studies in particular, despite a spark of interest in 2012, when the technique was centred in the *Journal for Digital Humanities* that devoted an entire issue (2012, Vol. 2, no. 1) to the use of topic modelling in the Humanities.

In spite of the attention it received in 2012 and several small studies afterwards (of which the most important ones are mentioned in Chapter 2), topic modelling in Humanities research has not (yet) established a foothold. This could be due to the not always unambiguous results topic modelling algorithms might present to their user, to their difficulty to set up, or their incomprehensibility (because of their strong foundation in algebra). With an increasing number of researchers participating in the Digital Humanities, doing "interdisciplinary" work by using techniques from the exact sciences (e.g. mathematics and computational sciences), I think the technique should be given a second chance, and I believe that the technique can have a valuable contribution to Literary research in such a way that both traditional researchers and scholars from the digital humanities can benefit from its outcomes.

Research (see Chapter 2 for examples) has shown that this technique can be applied to large corpora of newspapers, scientific journals and even literature. However, these studies do have some downsides. Firstly, almost all of these examples centre on the applicability or validity of the technique itself and are being used in a manner that is less appealing for researchers from History or Literary Studies.[3] Secondly, concrete case studies from the field of History and Literary Studies are scarce, especially research that is done on Dutch corpora, despite the Dutch digital data collections that are available nowadays.[4] Thirdly, it also appears that this technique is becoming less popular, despite its possibilities for textual analysis. To fill this void in examples for Dutch literary studies (and Digital Humanities in general), this thesis, therefore, provides several concrete examples of the appliance of the technique of topic modelling on a Dutch (historical) corpus.

De Gids

There can be no better corpus to put this technique of topic modelling to the test for Dutch literature than the corpus of the eminent general cultural magazine *De Gids* [The Guide] that was published from 1837 onwards and is completely available digitally (be it partly licensed). In its early days, this periodical has proven to be highly influential in the field of literature, culture and science. Nowadays, this corpus still functions as an archive for the evaluation of culture and science in the recent history of The Netherlands for researchers in the Humanities. It contains writing on the formation of the Dutch politics and constitution, it lists opinions on events from the European stage, and it shows that its contents and authors are up with the times with regard to new scientific developments in Germany, France, and The Netherlands. The periodical contains various genres of text, such as informative news items, opinion pieces, and literary and poetical pieces. What is more, these contributions are written by authors coming from a broad range of disciplines.[5]

So far, there has been little research on *De Gids* as a whole (Aerts et al., 1987; Aerts, 1997; Aerts, 2012), and since every volume of this periodical is available digitally, it comes as a logical and obvious choice to analyse this large, coherent and high-quality corpus with a technique that gives insight in the contents of the individual texts and the corpus itself. This technique makes it possible to discover new findings that would remain hidden with a more traditional approach to this source, and further makes it possible to validate and complement claims that have already been made about the contents of this magazine. The history of *De Gids*, its contributors, and the fact that it has been a constant (f)actor in the nineteenth- and twentieth century makes this medium an excellent object of research that can be examined with the use of a technique such as topic modelling.

[3] The technique is for instance used as substitution of human judgement and interpretation (e.g. as a method to automatically process and classify texts), instead of complementation.

[4] E.g. KB Krantencorpus and other sources in Delpher, DBNL, Nederlab, Parlementaire data, individual catalogues of university libraries, to name a few.

[5] The general cultural magazine could count on contributions from persons such as Nobel prize winners H.L. Lorentz (1853-1928) and H. van der Waals (1837-1923) [Science], influential politicians such as Johan Rudolph Thorbecke (1798-1872) and P.A.S. van Limburg Brouwer (1829-1873) [Politics], and renowned Dutch (literary) authors as Isaac Da Costa (1798-1860) and Jacob van Lennep (1802-1868) [Literature].

The goal of this thesis is thereby twofold: (i) it sheds new light on the corpus of *De Gids* which contributes to the total amount of literature and research that is available on this conspicuous periodical, and (ii) it serves as an example of a topic modelling approach on a Dutch (historical) corpus. To demonstrate the applicability of topic modelling as a technique to get a grip on a text, to use it as a powerful search tool, and to show its pitfalls for Humanities oriented research and specifically for research on *De Gids*, this thesis gives several exemplary case studies(see the sub-questions below), each increasing in scope and size, that provide a way to easily extract “themes from parts of this corpus. To set a boundary, these case studies will draw from the first 100 volumes/years of *De Gids* (The full corpus description can be read in Chapter 3). Each of the case studies eventually contributes to answering the main question of this thesis, which is as follows:

- **To what extent can topic modelling be considered a helpful tool in Dutch literary cultural research, seen from the example of analysing the cultural periodical *De Gids* (1837-)?**

This main research question is answered by looking at three sub-questions, all corresponding to a case study (experiment) in this thesis:

1. *How can topic modelling be used to validate claims about the contents of a corpus? (i.e. detecting the topics/themes that are known to exist in the corpus). Here applied in analysing the 1848 volume of* De Gids*, the year the Dutch constitution was reformed.*

2. *How can topic modelling be used as a search or indexing tool in a large corpus? (i.e. highlighting topics/themes and authorial contributions). Here applied in analysing the volumes from the years of the Dreyfus affair (1894-1906).*

3. *To what extent can topic modelling be used to discover new themes or shifts in topical content in a diachronic corpus? (i.e. highlighting differences in topical distribution within a specified time-frame). Here the first 100 years of* De Gids *are analysed from the perspective of cultural history and the history of academic disciplines.*

This research is new in the field of Dutch literary studies and adds an alternative example to the collection of existing research (mentioned in Chapter 2) with regard to the technique used. It will show the applicability of topic modelling to Dutch Digital Humanities research by centring on three concrete examples from *De Gids*. This thesis uses the work done by i.a. Remieg Aerts (1997) as a foundation, but complements this by providing new perspectives on three themes existing in *De Gids*. Not only do I hope this will result in valuable new insights in this fascinating periodical, its themes, and contributors, but I also hope this will contribute to the existing literature on the use and understanding of text mining techniques such as topic modelling in the Digital Humanities and literary studies. This research positions itself between Literary and Cultural Historical research, and the studies of Computational Literature and Linguistics. Therefore, I try to explain terms and techniques where needed, to make this thesis readable for scholars who are not familiar with the discussed algorithms or techniques.

The thesis is structured as follows: Chapter 2 gives an overview of existing and related topic modelling studies from the Humanities and gives an introduction into the corpus of *De Gids* that is used. After this, a historical background for each of the three case studies is given. Chapter 3 outlines the methods used for the three experiments, the programs and computational settings that are used and describes (the quality of) the corpus. Results from these experiments can be found in Chapter 4. Then, their quality and applicability are evaluated in Chapter 5. The conclusions that are drawn upon these evaluations are given in the final Chapter 6.

Chapter 2

Background and related work

To fully understand how the topic modelling algorithm that is used in this thesis works, it is necessary to get insight into its origin, its mathematical operations, and its parameters. Therefore, I have written a brief history of the technique below (2.2), together with a more detailed explanation on the topic modelling algorithm and the specific toolsets that are used in this thesis (2.3.1). Its explanation is more technical than the descriptions one usually finds in publications from the Digital Humanities or Literary Studies but is in line with what is customary in the field of Computational Linguistics. I have therefore chosen to also include the formal notation of the algorithm, hoping that it is understandable with help from the accompanying text.

The techniques that lead to the LDA algorithm, the most common implementation of a topic model nowadays, are described, after which LDA is explained in detail, together with its recommendations and fallacies. This part is followed by a section of Related Work (2.5) that features studies in which not necessarily LDA, but any topic model is used in combination with a corpus of texts. The approach or method and the corpus characteristics of some of those studies can be compared to and resemble what is implemented and used in this thesis. This chapter also gives an outline on the history and the varied contents of the general cultural magazine *De Gids* (2.6). Finally, this is followed by a historical background that is relevant for each of the three case studies.

2.1 Digital Humanities

Digital Humanities

The newly emerging (sub)discipline of the "Digital Humanities" gives shape to the increased digitisation of the humanities and serves as an umbrella term to encompass various strategies to tackle digital data, stretching from the field of art history to the field of computational linguistics. All the subfields within the Digital Humanities have in common that they do "something on the computer," but some of them are more adept at using computational techniques to produce analyses and results. Using computational tools which often have a strong foundation in mathematics is not always applicable to, or necessary for, research in the Humanities, but it is shown that in some cases these tools can be quite fruitful (see the related work section 2.5 or below for examples).

2.1.1 Digital text analysis

It is well known that the larger the object of research is, the less comprehensible it gets. There is the risk of being swamped by the material and it is not clear where one should start to get to know the research object. Fortunately for the 21st-century researcher, making use of computational tools can help to get a grip on a collected corpus, either to find research opportunities or to get insight into the nature of the accumulated texts. Computational tools can also be used to (more easily) validate prior claims about certain data. A good and innovative example from the field of literary studies is the advent of computational techniques in the field of stylometry to validate or falsify claims about authorship attribution by discovering a change in authorial style in prose (Kestemont, 2012). Other studies assess the "literariness" of (digital) prose by comparing texts according to their text-internal features (Van Cranenburgh and Bod, 2017) or focus on categorizing (non-fictional) texts according to their themes (Riddell, 2014).

The computational complexity of these research examples in the humanities differs to a large degree. Some are highly computational and rely heavily on statistics and maths, others try to assist the more "traditional" literary research and are in essence nothing more than handy search tools, which prove to be a tremendous help for literary scholars. Just searching through digital texts is one of the simplest ways possible of a computer assisting research. For instance, for a corpus of literary texts, searching for the term "fainting" would sum up all the locations where this term is used and maybe give a total count of how many times this word occurs in the text or corpus. Some other search utilities might even present this term in its context by showing an excerpt from the corpus of five words before and five words after the search term to benefit interpreting this theme.[1]

This search already becomes more sophisticated when one lets the computer incorporate word variants of the search term: inflexions or synonyms of the initial token. With this, it becomes possible to search on a more abstract level. Besides "fainting," words such as "blackout," "pass out," and "collapse" are used as input in the program and return in the search result, which ideally generates more hits.[2] To further abstract this search, one ideally asks the computer for all instances (in the case of literature) where "a person loses consciousness," or, even more generally, all instances in which there is a "(negative) medical condition of one of the characters" described.

[1] A well-known example in literary studies is the concordance program AntConc (http://www.laurenceanthony.net/software.html), but other, similar software exists as well (e.g. ATLAS.ti (http://atlasti.com) or Voyant (https://voyant-tools.org.)

[2] One may assume a (literary) text uses synonyms to describe accounts of the same events

topic

In this situation, one asks the computer to return all cases in which this ”topic” is exemplified.

latent

ambiguity

In the example above, one expects a machine to go beyond the surface representation of words and terms and know the meaning of a text. One wants to find instances in the text with the subject “medical condition,” such as the “fainting” example above, but this is never explicitly mentioned as “medical condition” in the corpus. This topic is latent and is therefore inferred from these kinds of descriptions, which are, of course, built up from a collection of resembling words. Humans go through this process of inference automatically when reading a text, but for a computer, this can be very difficult. It has to overcome issues such as word ambiguity (e.g. disambiguating between “lead” (verb) or “lead” (material)), cross-reference (with use of pronouns), or the use of figurative speech. At the same time, it has to be aware of a languages syntax, morphology, discourse, and much more. It has to process text in a way a human does, which is not (yet) possible.

2.1.2 Natural language processing

Natural Language Processing

annotating

Developments within the field of NLP (Natural Language Processing) and text mining have led to machine translation systems (e.g. Google Translate), summarisation systems and applications that can detect concept shift . This is just a small selection of the applications that are built to process and understand texts on a computational level. These techniques are mostly relying on prior (user) input in the form of training data, which makes developing such a system very time-consuming. One needs “gold data,” often created through human labelling or annotating. This can, for example, tell the computer what excerpts and word sequences in a text are important in the case of summarising. In cases in which human labelling is required, an interpretation step is added to the process. This makes a computational approach seem less independent and objective than it really is since the result still depends on human effort. To partly overcome this, one often composes annotation guidelines and uses measurements such as the inter-annotator agreement to harmonise the result of multiple annotators. This consistent human-built data is then agreed to be set as ground truth. The computer learns from this (human) input and tries to approximate the decisions of a human annotator as closely as possible when it encounters unseen data. When given enough training data and given the right instructions telling which features should be used in this decision-making, these models can be very robust.

unsupervised

Other techniques that are used in clustering tasks (grouping texts according to their similarity) rely on an unsupervised approach in which no prior model outcome is supplied. In these tasks, the computer tries to structure its input data based on text-internal characteristics or other (meta) data. Imagine the process of sorting a bowl of skittles by grouping by colour, or illustrated with a metaphor more within the scope of this thesis: sorting newspaper articles based on their length, typeface, and other features, such as the words that are being mentioned.[3] In any case, both supervised and unsupervised systems heavily rely on the use of statistics to process data and make predictions.

2.1.3 Topic modelling

topic modelling

One of the text-mining techniques that can give structure to and insight in large amounts of textual data is the technique of topic modelling that has a computer automatically find themes or "topics" in a piece of text. Topic modelling is a clear example of a technique that works in an unsupervised way,[4] since, in principle, no prior information is required in order to cluster or structure input data. To use a quote from David Blei, one of the creators of the now considered de-facto standard for topic modelling:[5]

> Topic modeling provides a suite of algorithms to discover hidden thematic structure in large collections of texts. The results of topic modeling algorithms can be used to summarize, visualize, explore, and theorize about a corpus. (Blei, 2012b, p. 7)

The term topic modelling thus covers several techniques that are able to perform a clusterisation of data. The particular technique I will be using in this thesis is based on the adaptation of Blei, Ng, and Jordan (2003), who introduced an improved method of existing topic modelling approaches in 2002. The history and maths behind this clustering technique are explained in the next Section (2.2).

This method, called LDA, works by generating a specified number of topics for a collection of texts (a corpus). These topics are not expressed by a (clear-cut) label (which is latent and manually assigned, just as in the example given above), but are represented by clusters of words, each varying in representativeness for the topic. In reference to the example in the previous section, words such as "doctor" or "hospital" are more representative for the topic "medical conditions" than the words "drugs" or "alcohol," which might also appear in the topic "narcotics" or "festivities." The user should, after the computer has generated its clusters, assign meaning to these groups of words (interpret them) and assign them a label. After this is done, there should be a clear outline of the corpus contents, its themes and proportions, in the case of a successful outcome of the topic modelling analysis. This is dependent on corpus-internal factors as well as on experiment related settings.

[3] I am not mentioning sorting based on subject/theme here, because that is one step too far ahead. This will become clear in the following paragraphs.

[4] In essence, this technique works unsupervised, but variations that use prior (user) input data exist.

[5] This follows both from the amount of literature on topic modelling that is using the LDA-technique by Blei, Ng, and Jordan (2003) or adaptations on this, as from the genesis of this technique (see Chapter 2). At least within the (Digital) Humanities LDA is the most used (Meeks and Weingart, 2012, p. 3). A (not any more updated) list of studies using topic modelling is listed in the *Topic modeling bibliography* by David D. Mimno (s.d.).

Topic modelling provides a method for textual research that combines an automated computational and highly statistical approach with a human element of interpretation. This combination proves to be quite applicable and fruitful in the Digital Humanities and opens up new possibilities for a new type of questions, again according to Blei:

Digital Humanities

> With probabilistic modeling for the humanities, the scholar can build a statistical lens that encodes her specific knowledge, theories, and assumptions about texts. She can then use that lens to examine and explore large archives of real sources. (Blei, 2012b, p. 8)
>
> The humanities, fields where questions about texts are paramount, is an ideal testbed for topic modeling and fertile ground for interdisciplinary collaborations with computer scientists and statisticians. (ibid., p. 9)

2.2 History of topic modelling

Topic modelling has its origin in applying a dimensionality reduction to a term-document matrix of simple word counts in documents from a corpus. This representation stems from an even simpler one: a list of words and their respective frequencies. Below, a unigram language model, LSA, pLSA and LDA are discussed respectively, which all prove to be extensions on one another. The LDA model by Blei, Ng, and Jordan (2003) that is described lastly provides a relatively easy to use and labour-free method to get a grip on a corpus of texts. This probabilistic model that is built on top of these prior topic models outperforms these on the aspect of speed, accuracy and generalisability. It is this model (of all topic models) that is most popular in research nowadays in the Digital Humanities and other disciplines.

2.2.1 Information Extraction

In order to retrieve information from a text in an automatic way, one could make use of software that counts words and provides word frequency information. The most frequent words (MFW) in general (function words) tend to have the lowest informational value, whereas words that appear less frequently, but still several times (i.e. words that are not hapax legomena), or words that only reside in one document in the case of a corpus analysis, might be distinctive for the theme of the text and have therefore a high informational value (content words).[6]

function words

content words

By looking at the most frequent words in a text, one could already get a general idea of the text's themes. An illustrative example of this is given in Figure 2.1 that

[6] This should be combined with e.g. word sort (POS) information to differentiate between content and function words in case of ambiguity issues.

1	de (172)
2	en (96)
3	van (88)
4	het (82)
5	den (50)
6	in (47)
7	der (44)
8	te (35)
9	een (31)
10	met (27)
11	dat (25)
12	aan (24)
13	wij (22)
14	eene (21)
15	niet (20)
16	door (20)
17	is (19)
18	ons (19)
19	voor (19)
20	zijn (18)
21	zij (18)
22	over (17)
23	bij (17)
24	die (16)
25	als (16)
26	op (15)
27	des (14)
28	zoo (13)
29	uit (12)
30	men (12)
31	dieren (11)
32	hebben (11)
33	ook (11)
34	schrijver (10)
35	geschiedenis (10)
36	maar (9)
37	heeft (9)
38	verhandeling (9)
39	welke (9)

FIGURE 2.1: Example of a word list sorted on frequency. Shown is that the first content word appears at index 31. The list is built up from the 39 most frequent words in the single article `_gid001183701_01_0002`.

$$\text{weight}_{t,d} = \text{tf}_{t,d} \cdot \frac{1}{\text{df}_t}$$

FIGURE 2.2: The tf-idf weighting formula for a word in a document.

contains a word list of a single article from *De Gids*.[7] This figure shows the most frequent content words `dieren`, `schrijver`, `geschiedenis` and `verhandeling` appearing at respectively position 31, 34, 35 and 38, and they are situated much further down than the non-informative function words from this text.

stop words

tf-idf

To further improve this frequency based information retrieval, one could start using a list of stop words to remove function and other unwanted or irrelevant and common (predefined) words and keep content words that are specific for the analysed text. In the case of an analysis of multiple documents from a corpus, one could look at the weight of the words that are being used in a text, which can roughly be explained as a number that indicates how representative or unique a word is in the corpus and in a particular text. The standard weighting scheme for expressing the weights of words in a corpus is a tf-idf measure (Daniel Jurafsky and Martin, 2017, p. 278). One could use this measure to spot important words in a text. By dividing the term frequency *tf* of term *t* in document *d* (the number of times a word occurs in a document divided by the total number of words in that document) over the document frequency *df* of term *t* (the number of documents the term occurs in in the entire corpus), one ends up with a relative weighted frequency (see Figure 2.2).[8]

This relatively simple approach of picking a top-term from a document and use it as the topic (label) nevertheless does not allow for describing more complicated topics. Just one term (such as *dieren* in the above example of Figure 2.1) might not convey sufficient information to capture what the topic is about. This could be due to the

[7] Potgieter, E.J., 'Verhandelingen en losse Geschriften van P. van Limburg Brouwer.', *De Gids* (1) 1-8. <_gid001183701_01_0002>

[8] Often, these weights are normalised by converting this function to a logarithmic one.

$$\theta_1 = \{w_1, \ldots, w_n\}$$

FIGURE 2.3: Probability distribution of words for a topic, given in a unigram language model, a type of multinomial distribution that expresses the probability of a word occurring in a topic.

fact that a general umbrella term that captures the topics meaning might be absent/latent in the corpus. This means that this way of defining topics might not unveil the necessary information to correctly label or classify a text. In short, this approach of picking one term as topic suffers from (i) a lack of expressive power, (ii) an incomplete vocabulary coverage and (iii) ambiguity issues (Zhai and Massung, 2016, p. 335). Instead of using just one word/term to describe a topic, using a probability distribution over words in which a topic is expressed by multiple words can solve these problems.

2.2.2 Towards a probabilistic model

distribution

Bag of Words

Instead of using a single word from a document to express a topic, a topic can be expressed as a list of words (a distribution) in which every term is given together with its probability of occurring in this very topic. Very distinctive words for a topic appear on top of the list with a high probability, and likewise, irrelevant words appear at the bottom with a very low probability.[9] The probabilities for all the words in a corpus in such a distribution sum to one. Since the model treats the text as a Bag of Words (BoW), the document's word order is not evaluated, but frequency information is used. Therefore, the probability distribution can be seen as a *unigram* language model, since the words are seen as individual tokens without word order information. A formal representation of such a model is given in Figure 2.3 in which w is a word from a document.

More nuance is now given to the representation of a topic: this allows for multiple words that together make up the description or definition of the topic. The implication of this is that it is now (for a human interpreter) less clear cut what the topic is about. Its label or meaning should be inferred from or interpreted on the basis of this combination of words. This could very well mean that the label that best describes this topic, which might be a concept or abstract term, does not occur in the text: it is latent. However, an advantage is that this representation is able to solve the issues that were described at the end of section 2.2.1, as (i) it is better in giving the topic its distinctive meaning, (ii) it is able to incorporate more words in its representation under the same topic, and (iii) word meaning (in case of ambiguous word) can be inferred from other members of the distribution that have a high probability (ibid., pp. 335-336).

[9] But not a probability equal to 0, since there is always a (small) chance of appearing in the topic.

probabilistic topic models

This is where probabilistic topic models come into play. This type of generative models[10] are constructed by estimating the highest probability of its model for a given distribution. In the case of a topic distribution, it estimates the highest probability of the topic by estimating the combined probability of the words given that topic. In the example above of a unigram model, there is just one topic in the text that is explained or represented by a distribution of words over topic. All words from the text occur in this distribution, only in varying probability. Highly frequent words (including function words) are favoured in this model and have a high probability of occurring in the topic due to their frequency.[11]

LSA

pLSA

To extend the capability of such models, in order to describe a text in which it is known that multiple topics exist, or to apply the same model to multiple texts from the same corpus, one could use the most basic topic model: probabilistic Latent Semantic Analysis (pLSA) (Zhai and Massung, 2016, p. 368). Instead of assuming a document is made up of only one topic, the pLSA model allows for an extra distribution: a distribution of topics that make up the document.[12] The topics in this model are still expressed by means of a distribution of words and their probabilities of occurring in this very topic. This model was first introduced by Hofmann (1999) as an extension to and improvement of the LSA model, which suffers i.a. from interpretability issues (ibid., p. 3).

LSA
SVD

Topic modelling through LSA (or often written as Latent Semantic Indexing (LSI)) applies a singular value decomposition (SVD) to a matrix of documents and terms (expressed in frequencies or other weights, for example tf-idf) (Deerwester et al., 1990). What is left after this operation is a denser and more significant representation of the data, since only the most significant terms are kept in this (truncated) dimensionality reduction (Daniel Jurafsky and Martin, 2017, pp. 286-289). Information about the text and its words is reduced into an approximation of the original data so that the most important words stand out. The term-document matrix as shown in Figure 2.4 is converted into two matrices in which each column (dimension) gives information on a topic: The number of dimensions (topics) to keep can be specified and allows for a huge reduction of the data: coming from a matrix that describes a

[10] A type of model that is able to describe or "explain" the data: it is able to (re)construct a text from a set of parameters that are often created by "training" the model.

[11] This can be overcome by filtering the function words (e.g. by using a stop list) prior to constructing the distribution, or by using a so-called "mixture model" in which a (representative) background corpus is used to filter out this "noise" of words that occur frequently in any text (Zhai and Massung, 2016, pp. 345-368).

[12] Instead of thinking of reconstructing topics that could best describe a text, one should think the other way round: the topics already existed when the text was written by its author, who scooped words from them when he or she wrote the text. The model tries to unveil them so that its outcome best describes the topics that the author had in mind when writing the text. This is more or less what a generative model is assuming and trying to do.

	t_1	t_2	t_3	t_4	t_5
d_1	1	1	1	1	2
d_2	2	5	0	0	3
d_3	0	0	3	2	0
d_4	5	0	2	5	8
d_5	1	2	0	0	1

FIGURE 2.4: Example Term-Document matrix (m-by-n) for SVD for five terms (words) in the columns and five documents in the rows. Numbers are frequencies (and in this example random).

30k word vocabulary, the number of dimensions might be reduced to 50, in which each kept dimension represents a topic and gives weights to various terms from the vocabulary, which is similar to the representation of a topic by a distribution that was described above, except for the fact that the model does not include every word from the vocabulary anymore.

pLSA

The pLSA model does not involve term-document matrices. Instead, it combines multiple probability distributions to express a probability P of a word w of occurring in document d and therefore gives the LSA model a proper statistical foundation (Hofmann, 1999, p. 1). Dimensionality reduction is not implemented, and, instead, the most significant words, i.e. the ones that most likely express what a topic is about, appear by means of sorting on highest probability. How this model functions can be expressed in a graphical model as is shown in Figure 2.5.[13] For every d (the document index) in the corpus a topic distribution is generated with words w belonging to topic z. Therefore (the figure can be read from left to right), for every document d, a topic z is chosen (the outcome of the conditional probability of the topic given the document $P(z|d)$) and a word w is picked from this topic (the outcome of the conditional probability of the word given the topic $P(w|z)$). This means that the process of generating words from a topic is generative, but that generating the topic distribution for every document is dependent on which document is sampled. In Figure 2.5 this is indicated by the outer plate D which indicates that the d is indexed. In other words: there is no other information than for the documents $\{d_1 \ldots d_i\}$ (i.e. every row from an m-by-n matrix) that were present during the model's initialisation and training. The inner plate N represents a similar indexing for every word in the document (i.e. every column from an m-by-n matrix).

This immediately shows one of the shortcomings of the pLSA model: it is not purely a generative model, since it is not able to cope with new (unseen) documents presented to it (Blei, Ng, and Jordan, 2003, p. 1001). It is dependent on the documents that are used to initialise, construct and improve the model in a training stage, where

[13] This way of modelling stems from Bayesian inference. This is further explained in section 2.3.1.

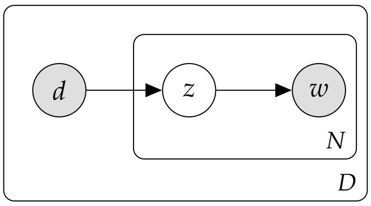


FIGURE 2.5: Graphical model of the pLSA model, as also shown in Blei, Ng, and Jordan (2003, p. 1000).

ideally a generative model should be able to function on previously unseen documents and give a prediction of topics over its contents.[14] A second culprit is the fact that the model's size (its parameters) grows proportionally with the number of documents that are analysed, which could lead to performance issues due to a lack of computational resources, but, more importantly, this increases the risk of overfitting[15] (Blei, Ng, and Jordan, 2003, p. 1001). This is important to take into account when working with a separate test and training set when one wants to apply the model to unseen data.

overfitting

These shortcomings are partly solved by introducing an alternative for the static document index *d* that is present in the pLSA model. By feeding the model a prior distribution of topics over documents that expresses the document's topic distribution, and another prior distribution of words over topics that expresses the topics' word distributions, the model is no longer dependent on both of its hidden variables[16]. This is done in further improvement of the pLSA model, in a model called Latent Dirichlet Allocation (LDA).

2.3 Latent Dirichlet Allocation (LDA)

The previous section described a short history of the pLSA topic model, which is not a purely generative topic model and which could encounter performance issues when working with large data collections. In order to solve these downsides, Blei, Ng, and Jordan (ibid.) extended the pLSA model by converting it into a Bayesian[17] model by providing it with two prior probability distributions, so-called *dirichlet distributions*. This type of distribution is named after the German mathematician Johann Dirichlet (1805-1859) and can be explained as being a “distribution of distributions” that helps the generative model in choosing several topics to draw from

Bayes

[14] This aspect is of lesser importance for this thesis, since the constructed topic model will not be used for e.g. a text classification task. The documents that are analysed are also available and used during the model's training stage. This, however, matters for the aspect of reproducibility.

[15] When a model is trained so well that it perfectly comes up with predictions for its training data, it is “overfitted.” This means that the model most likely is not representative anymore for use on other documents (Daniel Jurafsky and Martin, 2017, p. 97).

[16] Variables that are not directly observed in the data.

[17] This term comes from the Bayes' theorem which is used in statistics to calculate a probability based on existent knowledge: a prior probability.

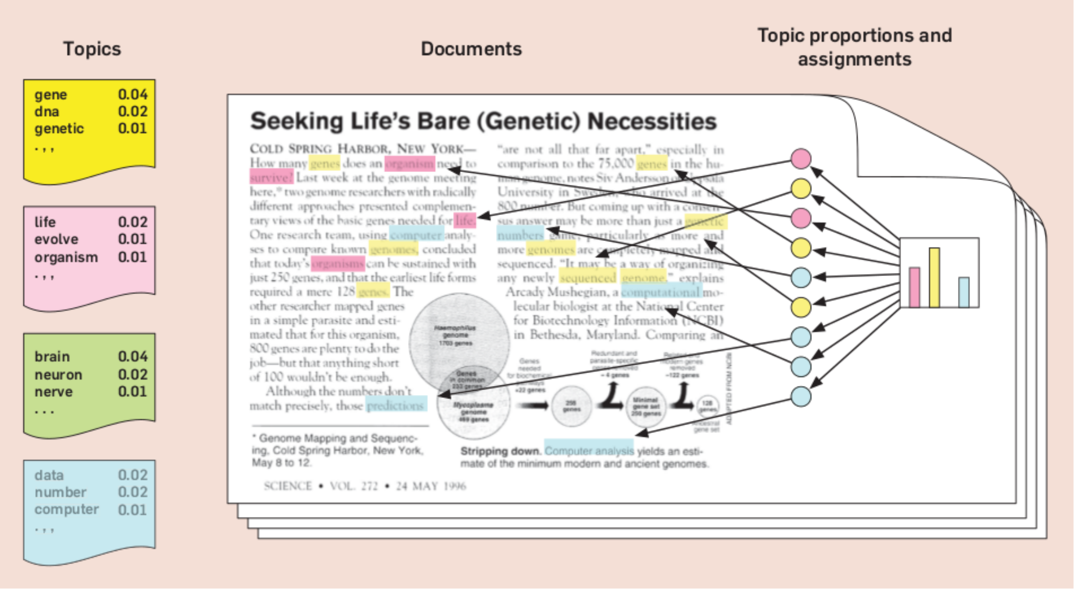


FIGURE 2.6: Illustrative example of how the LDA model works. Taken from Blei (2012a, p. 78).

when constructing a text using the first distribution[18]. Subsequently, the second distribution[19] helps in choosing the words from these topics.

It might help to visualise the workings of the LDA model with the metaphor Blei (2012a) himself uses. Constructing an LDA topic model is similar to highlighting parts of text with a fluorescent marker so that each (meaningful) coloured word or sentence corresponds to a specific topic. The image in Figure 2.6 illustrates schematically how the algorithm works. Its contents can be split up into three parts. One document from a corpus of many documents is shown. To the right, the topic distribution in this document is given in the form of a bar chart and represents the distribution θ of topics for the entire document, which is based on the topic weights in the dirichlet prior α. Every bar in this mini chart corresponds to the relative presence of a topic in the document. The coloured dots one step to the left correspond to every word w in a document that is assigned a topic proportion z. Based on this topic proportion, a word is picked that is used in the document itself (one step further, the highlighted words in the text). The sticky notes to the left are a combination of these two parameters: given the topic proportions for this single document, the words are picked. When one sums up every term picked in the document according to its topic, this overview of topics and (most) relevant terms is created. At the top of these lists are the most distinctive words given for that topic: those words have the highest probability of occurring in this topic. The direct representation of these topics is in the words that are here highlighted in the document, and these topics are not (directly) observable somewhere else: they are latent.

[18] α, which has a length of the k number of topics in the corpus
[19] β, which has the length of the $\overline{V}$ (vocabulary size) number of words.

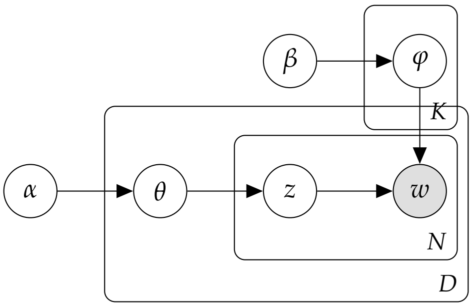


FIGURE 2.7: Further extended graphical model of the LDA in which a dirichlet prior is also used for the topics over words distribution (the second β prior to the K plate). Designed to resemble the graphical model of the smoothed LDA in Blei, Ng, and Jordan (2003, p. 1006).

More schematically, the LDA model can be described as a graphical model that shares and shows its similarity with the pLSA model. This is illustrated in Figure 2.7 in which both priors α and β are added and placed outside of the plates (which indicate some form of indexing and repetition) and an extra plate is added. These priors are corpus parameters and are generated once for the entire corpus, instead of for every document anew which is the case with pLSA (Blei, Ng, and Jordan, 2003, p. 997). The other variables are generated repeatedly, either for every document in the case of θ, or for every word in the case of z and w. The words stem from every φ, which is drawn for every topic z in θ. These word distributions are then drawn repeatedly for every topic that is specified by the distribution. This means that one document can contain words from multiple topics.

This model can also be read from left to right: For every document in plate D, a topic distribution θ is generated based on the information given in the prior α. Then, for every word that is supposed to end up in the document, a random topic z is picked from which a random word from word distribution φ is selected (Blei, 2012a, p. 78). The only difference between LDA and the previously described pLSA is the way in which it deals with the two distributions of words over topics and topics over documents. In LDA, the word distributions are drawn from outside of the model (from two Dirichlet distributions, shown in Figure 2.7 as α and β) and are drawn prior to the generative process in which documents could be generated from the model.

2.3.1 Training and improving the model

During the initialisation stage of the model, the priors α and β are built up randomly, which means that, at this point, the generative model is able to generate documents, but these are incoherent and most likely do not make any sense. After all, the model is based on random distributions. The model can be improved by iterating over several training documents (all documents in the (training) corpus), and by analysing

$$p(w,z,\theta,\varphi;\alpha,\beta) = \prod_{k=1}^{K} p(\varphi_k|\beta) \prod_{d=1}^{D} \left[p(\theta_d|\alpha) \prod_{n=1}^{N_d} [P(z_{nd}|\theta_d)p(w_{nd}|\varphi_k, z_{nd})] \right] \quad (2.1)$$

$$p(w,z,\theta,\varphi) = \prod_{k=1}^{K} p(\varphi_k) \prod_{d=1}^{D} \left[p(\theta_d) \prod_{n=1}^{N_d} [P(z_{nd}|\theta_d)p(w_{nd}|\varphi_k, z_{nd})] \right] \quad (2.2)$$

$$p(\theta_d) \sim \mathrm{Dir}(\alpha) \quad (2.3)$$

$$p(\varphi_k) \sim \mathrm{Dir}(\beta) \quad (2.4)$$

FIGURE 2.8: The LDA model from Figure 2.7 written in a Bayesian formula. The second formula has α and β removed from the equation. These are used and specified in the third and fourth formula.

the structure of these documents during every iteration in relation to predefined and learned "weights." This is done by inspecting only the words that occur together (as a Bag of Words) in a document. Depending on this analysis, the model's priors α and β are recalculated and updated to better fit the structure and thereby themes of the documents in the corpus. The model is updated in a two-step process that comprises an expectation (E-step) and maximization (M-step) of the weights in the model using the so-called Variational Expectation-Maximization algorithm (see below).

EM algorithm

Bayes

Looking back at the Bayesian plate notation of the LDA model in Figure 2.7, one can also write this model into a full equation, which is done in Figure 2.8. From left to right: the joint probability of the model is the probability of the word w, its topic z, the topic distribution θ and the word distribution of the topics φ, which are parametrised by the Dirichlet distributions α and β. The probability of the word distributions per topic φ can be expressed as the product of the probability of all distributions for all k topics in which $\varphi \sim \mathrm{Dir}(\beta)$. Further on, the joint probability of $p(w, z, \theta)$ is the product of the probability of the topic distribution θ, which comes from $\theta \sim \mathrm{Dir}(\alpha)$, and the joint probability of the topic z given the topic distribution θ and the probability of the word w given the word distribution φ (repeated for every word N). This multiplication is done for every document D.

posterior probabilities

The only known (shaded in the graphical representation) variable is w, which comes from the observable data (the words) in the document. The other variables should be inferred and these posterior probabilities[20] must be solved in order to correctly train the model. It is in these posterior distributions where the latent semantic structure of a text or corpus is encoded: what are the topics that are located in this corpus? Since there is a nearly endless number of possibilities if one wants to solve this equation, the posteriors in this equation cannot be calculated. It is, in fact, intractable, and only an approximation of the posterior is possible (Blei, Ng, and Jordan, 2003,

[20] The posterior probability is the result of the Bayesian equation.

p. 1003). Techniques such as (collapsed) Gibbs sampling (Heinrich, 2008, pp. 18-23) can be used to calculate the likelihood of the model, but since the LDA implementation used in this thesis (Gensim, see Chapter 3.3.1) is not based on this Markov Chain Monte Carlo (MCMC) variant, but instead on an "online learning" variant of variational inference, this technique is described below instead. Using this alternate technique means that it is not necessary to iterate over the entire set of training data in full (but this still can be done). The documents are analysed sequentially, which is more efficient than using traditional methods such as MCMC (Hoffman, Bach, and Blei, 2010, p. 1).

Online learning

The LDA implementation that is used in this thesis is slightly different from the standard one proposed by Blei, Ng, and Jordan (2003) and that is used in MALLET (A. K. McCallum, 2002). This implementation, which features a type of "online machine learning," is discussed below.

Online learning

Online learning is a training method in which a model is trained using data that is created during the process, or data that is not yet available when the training procedure is started (S. J. Russell and Norvig, 2010, pp. 752-753). In this method, documents are streamed to the model sequentially and are analysed, after which the model updates its parameters and is ready to start analysing a new batch of documents again. This is contrary to a more classical machine learning approach in which all available training data is analysed at once and evaluated afterwards, and is, considering the size of corpora used in topic modelling approaches, beneficial to the performance and flexibility of the model: it becomes scalable, as it can be improved or tuned over time, depending on the structure of the newly added documents, without the need to analyse previously seen documents again. This can be especially useful when analysing a diachronic corpus[21], considering its size and its presumable gradual changes (Hoffman, Bach, and Blei, 2010, pp. 1-2).

Variational Bayes

The online Variational Bayes (VB) algorithm[22] as described by Hoffman, Bach, and Blei (ibid.). It proves to be more efficient than traditional MCMC algorithms and just as accurate (ibid., pp. 1-2). The task of inferring the latent variables and thereby constructing the hidden topic structure of the corpus analysed can be reformed into an optimisation task so that these variables can be approximated. The online stochastic

optimisation task

[21] This could either be an already existent historical corpus, such as the corpus of *De Gids* used in this thesis, or be a "live" corpus that grows over time as more texts, such as news articles or tweets are published, for instance, online.

[22] Variational Bayes is the same as Variational Inference and is called Variational Bayes when it is used inside a Bayesian (hierarchical) model (i.e. a model that depends upon several posteriors that are formulated locally and globally, such as the model in Figure 2.7) (Hoffman, Bach, and Blei, 2010, p. 1). Variational Inference is a method used in optimisation approaches and is here used to approximate the latent variables in the LDA model, and thereby the density of the probability distributions in its priors α and β.

convergence

optimisation presented in this model is repeated several times for every document in the corpus until there is a minimal change observed in the tuning of the parameter settings. At this point, the model converges[23] and cannot be improved any further based on the data that has been presented to it.

E-step

M-step

hyperparameters

This is essentially the E-step of the learning stage of the model in which the true posterior (the local hidden variables θ, z and φ) is estimated by drawing a random topic distribution and by minimising the difference between the estimated distribution and the true posterior. Since the real latent parameters can only be approximated by the model, the challenge for the model is to minimize the difference between the estimated probability and the true posterior, which can be calculated by optimising the hidden variables in an expectation step. Based on these optimal parameters that are constantly changing and tuned with each iteration of the model, the model picks a new topic distribution in an M-step, in which the global (hyper)parameters α and β are updated by using a Maximum Likelihood Estimation, so that at the end of the procedure, the model has an optimal probability for all its weights and distributions, which translates into a model that ideally produces the most coherent output, since it best captures the (hidden) structure of the texts it was trained on. To which degree these hyperparameters are altered affects the model's sparsity: a large α means that documents consist of a large number of topics, which are thus more fine-grained and possibly more related. The same is true for the β on the concern of the topic-word distribution: a large β results in many less specific words per topic.[24]

2.3.2 Assumptions

Bag of Words

LDA is built upon some assumptions in order to work as instructed. First, the topic model treats the documents as a Bag of Words (BoW), which means that word order is not important when analysing a body of text (Blei, 2012a, p. 82). A Bag of Words representation only provides a list of all the words in the document together with their respective frequencies (in the document). This implies that any syntax information, and thereby perspective information, in the document is lost in this representation, but that the thematic contents of a document are still represented. The BoW representation is, in a way, related to the Vector Space Model (VSM) that, for example, can represent documents as a vector of features and their weights, which could be frequency information for each feature (word). To this extent, one can assume that documents with a similar BoW also have similar themes, assuming that

[23] Convergence is reached when the model cannot be improved any further. It might be the case that the model does not converge at all when the changes between the previously estimated parameter and the current become larger instead of smaller. This can happen when incoherent data is presented to the model, or when parameter settings are malformed or incorrectly chosen (e.g. when the model converges to a local maximum, instead of a global).

[24] How these parameters exactly affect the other parameters depends on the sort of Dirichlet distribution: symmetric or asymmetric. It follows from (Wallach, D. M. Mimno, and A. McCallum, 2009) that the model performs better when an asymmetric α is chosen, but when a symmetric β is picked. More on the implementation of these settings can be read in Chapter 3.3.1.

similar words carry similar meaning. What is more, the order in which the model analyses the documents is irrelevant, which could be a problem for sequential corpora in which topic distribution is also influenced by a temporal aspect.[25] Implementations to overcome this exist and are described, along with their own problems for Humanities' scholars in Section 2.5.

Second, the topic model assumes that all documents from a corpus can be expressed as a mixture of topics that are latent and therefore not appearing explicitly in the corpus. Occasionally though, concept labels from the corpus come to light as the most distinctive word belonging to a certain topic, but this is certainly not the case for all topics. Methods are developed to automatically label (and "identify" or "interpret") topics as produced by topic models. The most straightforward approach is picking the word with the highest probability from the cluster and marking it as topic label. More sophisticated approaches consist of using Named Entity recognition (Lauscher et al., 2016) and concept identification for topic labelling.[26] These techniques are not used in this thesis, since the topics are checked and labelled by hand, as described in Chapter 3.

generative model

Third, the model assumes that this mixture of topics exists prior to the text's construction by its author and it therefore treats the document as generated from the pre-existing topic mixture, with all available topics to choose from. This is the generative aspect of the model, as described above and essentially captures the goal of modelling in informatics in general: modelling the world in a set of algorithms. By modelling an LDA algorithm, one, in fact, creates a tool to generate a corpus of documents (as a human writer would do), each with a specific topic mixture with specific words attached to it. As was explained previously, syntax information is not captured by this particular model, since it relies on a BoW-representation of the data. However, it needs a dictionary (list) of words.

2.3.3 Limiting factors

Tang et al. (2014) provide a set of best practices, based on the limited factors of an LDA-model. They provide guidelines for choosing a suitable corpus that can be used for a topic modelling analysis. The criteria for this are that (i) the corpus should have a sufficient number of documents, (ii) the documents in the corpus should be long enough, and (iii) that one should be cautious for specifying too many topics to the model. Furthermore, (iv) the corpus should contain clearly defined topics and contain a small set of distinctive words, and (v) the model's hyper-parameters should be set correctly.

[25] Furthermore, this implies that (the probability of) each document is independent of other documents.

[26] For instance, picking a topic label by means of tf-idf (Lonij and van Eijnatten, 2016).

The article by Tang et al. (2014) shows that overfitting a model onto a corpus (i.e. specifying a higher number of topics than are present in the corpus) dramatically decreases its performance. One can learn from this that it is important to specify the right number of topics to the model. The closer this number is approximated, the better the model's performance, as indicated by the TPE (Topic precision error, (ibid.)). However, with regard to the performance measurements, the interpretability of the topics is another important factor. The more topics that are specified to or requested from the model, the more fine-grained the result can be. However, as can be read in section 2.4, a low error rate (i.e. the model is near convergence and an "ideal" distribution over topics and documents is made and reflected in the values of the model's hyper-parameters) does not guarantee that there is a result generated that also "makes sense" from the perspective of a human interpreter.

With regard to the first sub-question and case study in this thesis (see Chapter 3.6.1), it is important to investigate if the low number of documents might pose a problem. In this study into the topical contents of a single volume of *De Gids*, just 72 documents are used. However, these documents are quite lengthy with an average vocabulary size per document of 7,094 terms. A detailed description can be found in Chapter 3 in which the experimental settings for this question and the other questions are defined and explained. By and large, it seems that the results from training a model onto this low number of documents are decent or even fine: its topic clusters make sense. More on this can be found in the respective section of the results (4.1).

Additionally, data from the real world is usually messy and differs in length. This certainly is the case for the corpus of *De Gids* that is used in this thesis (see Chapter 3 for the corpus description), despite the fact that it was already available digitally. The recommendations from Tang et al. (ibid.) should be seen in perspective to this specific data. Based on their suggestions, it could be worthwhile to look at eliminating small documents (fewer than 50-100 words) from the dataset to boost the model's performance. The recommended setting should be fine-tuned on the contents of the corpus. The genre and structure of the corpus of *De Gids* that is used in this thesis resemble that of corpora used in previous research done on (scientific) journals. The settings that are used in the studies that are discussed under Related Work (2.5) might help to select the right parameters for the corpus and the model. Other settings can be deduced from trial and error.

2.4 Evaluating the quality of the topic models

Interpreting the outcomes of a topic model is not always straightforward. It is not always possible to attach a label to a cluster of words, nor are these clusters always similar to a human-based understanding of a topic. This is made clear in the article "Reading Tea Leaves: How Humans Interpret Topic Models" by Chang et al. (2009)

in which human interpretation (the art of assigning meaning to a group of words or recognising topics in a document) is compared to the construction of topics by several topic models in a quantitative study. It shows that there is not necessarily a good "fit" between the constructed topics[27] and the topics a human reader would highlight in a corpus. In fact, as this study shows, the more fine-grained the topics are, the harder it becomes for a human interpreter to identify the topic. At the same time, this study does show that the results from topic models approximate the way humans distribute documents to topics, which can be seen as justification for using topic models in Humanities research.

These findings also follow from Van Atteveldt et al. (2014, p. 2) who indicate that a topic generated by an LDA-model does not necessarily resemble a human definable topic, but they do find that their topic model is able to capture (Dutch) political issues in their corpus of Dutch newspapers quite well.[28] They even suggest that the topics from their LDA-model can substitute manual annotation for an automated approach, which makes clear that the use of an LDA topic model in research involving a large collection of (unlabelled) data is a fruitful method.

qualitative evaluation

The evaluation of the constructed topic models in this thesis will be qualitative. Giving the model's perplexity or its held-out likelihood and using this as an indication of the quality of the topic model is not an accurate measurement, as follows from Chang et al. (2009). These values will, however, be used in the model's training stage, as an indication of the training time that is needed in order to construct the best model. This optimal model is picked by manual evaluation, which is done by inspecting the top-words and top-associated-documents from each generated topics. If the topics include words that are related or are coherent within a particular subject, and the top-documents are about this theme, then it means the model most likely has a good fit within the corpus and its themes. More on the method of evaluation can be read in Chapter 3.5.

2.5 Related work

2.5.1 Journal for Digital Humanities

The interest in topic modelling in the Digital Humanities spiked in 2012, when the entire winter volume (II, 2012)[29] of the *Journal for Digital Humanities* (which quickly ceased to exist after) was devoted to the use of topic modelling in the Humanities. This issue provides an introduction to the technique, its emergence in the (Digital)

[27] Topic models are often trained until a certain perplexity value or a certain held-out likelihood is achieved, which could be a sign of the construction of coherent topics.

[28] This paper was part of a bigger project that is described in C. Jacobi, van Atteveldt, and Welbers (2016).

[29] http://journalofdigitalhumanities.org/2-1/

Humanities and some useful hands-on examples of topic modelling applications in research. It contains contributions written with the Humanities' scholar in mind by David Blei (Blei, 2012b), as well as from figures from history and literary departments. This special edition was meant to "[...] push the conversation on topic modeling and also to reflect on the larger community in which it is situated." (Meeks and Weingart, 2012).

According to its introduction, the interest in topic modelling in the Digital Humanities started around 2010, when renowned DH bloggers such as Matthew Jockers (2011) and Cameron Blevins (2010) started talking about the technique in their online writing. For a long time, their work served as an instruction set for using the technique of topic modelling in the DH, due to a lack of (Humanities oriented) literature on this subject. With the creation and rise of easy to use programs such as MALLET (MAchine Learning for LanguagE Toolkit) (A. K. McCallum, 2002), the technique was quickly adopted by the pioneers in the field and eventually found its way into DH anthologies (Weingart, 2011; Graham, Milligan, and Weingart, 2013; Jockers, 2013; Underwood, 2014). These pioneers from the field were asked to contribute to this special issue of the journal, allowing the technique to further integrate among people working in the field of DH. Several styles of research are presented by these authors, ranging from literary network analysis to historical approaches.

One of these approaches is written by Megan Brett (2012), who gives a brief overview of the technique but also describes the necessities for a successful topic modelling approach, which translate to a set of recommendations: (i) one needs a large corpus, preferably of at least 1.000 documents, although this depends on the length of each document (ibid., p. 12). This corpus has to be stripped of punctuation, numbers, capitalisation, and stop words. Moreover, (ii) one has to be familiar with the corpus so that a researcher is able to interpret the more complex topics that are returned by the model (ibid., p. 13). Next, (iii) one needs a piece of software to build and train the topic model. It is also important to find the right number of topics, which is one of the few parameters that should be set by a researcher. This number can be found by trial and error (ibid., p. 14). Finally, (iv) one needs to find a way to analyse these topics and make them insightful through a visualisation of the results.

Another study was done by Lisa Rhody (2012), who applies topic modelling to a corpus of 4.500 poems. She notes that the process of writing an ekphrastic poem is similar to the workings of the generative LDA model: picking from tropes and conventions to build up the poem by a poet (ibid., pp. 18-19). However, poetry is fundamentally different from the standard corpora that are used in topic modelling studies, because of its use of figurative language. Where a topic model outputs topics that resemble themes in non-figurative texts, the model might come up with different results in the case of more literary documents. It is important to be aware of this when one uses a corpus of mixed text types, or when one uses LDA on a corpus of literary texts only as Rhody (ibid.) does. She removes stop words from her corpus

poetry

and trains a 60 topic model. Because of the nature of poetic texts, a label to describe a topic does not suffice. Instead, the topics can be regarded as modes of discourse (Rhody, 2012, pp. 29-32), such as (i) topics consisting of corpus artefacts or words from other languages that make up a small amount of the corpus that are being clustered, (ii) topics that are dominated or skewed by words that occur multiple times in longer documents, (iii) topics that seem very clearly defined, although close-reading or interpretation in light of the documents "discourse" might tell otherwise, and (iv) topics that cannot be identified without close-reading of similar texts, which might have different themes, but the same discourse. Looking at this study, it can be concluded that topic modelling can still be applied to a literary corpus, but that the result is much less straight-forward and depends upon analysis on the level of discourse, rather than theme.

Literary text in the form of diary entries was analysed earlier by Cameron Blevins (2010), who writes about a small-scale experiment of applying topic modelling (as part of many other ways to extract information from this corpus) to the diaries of Martha Ballard, an American midwife born in the 18th-century. He generated 30 topics on his corpus of more than 10.000 (small sized) entries by using MALLET and labelled them according to their theme. With the date information from the diary entries, he is able to see a fluctuation of topics over time. He uses the corpus as is, although he most likely makes use of the built-in stopword list of MALLET.

Such an addition of a temporal dimension applied to contributions to the journal of the Modern Language Association (PMLA) is done by Goldstone and Underwood (2012), who group their most interesting findings of their experiments on this corpus. They, for instance, show that they can model the rise and fall of the Structuralism movement as a trend in research, but fail to explain its rise that is seen much earlier in the corpus than is commonly presumed in literary history: the topic model does not grant an explanation for this. This study into the history of literary trends is intensified in Goldstone and Underwood (2014) and is briefly described further below. Furthermore, they experiment with representing a distribution of topics in a network graph, but at the same time warn for drawing hasty conclusions from the seemingly meaningful network diagram (Goldstone and Underwood, 2012; Underwood, 2012b). To compare similarities between topics, or to plot topics on a two-dimensional scale, one can use a PCA, although this still has the issue of obscurity and interpretability, or hierarchical clustering (Underwood, 2012b, comments section).

trends

The final contribution to the journal is given by Schmidt (2012b) (which builds on Schmidt (2012a)). He states that interpreting the topic on the basis of inspecting the coherence and meaning of the top (five) terms is not the right method, as much information remains obscured this way: recall that every word from the corpus is included in the distribution in varying probabilities. Wordclouds already add some more information in the form of scaling according to word frequency or another weighting, but this representation, according to Schmidt (2012b, pp. 51-52) still relies on the textual representation and its semantics. To make this potential problem more transparent, he uses a corpus of coordinates rather than texts and shows that the most common coordinates (instead of words) in his case do not come close to the centre of the collection of coordinates that makes up the topic. In other words: the most common terms do not have to be the terms that make up the meaning of a topic (ibid., p. 54). This is related to the existence of so-called "chimaera topics," that occur when the model appears to have described two themes in a single topic, of which one would then be missed when only the top terms are analysed. This shows the importance of interpretation in context: preferably, one has to look at a large portion from a topic word distribution, preferably in context. I should point out that these extreme examples that are given are all selected specifically to illustrate the problems of an LDA model for his article and might not come up in a real, well-tuned topic analysis of textual corpora. The remarks that Schmidt (ibid.) gives on dealing with diachronic corpora are discussed below in Section 2.5.4.

2.5.2 Handbooks and anthologies

Shortly after this journal was published, another handbook on this topic came out. The handbook *The Historian's Macroscope* by Graham, Milligan, and Weingart (2013) was written on the same entry level as the Journal for Digital Humanities. This book provides an accessible entry, including a tutorial to topic modelling for the history scholar/student at a very basic level (using MALLET). They state that the main reason for using topic models in Humanities research is to gain new insights on some (text) collection: "[...] to generate new ways of looking at our materials [...]" (ibid., pp. 119-120), and not to use the outcomes of a model for fact-checking, since the model's configuration is prone to ambiguous interpretations and cherry picking. One really needs the context, instead of ordered lists of words, even if the lists are constructed as word clouds, providing a bit of context through a shared presentation. This follows from a blog post by Schmidt (2012a), who visualises how the (standard) settings of a topic modelling tool can affect inference on the results by modelling (non-textual) whaling routes, although this example is largely limited to picking the "right" number of topics.

A similar approach and introduction for the literary scholar or student is provided by Matthew Jockers (2013) in his *Macroanalysis*, in which he gives a brief overview from the perspective of literary research into themes in literature. He sees topic culturomics modelling as the solution to finding a better entry into the field of "culturomics", a term that was launched by Michel et al. (2010) in relation to the Google N-gram tool[30]. With topic modelling, a researcher is a step closer to automatically attribute

[30] A tool that displays the relative frequency of one or multiple words over time. See: https://books.google.com/ngrams

meaning to larger excerpts of text, since topics that are created by a (probabilistic) topic modelling algorithm overarch the meaning that is attributed to uni-, bi- or trigrams into bigger, more conceptual clusters of words (=topics), so that they are, in theory, able to capture the themes and motives of (literary) texts (Jockers, 2013, p. 122). Jockers sees the potential of this technique for an application in literary analyses, and mentions some example studies of topic modelling in the humanities. The contents of some of these example studies are briefly discussed below.

As an example of a topic modelling approach, Jockers analyses a corpus of 3000+ books from the Stanford Literary Lab[31] with MALLET (A. K. McCallum, 2002). Using a relatively high number of topics (500) in his settings, he finds that most of the automatically constructed topics are relevant or make sense. It is therefore important to notice that the occurrence of irrelevant or scrambled topics is not detrimental for the overall result of the analysis, since the clear topics still prove to be useful for the interpreter (Jockers, 2013, pp. 128-130).[32] He also shows the benefits of using a topic modelling approach over a word collocation search for literary analysis, as the first might be able to identify words belonging to a certain theme that a keyword search might miss, for instance, due to the fact that the words are placed too far apart (ibid., p. 127). Jockers uses wordclouds to visualise a topic's word distribution, which also provides an accessible manner to analyse the semantics of a topic for himself and for the reader (ibid., p. 130). At the same time, the word clouds are a transparent way to support the labelling choices made.

wordcloud

Jockers further uses his constructed topics and their labels to highlight themes in male and female fiction (ibid., pp. 136-146), and to show theme proportions in American, British and Irish literature. He also gives an example of a thematic trend analysis over a 100 years of novels (ibid., pp. 145-147) and tries to associate themes with author and nationality (ibid., pp. 147-153) using a classifier. From this, although his experimental settings are not that well described,[33] it can be read that Jockers has separated his corpus into groups of gender and nationality. He also feeds chunks of 1000 words to his topic model, instead of entire novels at once. This might be beneficial for finding smaller and more specific topics since those will not be "lost" in the general BoW of the entire novel this way (ibid., p. 134). There is, however, no strict guideline or recommendation regarding text segmentation when using LDA. The LDA model allows for multiple topics per document. Thus, using documents or chunks that are too small might be detrimental to the creation of topic word clusters. Chunking on marked text segments (e.g. chapters) could be the most intuitive option when dealing with large texts. One benefit of this approach might be that chunking

chunking

[31] `https://litlab.stanford.edu/`

[32] The incoherent topics might be coherent for the algorithm on some aspect that is not valued by a human interpreter.

[33] Instead, this is done in a separate article "Significant themes in 19th-century literature," see Jockers and D. Mimno (2013).

the novel also allows for an analysis of a shift in topics in narrated time, and thus a change of topics over the length of the novel (and not necessarily narrative time).

The more technical and related paper by Jockers and D. Mimno (2013) shows that the results regarding gender and whether a work is published anonymously are statistically significant.[34] They only analyse nouns, which works well for their approach, but is not recommended as a general step in preprocessing (ibid., p. 754). Moreover, it is noteworthy that they do not lemmatise their corpus and continue with singular and plural forms of nouns. Finally, the metadata that they are using is never used in training the model and serves purely for answering their hypotheses.

Another tutorial-like guide can be found in Graham, Weingart, and Milligan (2012) and a practical approach for humanities' scholars can be found in Weingart (2012).

Literature on topic modelling in the Humanities mainly deals with corpora with non-fictional or non-literary texts. More attention is given to analysing periodicals, such as (scientific) journals, newspapers and conference contributions and proceedings. Another example of this is the small study on the history of German Studies by Allen Riddell (2014), who analyses over 20.000 articles and book reviews from various German Studies journals. In his article, he explains his method to infer topics from his corpus by using an LDA topic modelling approach and he visualizes the prevalence of certain topics in his corpus in which he is able to detect a sudden surge or decline in topic popularity per year in his corpus. This prevalence is expressed by the relative topic portion in contributions belonging to a certain year. Topics in his study are expressed by groups of characterising words, which he himself annotates with labels (e.g. "Goethe" or "folklore") by means of interpretation. Where possible, he explains these surges and declines in interest by connecting them to societal events or other factors such as birthdays and special issues of Goethe and Grimm respectively. Apart from signalling some notable shifts in topic proportion, Riddell only suggests further exploitation of the topic modelling technique, for instance, to check for existing accounts of (German) history in his corpus and verifying hypotheses regarding these accounts (ibid., p. 109).

Another paper from the field of literature is written by Goldstone and Underwood (2014), who write about the history of Literary Studies with regard to research trends. They use topic modelling on a corpus of 21.000 articles from the past 120 years in order to reveal "important but hitherto unarticulated trends in literary scholarship" (ibid., p. 360) and indicate that topics could be seen as much broader constructs than themes alone: idiom and discourse could possibly be captured by the algorithm (ibid., p. 361). They prevent articles of less than 1.000 words from ending up in their corpus, they focus solely on scholarly articles, and they use a stop word list that contains frequent words and several proper names. The LDA implementation that they use is the standard one present in MALLET. In the results of their model, they can see

[34] They, for instance, implement a permutation test for the gender differences and they bootstrap for the variance in the corpus.

a change in research style (e.g. interpretation) and see several fields of interest (e.g. formalism) emerging throughout the twentieth century. They attempt to interpret these results in light of what they know about the history of Literary Studies. Overall, it seems that their model is able to capture the changes in literary disciplines that are already known, it signals a change in literary vocabulary, and it also provides aid in investigating if changes in style occurred that abruptly in history as is claimed about the return to Philology or New Formalism (Goldstone and Underwood, 2014, p. 375). Furthermore, Goldstone and Underwood (ibid., pp. 365-366) discuss the urgency of scholarly interpretation when working with statistical models and the need to always provide a context with the search result. This can simply be accounted for by looking back at the material and by becoming aware of the outlines of the corpus that is used.

The studies described above show that the authors from the literature and history side of the Humanities all see topic modelling as an instrument that aids interpretation, as a heuristic device, in which uninterpretable results can be neglected in favour of the analysis of coherent word clusters that a topic model produces. It offers a way to gain new insights into a corpus' contents. This is contrary to the more linguistic and exact approach by Newman and Block (2006), Hall, Daniel Jurafsky, and Manning (2008) or Hoffman, Bach, and Blei (2010) who use the technique as a clustering method and even as a verification/falsification device for hypotheses regarding topic trends in diachronic corpora.

2.5.3 Looking at diachronic corpora

Observing changes in topic distribution diachronically was done by Newman and Block (2006), who investigate volumes of an 18th-century colonial American newspaper: the *Pennsylvania Gazette* (1728-1800). In their paper, they compare several topic modelling techniques (LSA, pLSA and K-means clustering) by running a topic modelling analysis on their corpus. They see topic modelling as a useful technique to confirm known trends (ibid., p. 754). Instead of using common validation techniques such as measuring the model's perplexity, they use a qualitative human evaluation and see which topics make the most sense from a human's perspective (ibid., p. 756).

The steps in their analysis of this newspaper are clearly described. They implement stop word removal with a custom-tailored list of stop words, words smaller than three characters and words that appear less than 6 times in the corpus. They also implement stemming in their pre-processing, but this was only limited to removing the English plural s as a word's suffix. (ibid., p. 757). They set the number of topics to 40 and provide a manually annotated label for each topic. Picking this amount was done by inspecting the likelihood of each model with several numbers of topics, but the interpretability of the topic and whether the topics "looked good" were the deciding factors in this (ibid., p. 764). The interpretation of the topics was done by

stemming

a historian who is familiar with the contents and contexts of the *Gazette* and who is able to connect historical changes and political influences to the newspaper to the generated topics. When it comes to running an analysis over time, Newman and Block (2006, p. 765) cluster the articles from the newspaper per decade and give two examples of an economy and culture related trend in their corpus, which shows the possibilities of using a probabilistic topic model for historical research.

On a more conceptual level Hall, Daniel Jurafsky, and Manning (2008) investigate historical changes in contributions to various linguistic organisations (such as the ACL and COLING) by following the idea of Kuhn's paradigm shifts. This is one of the studies that verifies existing claims and hypotheses with the use of topic modelling. Instead of limited (anecdotal) evidence, this technique is able to provide statistical relevant verification of the hypotheses. The authors do so by training an LDA model on the contributions to ACL over time. They are able to detect obvious (new techniques, computerised methods) and more subtle changes to the field (e.g. a more theoretical focus). The topics they investigate are partly built up automatically by the 100 topics model and partly by manually picking terms and providing this as prior information to the model.

What is more, in the paper that described the online learning technique by Hoffman, Bach, and Blei (2010), an example study is done into the topical contents of a corpus of *Nature* and Wikipedia articles. This is primarily done to test the added value of the online learning approach, but also to provide an example of a topic modelling approach using this technique on large-scale corpora: more than 350.000 *Nature* articles and 100.000 Wikipedia articles. They implement pruning of these documents by keeping only a selected vocabulary of common terms of the texts, by which they reduce the size of the documents significantly without compromising on the performance of the model (ibid., pp. 7-8). They used different settings but a fixed running time of 5 hours per model. The evaluation in their paper is done by measuring the model's fit by looking at the perplexity of the model on held-out data.

SLDA

Nelson (2010) also looks at a large historical periodical in their *Mining the Dispatch* project. A paper related to this project was published by Templeton et al. (2011) who apply a slightly altered version of LDA on this corpus. By using a "supervised LDA" approach, they transform the "undirected" topic search into a direct one, with which they are able to steer the models aim to shed light on predefined themes. This might be helpful when a researcher already knows which themes are prevalent in the corpus. Steering the model towards these themes that are most likely going to be seen in corresponding topics might lead to a better or more correct topic distribution. Using this supervised approach also gives substance to any metadata that is available for the corpus that otherwise remains unused. The supervised approach of SLDA is used by Templeton et al. (ibid.) to discover wartime topics in the *Richmond Daily Dispatch*, a Confederate newspaper. As extra information to the model, they give the Confederate casualty count during the Civil War per week for a published document from the corpus, which should correlate with the number of war-related topics in a document. Again, such a supervised approach is only usable if there is sufficient knowledge about the corpus and some form of labelling or metadata is existent. In this case, this data came from a different corpus-external source. It can be read in Rhody (2012, p. 24) that the contents of the corpus of the *Dispatch* included all kinds of text types, such as poetry, opinion pieces, and political reports, that were all modelled together.

A study that includes historical, as well as more modern newspaper contributions was written by T.-I. Yang, Torget, and Mihalcea (2011). They topic model a corpus of Texan newspapers from 1829-1930. Every time period in their corpus is described with background information from a historical point of view, which helps them to interpret the results that are returned by their model. In their pipeline, they implement spelling correction (to correct OCR errors), Named Entity recognition and stemming, but they also build a topic model using MALLET after each of these steps for evaluation purposes. The standard stop word list from MALLET is used, together with their own additions. Based on the three resulting models, they discard the NER and stemming results, since they find that stemming returns uninterpretable words and just the spelling corrected results seem fine too. The results are evaluated qualitatively by a historian, who signals topics that are related to economic aspects and were expected for this corpus. An accuracy score is given for each of the topic groups they analyse and for each of these topics an explanation or context is given. Besides signalling known events, the topic model also comes up with new insights, but this had to be spotted by the expert, which again shows that some degree of expertise on the material is mandatory.

2.5.4 Tracking topics over time and objectivity

A word of caution regarding the extension of the original LDA algorithm to better accommodate sequential corpora is uttered by Ted Underwood (2014, pp. 71-72), who reflects on diachronic topic modelling from a historian's perspective. He mentions Dynamic Topic Models (Blei and Lafferty, 2006) and Topics Over Time (Wang and A. McCallum, 2006) in relation to the idea of "historical continuity" in particular. He writes: "Historians are probably better advised to rely on a simpler algorithm like Latent Dirichlet Allocation, which remains blissfully ignorant of dates and yet in practice tends to produce coherent diachronic patterns." (Underwood, 2014, note 15). This fallacy is also mentioned in Goldstone and Underwood (2014, pp. 366-369).

Schmidt (2012b, p. 56) also mentions the problematic aspects of such models for a "humanist." These kinds of models assume that there is a historical change in topics, which influences the topic distribution if there is no temporal dimension present for a theme. This produces a result in which topics are created that in fact do not really exist. Furthermore, they assume that historical events happen continuously and penalise documents and topics that do not fit the pattern of occurrence and do not occur in these time periods (Schmidt, 2012b, pp. 63-64). In other words, these topic modelling techniques might be, from the point of view of a humanities scholar, not the most ideal way to deal with sequential corpora that most likely contain historical changes. Although Wang and A. McCallum (2006) claim that their technique is an improvement of the standard LDA approach, trading in this upgrade for a more refined, historically more valid, and unbiased result might be better.

This difference in how historians and literary scholars deal with topic models compared to linguists or researchers from computational studies also becomes visible in the way the technique is used. It seems that historians and literary scholars deal with their material with much more caution, and are hesitant to use statistics from a topic model to answer hypotheses. Topic models might give the illusion of an objective technique. However, the reality is quite different. Ways of using the technique in "humanistic research" can roughly be categorised in three categories coming from Roland (2016): "Topic modeling's use in humanistic research might be thought of in terms of three broad approaches: as a tool to guide our close readings, as a technique for capturing the social conditions of texts, and as a literary method that defamiliarizes texts and language." All the studies that were described in this chapter are examples of these styles.

2.5.5 Example studies using Dutch corpora

So far, the literature on topic modelling provides small-scoped case studies with mainly English corpora to illustrate the way it works, or to present a development of the technique. Studies that are merely focused on the technique itself or criticise it are more frequent than studies that actually apply the technique by outlining the contents of a periodical or fictional work. The number of studies that are using topic modelling on Dutch corpora is even scarcer, but a few can be described.

The ePistolarium project of the Huygens Institute for the History of the Netherlands[35] experimented with topic modelling (testing LDA, LSA and Random Indexing) and used it to calculate a similarity score between two letters from a corpus of seventeenth-century scholarly correspondences (Wittek and Ravenek, 2011), and dealt with multilanguage and spelling issues. They implemented the technique as a classification device, which is used in the online exploration module to assist the user in finding his or her way through the corpus.

A stand-alone tool that is tested on the KB's newspaper archives was made by Huijnen and Lonij (2016), who built a "Keyword Generator" that allows a researcher to find broader themes in the newspapers than are explicitly described: it generates the latent themes from the corpus based on extracting a dictionary of terms. To build

[35] http://ckcc.huygens.knaw.nl/epistolarium/

this word list, they implemented several topic modelling techniques, of which LDA is one. The tool then offers its user the possibility to discriminate several irrelevant topics that are returned, after which the KB's Dictionary Viewer[36] is used to search for these terms, which cause the topics and their occurrence in the corpus to be visualised on a temporal scale. Thus, the technique is used as an extended search tool.

A small scoped study on using topic modelling for analysis on Dutch literature is described by Jautze, van Cranenburgh, and Koolen (2016) who investigate what a topic modelling algorithm can do to explain the judgements on the literary quality of a novel by looking at what they call "mono-topicality." They implement lemmatisation and stop word removal of function words and names, and chunk the text in portions of 1.000 tokens, after which they use MALLET to create a 50 topic model. To link topics to literary themes, they build upon the ideas by Jockers and D. Mimno (2013) and show that an even topic distribution correlates with a higher literariness. To explain this, they are using the created word clusters and their presumed meaning and compare novels with similar topics, so again the technique is used as an (implicit) classification mechanism.

It is difficult to find other examples of studies that use topic modelling techniques on Dutch cultural (historical) data. More extensive studies that use the technique are carried out in de field of the Social Sciences, in which researchers analyse political or juridical data. An example of this is the paper by Van der Zwaan, Marx, and Kamps (2016) that is aimed at the validation of "Cross-Perspective Topic Modeling," an extension to a topic model that is used for opinion mining. They use a corpus of Dutch parliamentary texts[37] in which they analyse nouns (corresponding to topics) and adjectives (corresponding to opinions) separately in two models. The model's vocabulary is pruned by removing words that occur less than six times, as well as the top 100 terms. On their corpus of around 20.000 documents, they train a model of 100 topics. Statistical tests then show that the topics and opinions that are created make sense from a machine's point of view: high-quality topics are returned, which advocate the use of this technique in the field of Political Science. But, as the authors indicate, even results that are backed with statistically valid data have to be checked to make sense from a human perspective (Ibid., p. 35).

The lack of other studies that use Dutch corpora could be related to a lack of digital and accessible datasets, their quality, or to difficulty with the contents of these corpora, which is probably connected to finding suitable and verifiable hypotheses to work with. In any case, this void illustrates the necessity of new studies into Dutch cultural texts.

[36] `http://www.kbresearch.nl/dictionary/`

[37] Gathered in `http://search.politicalmashup.nl/`.

2.5.6 Derivatives and extensions of LDA

There are numerous extensions to the original LDA algorithm. A small implementation, such as the more effective learning method described above, is an example of this, but there are more severe changes to the model made to add more ways to (automatically) structure information in a given corpus. This includes the use of metadata, such as the author of a text. The author-topic model as described in Rosen-Zvi et al. (2004) adds another layer of information on top of the topic and word distributions from the standard LDA model, so that a distribution of topics per author is created. This way, it becomes clear which subjects an author writes about and how his writing relates to another author. This paves the way for authorship attribution through topic modelling, since the model can be used as a classification device that can assign the most probable author to a document. The model functions on the assumption that an author has a specific or unique way of building a text from various topics, and that this author uses the same topics in all his or her writing.[38]

Similarly, in Structured Topic Modeling (STM) (Roberts, Stewart, and Airoldi, 2016), any document attribute can be added to the model as a covariate[39], which tightens the use of this model to any available metadata on the documents. The idea is that the structure, i.e. the prevalence and content of topics, is influenced by the text's metadata, such as date or author information. How labelled data is used is also shown in the Supervised LDA (SLDA) algorithm by McAuliffe and Blei (2008) who test this method on predicting film ratings from a review (sentiment analysis) and finding the political tone for political amendments. From their study, it turns out that SLDA outperforms traditional techniques in prediction tasks, such as predicting the score of a review. Thus, when one works with a labelled corpus, which is crucial, one is best off by using a supervised version of a topic modelling algorithm. However, other (supervised) prediction techniques might perform better in such tasks. It is ultimately the non-supervised and non-extended LDA that is strong in unsupervised tasks.

One of the latest developments includes the use of word embeddings in combination with the LDA algorithm. Moody (2016) describes a system in which the word vectors of a word2vec model are incorporated into a topic distribution of the LDA model, which creates clearly defined and highly interpretable topics over a corpus. The downside to this approach is that this technique needs a lot of processing power and resources, and it does not necessarily produce more coherent topics for an interpreter, compared to an original LDA approach. Nevertheless, this development of integrating word embeddings into known techniques such as LDA is promising.

word2vec

[38] If one thinks of a news corpus, specialists on a subject, such as foreign policy, economics, literature and arts, each tend to draw from these same subjects in all their writing.

[39] The authors of this model have released a package for R, see also http://www.structuraltopicmodel.com/ for a list of publications using this technique.

2.5.7 Other approaches

Apart from using topic modelling techniques to get insight into the themes of a text, the technique can also be used as a classification mechanism in the field of authorship attribution (M. Yang et al., 2018). This means that the technique can possibly be used in finding the author of an anonymously published (literary) text, or in finding the proper author of a letter in the juridical domain. Different settings and different parameters are needed in such applications, since it turns out that stop words, that are mostly filtered in approaches that try to give information on the contents of a corpus, prove to be a strong marker for authorship. The assumption here is that the same author uses the same topics in their texts, or the same distribution of topics that do not necessarily relate to human interpretable topics.

It was already stated that the technique is not limited to textual data. A good example from the field of the Spatial Humanities is the study by Schmidt (2012a) who plot topics in the form of ship routes on a map, but topic modelling is also used in the field of musicology (Shalit, Weinshall, and Chechik, 2013) or outside the Humanities for modelling, for example, DNA sequences in bioinformatics (Liu et al., 2016).

2.6 The periodical: *De Gids* (1837-)

2.6.1 The foundational years

Establishing a new general cultural magazine in the nineteenth century was not that easy. Responding to a zeitgeist that produced bland and languid publications in literary magazines such as *Vaderlandse Letteroefeningen* (1761-1860), the publication of *De Gids* answered the shriek for new literary trends and developments, as well as the need for a critical and innovative periodical in 1837 (T. Jacobi and Relleke, 1987, pp. 8-9). The first issues, up to the third volume of *De Gids* were published with the subtitle "Nieuwe vaderlandse letteroefeningen," which could be seen as a taunt directed at the magazine that existed from 1761-1876, the leading literary journal at the time which was *De Gids*s main rival, according to its first editors. After two years, this subtitle was removed from the cover and the magazine continued to appear on a monthly basis.

De Gids had to be more innovative than other magazines and had to take the discussions on arts and sciences to a new level, giving attention to developments from abroad in the field of literature, culture and science. To accomplish this, authors, scientists and other cultural agents were invited to contribute to the magazine, though the main contributions were written by a small group of authors gathered in *De Redactie*, of which Johannes Potgieter (1908-1875) was by far the most active and polemic (Van den Berg and Couttenier, 2009, pp. 205-208). Together with Reinier liberalism

Cornelis Bakhuizen van den Brink (1810-1865) Potgieter created a periodical that distinguished itself by means of its liberalist ideology, focussing on objective criticism and individualism. It joined the cultural and scientific profiling of other European nations in the nineteenth century by exploiting and cultivating Dutch history and culture (Van den Berg and Couttenier, 2009, pp. 208-215). Prose, and not poetry, had a large role to play in this, acting as an ideal method to convey contemporary science (as education mechanism) and artistic style, as could be deduced from the reviews in the magazine that included contributions from specialists from various (scientific) fields.[40]

2.6.2 Change of guards: a change in editors

What had started as a project by Potgieter and Robidé van der Aa (1791-1851) that in the first editions mainly contained self-written literary contributions and reviews, had become a commonplace for experts from all kinds of disciplines, ranging from literature and politics to health and physics. This made *De Gids* a strong influencer amongst scientific journals (T. Jacobi and Relleke, 1987, p. 34), especially after the proposed changes by jurist Gerrit de Clercq (1821-1857) to the magazine in 1847. He wanted to transform it into a more general medium, offering more than just book reviews and fine literature, such as reflections on contemporary foreign and internal politics, arts, and sciences (Aerts, 1987f, pp. 31-32). From 1848 onwards, the magazine was propagating discussion and a freedom of speech with respect to judging the quality of one's work without taking the author's rank and position into account. This was accompanied by a change in contents: fewer reviews of professional or academic journals and more writing was meant to uplift and educate the reader. The periodical did not eschew controversial or socially relevant subjects at the time. This made *De Gids* one of the driving forces behind the liberal movement. Unsurprisingly, it also supported the tenure of prime-minister and former contributing author Johan Rudolph Thorbecke (1798-1872) from 1849-1853 and his party and ideals after (Aerts, 1987f, pp. 34-35; Aerts, 2018, p. 323). This liberal movement also turned out to be the saviour of *De Gids*, which had nearly met its end a number of times in its first years (Aerts, 2012).

Modern Theology

De Gids continued to interfere with political developments, such as the Dutch constitutional reforms, the exploitation of the Dutch colonies, theological debates connected to the discussion around science and religion, modern theology, and literature. Contributions regarding the latter can be ascribed to writer and critic Conrad Busken Huet (1826-1886) (Aerts, 1987f, pp. 38-47). He also dared to criticise the Dutch government and king in 1865, which was, at that time, again lead by Thorbecke, with whom the magazine sympathised at the start (ibid., pp. 49-52). Freedom

[40] This did not make *De Gids* very suitable for a large audience. The number of subscribers fluctuated in the first years around 220 and did not increase. See: T. Jacobi and Relleke (1987, pp. 19-21) and Van den Berg and Couttenier (2009, p. 214).

of thought and open discussion outweighed respect for the constitutional monarchy and its actors, as well as the position of fellow contributors. This lead to the resignation of Potgieter and Busken Huet as editors of *De Gids*. As a result, the focus of the magazine further shifted towards informing about and reviewing developments from the arts and sciences in The Netherlands and abroad. By doing so, it aimed at a well-educated liberal audience (Aerts, 1987b, pp. 54-55).

What *De Gids* had done with *Vaderlandse Letteroefening* was more or less repeated in 1885 by the founders of a new magazine called *De Nieuwe Gids* (1885-1893[41]), which was put on the market as “a magazine for literature, art, politics and science” (Van den Berg and Couttenier, 2009, p. 580). The people behind *De Nieuwe Gids* pointed out that *De Gids* was paying no attention to the literary developments from the *Tachtigers*, despite the fact that the periodicals focus during these years was on literature, instead of arts and sciences (Aerts, 1987b, p. 74). *De Gids* opposed the Movement of Eighty and disputed Naturalism and Pessimism in literature, which caused the magazine to be criticised for showing little innovation (ibid., pp. 77-79) at this time.

2.6.3 During the fin-de-siècle and the twentieth century

Literature and art in the form of music, theatre, and painting obtained a prominent place in *De Gids* from 1890 onwards. The magazine continued to function as a springboard to the publishing of integral novels in this time, as it, for instance, featured parts of unpublished works by Louis Couperus as feuilleton (Bel, 2015, pp. 69-70). What is more, the increased attention for new scientific developments could be seen in the contributions from authors such as Jacobus van ’t Hoff (1852-1911) on newly discovered elements (chemistry), Hendrik Lorentz (1853-1928) on X-rays (physics), Gerard Heymans (1857-1930) on experimental psychology and anthropology, and Ambrosius Hubrecht (1853-1915) on zoology (biology) (Aerts, 1987a, pp. 96-97). During the end of the nineteenth century, the magazine also featured contributions on the Feminist Movement, with, for instance, Aletta Jacobs (1854-1929) as contributing author (ibid., pp. 100-101). Most space in the magazine was still reserved for literary contributions (40%), followed by contributions related to arts, sciences, and societal matters (30%). De rest of the magazine was reserved for texts about literature from Dutch, as well as foreign authors (10%), other news and arts-related matter, and bibliography (20%) (Aerts, 1987d, pp. 120-121).

The number of authors with a scientific background in *De Gids* was low and did not reflect the attention there was for the sciences in society at that time (Aerts, 1997, p. 452). Therefore, the editors decided that *De Gids* needed a permanent staff that

[41] In fact, the magazine continued to exist until 1943, but not in its original form and not with its original staff. Willem Kloos (1859-1938) took the lead when he got rid of the other editors Frederik van Eeden (1860-1932), Frank van der Goes (1859-1939) and Pieter Lodewijk Tak (1848-1907) in 1893 (Van den Berg and Couttenier, 2009, p. 617).

could write on subjects related to the natural sciences, and in 1916[42] they found physician Johannes Petrus Kuenen (1866-1922) willing to join the board of editors. He produced, amongst other things, an article on Einstein's Theory of General Relativity. Contributions by other authors included the history of the natural sciences and publications, in moderation, on philosophy and ideology (Aerts, 1987d, pp. 122-123). In an attempt to restore *De Gids*'s connection to the literary world, Adriaan Roland Holst (1888-1976) was responsible for the poetry section of the magazine from 1919 onwards. He succeeded in persuading new and upcoming names such as Hendrik Marsman (1899-1940) and Jan Slauerhoff (1898-1936) to contribute to the magazine (ibid., pp. 125-129). He was later joined by Martinus Nijhoff (1894-1953).

The magazine criticised internal politics more and more often, but political developments in Germany and Italy were watched very closely as well. These developments were commented on and mostly criticised. In 1933, editor Herman Theodoor Colenbrander (1871-1945) was accused of plagiarism in one of his contributions. The discontent of other contributors and editors about this issue lead to the instalment of an almost entirely new board of editors by Colenbrander in 1933. The continuity of *De Gids* was saved, its composition was renewed and it carried on publishing on literature, arts, sciences, and history (ibid., pp. 137-139). Its main content consisted of essays on cultural, philosophical, and historical topics, as well as history of science related matters when the magazine reached its 100th anniversary.

2.6.4 Nowadays

During the Second World War the *De Gids* refrained from publishing political matter, but continued to be published, albeit in a watered-down version, until 1944 (Aerts, 1987g), to start again in 1945. The magazine recovered after the war, but almost no room was left for literary content. Instead, the magazine was reproached of being too academic (Calis, 1987, pp. 162-163). A fusion with another magazine and multiple changes of publisher eventually modernised *De Gids* from the 1950s onwards and it regained its literary character (ibid., pp. 166-175). Moreover, scientific and urgent subjects were barred from the magazine or referred to separate editions to make way for literature and culture, although a mixture of science related matter and contributions on culture were always present up to the end of the 1980s.

Despite all the times that the magazine changed editors, faced declines in subscription numbers, and saw the rises (and falls) of other literary or cultural magazines, it is to this day still published on a regular basis. *De Gids* has shifted its focus towards the arts and it now mainly features short literary stories, poetry and essays. It is published six times a year and is bundled with an issue of *De Groene Amsterdammer*

[42] Huizinga, Johan. 'Colenbrander als gidsredacteur', *De Gids* (105) 1941 2-5. <_gid001194101_01_0002>

(1877-), which gives the magazine a print of around 20.000 copies. Meanwhile, it is trying to increase its online presence on https://de-gids.nl/.

2.6.5 Literature on *De Gids* and cultural-historical periodicals

Research into *De Gids* and its time frame has mainly been done by Remieg Aerts, who recently published his biography on Thorbecke (Aerts, 2018), which I use in Chapter 2.7.1 to give the historical background for a volume of the magazine. The most influential work on *De Gids* has been published by Aerts (1997) on the interplay of the magazine, its editors and contributors, and their engagement with the liberalist movement until the end of the nineteenth century. In this work, he explains how *De Gids* could have become a popular, nation-wide cultural magazine by holding on to the ideas of liberalism and individualism, which could be seen in the ideas of its founders, its wide range of contributors and the articles that were published by the many editors that *De Gids* has had in this time period. This is very much related to what is written in the significant publication on the history of *De Gids* that was published by Aerts et al. (1987) in association with Meulenhoff publishers to commemorate the 150th anniversary of the magazine in 1987. Further publications about *De Gids* have mainly been done in the periodical itself and recur in several anniversary years.

Multiple scholars look at this periodical in their research, but such a self-contained work as *De Letterheren*, the work by Remieg Aerts (1997), which discusses the involvement in and influence of liberalism in the magazine in the nineteenth century, and which also includes the recent developments of *De Gids*, has not yet been published. Nowadays, the corpus of *De Gids* serves as an interesting and valuable corpus in the Humanities, which is shown by the number of people using it in their unpublished experiments and research.[43] And although this thesis is just an explorative study, developments in the Digital Humanities could have a role to play here.

2.7 Historical background for each of the case studies

In the three sections below the historical background for each of the three case studies in this thesis is presented. These include information deemed necessary to interpret the topics from the topic model, or provides information on existing literature about the aspects that are analysed in the results section in Chapter 4. They furthermore sketch a historical background for the constitutional reforms in Europe and The Netherlands in particular in 1848, give an overview of the Dreyfus affair in France and in Dutch media from 1894-1906, and it discusses how and which academic disciplines coexist in the first 100 years of *De Gids*.

[43] This follows from conference contributions and corridor chats.

2.7.1 Case study 1: Constitutional reforms in 1848

An earlier attempt in 1844 by nine members of the Dutch parliament, the liberal *Negenmannen*, to renew the Dutch government and its constitution, which was slightly reformed in 1840, was shot down on the premise that only the Dutch monarch could proclaim reforms (Te Velde, 2013, pp. 112-113). In 1847 and 1848, the Dutch king, under pressure of the ongoing revolts in Europe and The Netherlands, finally decided to allow, among others, Johan Rudolf Thorbecke (1798-1872), who was one of the *Negenmannen*, to formulate a revision to the current constitution (Aerts, 2018, p. 355). Once implemented, members of the Dutch House of Representatives were to be chosen directly, and the Dutch Senate would be appointed by the States-Provincial. The independent political position of the king was traded in for a system of ministerial responsibility (Te Velde, 2013, p. 115). Other proposed changes were meant to elucidate existing decrees on colonial finances and the rights of the parliament, but also included "functional" civilian rights, such as the freedom of press and freedom of religion (Aerts, 2018, p. 384).

1848 was a politically tumultuous year not only in The Netherlands, but in all of Europe. The constitutional reform of 1848 would have been unthinkable without the then administered political liberal reforms in France and the German states as a consequence of the *Révolution de Février*[44] and the Berlin *Märzrevolution*[45] respectively (Waling, 2016). To anticipate to domestic revolts and to prevent an uprising as was seen in these other Revolutions, Willem II agreed to a constitutional reform, while being urged by members of the liberal movement to fulfil the promises he made in his kings speech of 1847 (Aerts, 2018, pp. 355-360). This eventually led to the publication of the new constitution in November 1848.

The parliamentary reform of 1848 can altogether be seen as the basis for the current parliamentary democracy in The Netherlands, and a step towards a participatory democracy, which, to a great extent, can be credited to the efforts of Thorbecke. This was also the start of a change in politics in The Netherlands on which the liberal movement as driven by Thorbecke set its mark. It implemented reforms which further secularised the government, and had its influence on the economy (Te Velde, 2013, p. 117). It took another full year for the first liberal cabinet to be formed that could work with the new freedoms that were created in the constitution: Thorbecke I, 1849-1853.

Themes regarding this constitutional change and ideas on liberalism can be expected in this volume of *De Gids*. Several frontrunners of the liberal movement, such as

[44] The revolution in February 1948 in Paris during which king Louis-Philippe I was dethroned, which ended the French monarchy and installed the French Second Republic. This was the start for many democratic and social reforms, such as universal suffrage (for men) and free press (Waling, 2016, pp. 7-11).

[45] Influenced by the revolution in Paris, the Prussian king Friedrich Wilhelm IV was forced to implement liberal and democratic reforms in Berlin. The king was allowed to stay, but his power was severely reduced (ibid., pp. 13-22).

Gerrit de Clercq, Johannes Kneppelhout and J.E. Goudsmit, were also contributors to *De Gids*, though not all in 1848. What did get published in this year as the first article of the first issue was a review by Thorbecke, in which he discussed the role of the French in the modernisation of the Dutch state, which formed the preamble to his ideas on a new constitution,[46] and a vision on the reform and taxes by J.J. Rovère van Breugel.[47] An essay on the proposed idea of direct elections by De Clercq can also be found in this volume,[48] along with contributions that applauded the reforms initiated by S.J. van den Bergh[49] and J.P. Heije[50] (Aerts, 1987f, pp. 28-29, 32). The rest of the articles are devoted to the standard range of themes for *De Gids*: art and science.

More information on what happened in 1848 (experiment 1, Section 3.6.1) can be found in Te Velde (2013) and Waling (2016). For the time period 1894-1906, the time period of the Dreyfus affair (experiment 2, Section 3.6.2, information can be found in Read (2012), Wesseling (1987) and the biography of Zola by Mitterand (2002). These sources all describe an aspect of the affair, such as the historical background, the attention it got in France and abroad, and the engagement in literary circles. Knowing this hopefully helps to with the analysis and interpretation of the model's outcome.

2.7.2 Case study 2: The Dreyfus affair (1894-1906)

When, in 1894, an incriminating note was discovered by French military intelligence in the German embassy in Paris, the French and Jewish[51] military officer Alfred Dreyfus (1859-1935) was found guilty of high treason. The found *bordereau* contained information that only an officer that has access to documents from the General Staff could know. Moreover, the author of the note should have been an artillery officer, as was Dreyfus, given the character of the information (Read, 2012, pp. 179-185). Without proper evidence of him passing secret information to German intelligence, Dreyfus was stripped of his honour and military rank and was convicted and transported to Devil's Island, a French penal colony off the coast of French Guiana (ibid., p. 325). Dreyfus's conviction had all the hallmarks of being influenced by anti-Semitic publications in the media, as well as political scheming. Since France had lost the last

[46] Thorbecke, J.R., "Karel Hendrik Ver Huell en Rutger Jan Schimmelpenninck.", *De Gids* (12) 1848 1-35. <_gid001184801_01_0001>

[47] Rovère van Breugel, J.J., "De Grondwets-herziening en ons Belastingstelsel.", *De Gids* (12) 1848 320-362. <_gid001184801_01_0054>

[48] De Clercq, G.,"Twee Nieuwe Bijdragen tot het vóór en tegen der regtstreeksche verkiezingen .", *De Gids* (12) 1848 764-778. <_gid001184801_01_0039>

[49] Van den Bergh, S.J., "Koning en Volk.", *De Gids* (12) 1848 478-479. <_gid001184801_01_0024>

[50] Heije, J.P., "Twee Gedichten van Piet Bogcheljoen", *De Gids* (12) 1848 474-477. <_gid001184801_01_0023>

[51] He was not only Jewish, which would work against him, but he was also born in the Alsace, which was part of France at the time of his birth, but had become German after the French-German war of 1870. Him being an Alsatian automatically made him a potential spy in the light of those days (Read, 2012, p. 162).

confrontation in the German-French war of 1870-1871, anti-German and nationalist feelings were omnipresent, besides the already present anti-Jewish sentiments. In Dreyfus, the military and the public had found the ideal scapegoat.

Dreyfuss case and fate had almost passed in silence. However, after more than a year since he was isolated on the *Île du Diable* in March 1895, in September 1896 the French newspaper *L'Eclair* disclosed the existence of secret information that was kept back during the trial two years earlier (Read, 2012, pp. 494-496).[52] In the meantime Dreyfus's brother Mathieu was working together with, among others, the Jewish publicist and writer Bernard-Lazare to trace the real perpetrator and to free his brother. It would take another full year for anything to happen to the case.

In this following year, it became more and more evident that Dreyfus's trial was based on nothing but rumours, false accusations, false evidence, and attempts by the military establishment to save their face. Even a forged letter (the later called *faux-Henry*) is produced to prove Dreyfuss guilt, and to show that the trial did pass correctly (ibid., pp. 504-505). Georges Picquart, meanwhile appointed as chief of the army intelligence, discovered in the course of 1896 that the *bordereau* that was seen as the core evidence of Dreyfuss conviction, was, in fact, written by the French Hungarian officer Esterházy. In his attempts to redress Dreyfuss situation, Picquart found himself expelled and awaiting a trial (ibid., p. 612).

Émile Zola (1840-1902), influenced by Dreyfuss brother, Picquart's lawyer, and publicist and critic Bernard-Lazare, decided to join the growing group of people that advocated the reopening of the case against Alfred Dreyfus. Although hesitant to interfere with the course of events in the first place, he publishes his first of a series of texts on the matter in November 1897 in *Le Figaro*, which, a while later, would lead to his famous public allegation (Mitterand, 2002, p. 339). In this article, titled *M. Scheurer-Kestner*, named after the vice president of the French senate, Zola urges the senator to reveal the name of the real delinquent and asks the French government and military to admit and correct their mistake of this miscarriage of justice. Lastly, he speaks out against the anti-Semitics and the anti-Semitic media.[53] It was known at the time, in the circle around Scheurer-Kestner, that he wanted to address the questionable proceedings in the Dreyfus case, or at least that he was a *dreyfusard* (Read, 2012, pp. 521-533). Addressing this pledge to a sympathiser in such a high position would most likely progress the case.

It was January 13, 1898, when Zola published his pamphlet *J'accuse*, directed to the president of France, Félix Faure, on the front page of the French newspaper *L'Aurore*, in which he condemns the key players at the Ministry of War for their role in the affair (Mitterand, 2002, pp. 373-379; Zola, 2001, pp. 83-100). Two days earlier, a

[52] A few days earlier however, a false story was published in the (British) papers of an alleged escape of Dreyfus from Devil's Island to raise the interest for the case (Read, 2012, pp. 473-474).

[53] Zola had already published on this theme in his in 1896 in *Le Figaro* published article *Pour les juifs* (Mitterand, 2002, p. 311).

military trial had rejected the evidence that was pointing towards Estherázy as the perpetrator, which possibly could have led to the acquittal of Dreyfus. That the letter immediately led to a lot of commotion is seen in the following anti-Semitic revolts (Read, 2012, pp. 644-648), but also in the endorsements on the letter that appeared immediately afterwards. On the same spot in *L'Aurore* the next day, a small text was published, declaring the support of several "intellectual" signatories for what had become Zola's case (Wesseling, 1987, pp. 89-92). In the following days, the newspaper kept publishing these lists of supporters or *Dreyfusards*, which, in the end, accumulated the total number of listed supporters to more than 3,000. This and following publications by Zola[54], and the letters of support, however, had little effect, other than that Zola was charged and convicted several times for dishonouring the army. In July 1898, he decided to go and live in exile in England (Mitterand, 2002, pp. 492-495).

A newly instated government in France finally opened up the case again. A closer look at the evidence that Picquart had collected led to the discovery in August 1898 that the evidence that was kept back in the first Dreyfus trial, the *faux-Henry*, turned out to be fake (Read, 2012, pp. 740-742). Several other attempts to get to the bottom of the case by a change of guards in politics and jurisdiction eventually led to the nullification of Dreyfus's verdict in June 1899. He was allowed to return as "just a prisoner" to continental France for the reopening of his case in Rennes in August 1899 (ibid., pp. 826-834). This second court-martial attracted many Dutch spectators that had been watching the case closely (ibid., p. 838). Just before the case was reopened Esterházy (residing in England) admitted in a letter that was leaked to the press that he was the author of the *bordereau* (ibid., pp. 842-843), so it was against the odds that Dreyfus was convicted guilty for a second time on 9 September. Ten days later he was pardoned by the French government, which did so in an attempt to prevent a loss of prestige, given the international attention the martial court had received and the *L'Exposition de Paris 1900* (ibid., pp. 899-907). This did not mean he was acquitted from any conviction: it would take several trials and scandals in France until he was fully absolved from the charges of 1894, which finally happened in 1906.

Zola had returned from his self-imposed exile to Paris in June 1899, after which he continued to write on the case, even after Dreyfus was pardoned (Mitterand, 2002, pp. 619, 643). He published his final letter on the case *Lettre à M. Émile Loubet* in December 1900 in *L'Aurore*. Its structure is similar to *J'accuse*, but he now urged president Loubet to put a halt to his trials and to properly acquit Dreyfus (Zola, 2001, pp. 192-211). Zola died in 1902 from coal-fume poisoning. It was believed to be an accident at the time, but there is speculation that he was murdered on political motives and that his chimney was blocked on purpose, although evidence for this is rather scarce (Mitterand, 2002, p. 813).

[54] Zola later bundled his writing on the affair in his work *La Vérité en marche* (1901) (Zola, 2001).

The Dreyfus affair can be seen as the first time in France that a group of "intellectuals" collectively took a stand in a course of events and were "engaged" with what was threatened the most at this time: the fundamentals of scholarship, knowledge and a quest for the truth (Wesseling, 1987, p. 88). They acted from their beliefs in correct (scholarly) procedures, which do not allow fiddling with truth, justice, and objectivity. From this came the urge to support Zola during his trial in 1898 and the case of Dreyfus, mostly not out of political motives, but out of engagement. (ibid., pp. 98-100, 106-107).

2.7.3 Attention in The Netherlands

The engagement with the affair was also seen outside of France. H.L. Wesseling (ibid., 92-93, footnote 17) mentions that amongst the signees of the declarations in *L'Aurore* were around 2,300 Frenchmen. The other 750 names were foreign. This already gives an idea of the international appeal of the affair when Zola stepped in. Except for mentioning that "76 women from Amsterdam" signed the notes in the newspaper, he did not mention any further Dutch involvement with the affair. To which extent publicists in The Netherlands were engaged with the matter was investigated by Büch (1983), who signals a lack of publications on the affair, especially when compared to its (known) international attention. He finds publications by politician and chief-editor of *De Kroniek* P.L. Tak., writing by Frederik van Eeden and Lodewijk van Deyssel, which were published after the affair had landed, and Dutch translations of the letters written by Zola. (ibid., pp. 7-8). Büch concludes that there was more attention for the person of Zola and the Zola trials than for Dreyfus when it comes to the few Dutch publications and pamphlets he could find (ibid., pp. 9-12). One of these, published in 1898, mentions a high interest in the Dreyfus case and the Zola trials in The Netherlands, as seen in the number of daily publications in the newspapers, but it is uncertain how accurate this claim is.

However, it can be concluded from searching Delpher[55] that almost 4,900 newspaper articles (i.e. articles, advertisements, other) and around 270 articles from periodicals can be found when searching on "Dreyfus and Zola" in the time period of the affair.[56] These all mention the two search terms in the same entry. These numbers do indicate that there was attention for the case in The Netherlands, but the findings by Büch (ibid.) also suggest that this attention merely came from journalists, and not so much from Dutch scholars, literary figures or other "intellectuals." Oddly, he does not mention the play *Dreyfus, de martelaar van het Duivelseiland* by Anton van Sprinkhuysen that was staged in Amsterdam in December 1897, even before

[55] https://www.delpher.nl/

[56] These numbers have to be seen in perspective to the total number of journals and newspapers that are digitally available in the KB (and of course to what is not available).

Zola published his *J'Accuse* (1898). The Theaterencyclopedie[57] writes that this play was well received in the press and that it was considered for foreign (even French) translations (Theaterencyclopedie contributors, 2018). Although Dutch newspapers and the audience were predominantly, if not all, in favour of Dreyfus and the *dreyfusards*, French right-wing newspapers were also publishing on the Dutch performance, which made diplomatists interdict the staging of the play in The Hague and Rotterdam. Van Sprinkhuysen altered the play after the second Dreyfus trial in 1899, two year after its premiere, to accommodate the change of events in France. Shortly after, the attention for the affair was lost in the mire on the beginning of the Second Anglo-Boer War.

The staging of this play, the fact that it had sold out 150 times in its first-year (ibid.), and the fact that it was altered in accordance with the events in France, suggests that the affair was being watched very closely in The Netherlands. It also found its way into other art forms. The affair was mentioned in passing by Couperus in his *De Stille Kracht* [The Hidden Force] (1900), which was also published in *De Gids* as feuilleton (Bel, 2015, p. 33), and Albert Verwey also mentions the affair in a sonnet (ibid., pp. 61, 153). Although attention for the affair and anti-Semitism rests on publications on Jewish life and literature in The Netherlands, there is no clear evidence of extra attention for the case of Dreyfus within a group of Jewish authors that were writing on "Jewish themes" around 1900 (ibid., pp. 152-155). Their novels or publications in periodicals were discussed as normal, including in *De Gids*, without anti-Semitic critique in this time period.[58]

It can thus be concluded that the Dreyfus case was a hot topic in the press, but only received a little bit of attention by authors and writers in periodicals and literature.[59] Specific contributions on the case made in *De Gids* are not named in the above-mentioned sources, but some of its (regular) contributors are.[60] Based on this information on the attention the affair received in The Netherlands, it can be expected that it received attention in *De Gids* at the time as well. The goal of this second experiment is to find documents or parts of documents from the Dreyfus affair time period (1894-1906) in the magazine that mention or refer to the affair in some form. Hopefully, topics directly related to this matter can be discovered, although these might be hard to find. More subtle references could maybe be found in

[57] Largely based on the enthusiast digging in the newspaper archives of the KB by user F. Hers d.d. 31 March 2018. See: `https://theaterencyclopedie.nl/wiki/Anton_van_Sprinkhuysen`.

[58] An exception is the anti-Semitic review of Jacob Israël de Haan's *Het Joodsche lied* by Carel Scharten, but this was in 1915 and not in *De Gids* (Bel, 2015, pp. 156-157).

[59] This also follows from a newspaper article by Henriëtte Boas from 1998 in *Trouw* who mentions that she was not able to find any writing on the affair by leading Dutch authors and politicians at the time (Boas, 1998).

[60] An article on socialism and anti-Semitism in the fin-de-siècle by Stutje (2014) discusses Domela Nieuwenhuis's ambivalence towards the case of Dreyfus. The play as was described above is mentioned, but also the media that frequently published on the matter: the *Algemeen Handelsblad* reported extensively on the case with articles by its correspondent in Paris, Louis Israëls (ibid., p. 24). This newspaper's liberal tenet is comparable to that of *De Gids* (Aerts, 1997, p. 129), which could roughly be extended to its audience.

themes/topics on Judaism, anti-Semitism, or Zola. This should, if these do appear, return more results than when one ordinarily searches for key terms from the event. Knowing the course of events as described above helps in identifying these topics.

2.7.4 Case study 3: A disciplinary overview of the first 100 years of *De Gids*

Literature that is used in the final experiment (experiment 3, Section 3.6.3) is connected to studies on the history of disciplines in The Netherlands and Europe, and touches upon the studies that are focussing on signalling historical trends, of which the above-mentioned studies on topic modelling by Newman and Block (2006), Hall, Daniel Jurafsky, and Manning (2008) and Goldstone and Underwood (2014) feature a cultural-historic component.

The nineteenth and early twentieth century are, in recent European oriented research into the history of science and cultural history, seen as the age of increasing academisation, disciplinarisation and knowledge popularisation (or knowledge dissemination). Research was centred in and around universities, and different research areas were transformed into self-organised disciplines, here seen as a restricted and specialised field of research, each with its own community, publication media and shared culture and identity (Abbott, 2010, pp. 122-144). There were frequent publications about the newly acquisitioned knowledge and the developed knowledge infrastructure in various newspapers, magazines, and other media meant for a general audience. This stream of information was not limited to one certain field, but included an information stream from both the sciences and the arts:

> It was neither in the 18th-century nor in the present that popular science reached its heyday, but in the 19th century. This is due not so much to the amount of popular literature produced and read in that century, but to the significance of popularization for the self-image of the period. The natural sciences were considered to be the motive force of progress in all areas of social life; whoever wanted to be 'up with the times' had to be familiar with their success and method of thought (Bayertz, 1985, p. 209).

These major transitions are represented in the general cultural and authoritative Dutch magazine *De Gids*, which is evident from the research of Remieg Aerts (1997) and Klaas van Berkel (1998). Firstly, new disciplinary approaches in the magazines of this time were introduced and characterised from a solid international European perspective (e.g. stereochemistry, palaeontology, astronomy, and a range of theories of evolution, including the, at that time, highly controversial ideas of Charles Darwin). Readers were informed about the groundbreaking work of various specialists or specialised knowledge communities of which they are part (Peperkamp, 2004).

And secondly, the magazine provided a scope for public - highly normative - evaluations of these changes in the knowledge landscape, especially when established knowledge from the fields of the traditional humanities (e.g. theology, literature, philology, linguistics) was being confronted with newly acquired - often empirically grounded - insights from the natural sciences. The heated discussions about the meaning of the so-called "modern theology" (Fyfe, 2008, p. 121) which puts the authority of the "Word of God" in a different perspective (cf. "Conflict Thesis" or "Draper-White thesis" (C. A. Russell, 2003)), and the origin of species can exist side-by-side in this dynamic context.

Research into the dissemination and popularisation of knowledge in the nineteenth and early twentieth century in the Netherlands and Europe is scarce, and is often focused on the organisation and autonomisation of a single or a few discipline(s) (e.g. physics (Van Berkel, van Helden, and Palm, 1999)). It therefore pays no attention to the dynamics of the scientific field as a whole in this age period. As a result, sensitive discussions about the relative and changing mutual authority of disciplines have remained underrepresented.

It is assumed that the transitions in the academic landscape are, among others, reflected in the contributions to *De Gids*, as can be seen in the literature by Aerts (1997) and Van Berkel (1998). Claims about the relation between publications from the two cultures[61] have so far been formulated on the basis of intuition, and a full overview of the distribution between arts and sciences is lacking. Although it is not within the reach of this thesis to give a full analysis of the disciplinary or academic landscape in the nineteenth century, it is possible to make a first step in analysing the interplay between the traditional field of the arts and the newly developed fields from the natural sciences, by looking at a periodical from this time: *De Gids*.

In the description of the history of the periodical in Chapter 2.6 it already becomes clear that the magazine knew different compositions of the editorial board and that these boards each left their mark on the theme of the magazine's contents. Aerts (1997, pp. 449-450) writes that the main focus of *De Gids* was no longer on historical matter from 1865 onwards, but instead on developments and ways of thinking in the sciences. This increased scientific approach could be seen in how publications on ethics, societal issues, and religion were influenced by this new "scientific reasoning". He also writes that two separate worldviews circulated in *De Gids* around 1870: one that was tied to a firm belief in the natural sciences, its logic and its understanding of nature and the human, and the other, at that time the favoured movement, was leaving more room for a spiritual and ethical worldview, aiming at individual freedom and responsibility (Aerts, 1987b, p. 64). The main focus of this "general cultural magazine" had always been on culture and arts, and publications on or reviews

[61] I.e. Arts and Sciences, to speak in terms of C.P. Snow (1961) and his famous Rede lecture from 1959.

about developments in the natural sciences were always written from a societal analytical view. They investigated what impact the scientific developments had in daily life, on culture and the arts (Aerts, 1997, p. 454).

It is this interplay between the arts and sciences that fluctuates in the first 100 years of *De Gids* (Aerts et al., 1987), and after.[62] Not only does the distribution of contributions on the arts and sciences change during this timespan, but the magazine also had to cope with newly formed disciplines, a change in religious beliefs, and scientific discoveries in this era. In the first two decades of *De Gids*, contributions were written with a traditional Christian view in mind. The editors tried to facilitate and defend traditional Christian beliefs in a scientific way (also influenced by views from the Réveil), initially retorting to radical Textualism and speculative philosophy coming from Germany,[63] but as time passed they were allowing more and more other (Christian) interpretations and views in their magazine (Aerts, 1997, pp. 239-242). From the 1850s onwards, their traditional beliefs were also endangered by developments from the natural sciences that cause a shift from supranaturalist to naturalist worldviews. This signals the start of the discussions around "modern theology" that were also held in *De Gids*.

Natural sciences gradually became of more importance in *De Gids*, but until the 1850s, contributions on this matter always had to be accompanied by explanatory writing on its relation to a metaphysical worldview (ibid., pp. 243-245). It is in this period that authors such as psychiatrist J.N. Ramaer[64] (1817-1887) and physiologist F.C. Donders[65] (1818-1889) already wrote about the independence of the natural sciences, presented as free from any theological matter. Publications that directly interfered with religious matter were considered too radical for the public. Similarly, a refreshed view of religion was published, in which the old traditional perspective had to make room for this modern theology (ibid., pp. 247-249). From 1860 onwards, publications on the natural sciences in *De Gids* were no longer constrained by religious influence, provided that they were not too radical and that they were not interfering with speculative claims on religion (ibid., pp. 251-253). However, in practice, scientific publications were still accompanied by accounts on the field of physicotheology and other religious reading. In 1862, the editors of the magazine decided to minimalise the number of publications on religion, which lead to criticism from several authors belonging to various religious movements.

The number of scientific texts in *De Gids* had increased in the last 25 years of the nineteenth century, and scientific knowledge and its way of thinking was now also enforced in the field of ethics and reasoning, which traditionally had been in the

[62] It is also worth noting, although this is not covered by this thesis' corpus, that the magazine has now (anno 2018) been transformed into an entirely cultural magazine.

[63] Introduced in the Netherlands and in *De Gids* by publications of among others C.W. Opzoomer (1821-1892) and Van Limburg Brouwer (1829-1873).

[64] Ramaer, J.N., 'De Wetenschap der Natuur', *De Gids* (12) 1848 1-34. <_gid001184801_01_0042>

[65] Donders, F.C. 'Natuurkunde van den mensch', *De Gids* (10) 1846 750-782 <_gid001184601_01_0033>, 863-900 <_gid001184601_01_0038>.

domain of philosophy and theology (Aerts, 1997, p. 450). The magazine's contents eventually reflected societal interest in scientific fields, such as for biology at the end of the century. However, from its inception the magazine already featured articles from all disciplines and fields, varying from highly specialist reviews before 1848, to texts on scientific developments meant to educate commoners (ibid., pp. 452-453). This was the case until the 1880s when the *Bibliographisch Album* section in every issue, which featured reviews on all kinds of publications from the arts and sciences, was replaced by the *Letterkundige Kroniek*. The number of reviews and publications on subjects other than health and biology had already shrunk to a minimum. It seemed that the magazine averted itself from writing on the sciences, but it was also the magazine's custom to honour and commemorate previous authors and other leading scientists with extensive essays. Thus, in one way or another did these subjects find their way into the periodical (ibid., pp. 453-454).

After 1865, *De Gids* played a large role in the discussions that were held amongst various religious movements, and functioned as a publication platform for texts on the introduction of a "modern theology," as well as the role of theology in the academical field (ibid., p. 457). Because of its liberalist view, the magazine later acted as a medium for publications on other religions and comparative theology, which contributed to the creation of theology as academic discipline (ibid., pp. 461-462). Nevertheless, there was an increasing struggle between the scientific and religious worldviews, which was also visible in *De Gids*. This could be seen in hereto related philosophical writing on materialism and positivism. With respect to the discussions mentioned earlier around Darwinism, the theories of Darwin, together with their reception in German and English science, were discussed in the magazine by zoologist T.C. Winkler[66] (1822-1897) in a very careful manner, to not interfere with religious beliefs (ibid., pp. 472-473). The theories were quickly adopted, but also found opposition in the following years.[67] These publications appeared during a trend of further increasing interest in science, at the expense of writing on religion and philosophy. Metaphysics was later ejected to the personal domain and it appeared that science now also took the place of philosophy and religion when it came to ethics and morality (ibid., p. 483). This is in line with the interest in biology, and later psychology.

This is where *De letterheren* stops its description of science in *De Gids*. A recent special issue of *De Gids* by Saris and Visser (2005) offers some highlights from 100 years (1900-2000) of scientific contributions to the magazine, and notes that just 8% of the total contributions in the twentieth century of the magazine is devoted to the sciences. Its attention mainly goes to physics, but not so much to biology. The editors of this magazine state that the subjects that are discussed in *De Gids* are not representative of the scientific domain in the previous century, but that they do give an

[66] Winkler, T.C., "De leer van Darwin.", *De Gids* (31) 1867 22-70. <_gid001186701_01_0063>

[67] For example, in publications such as Spruyt, C.B., "Natuurkundige phantasieën", *De Gids* (38) 1874 401-440. <_gid001187401_01_0012>

overview of the interest there was in science in Dutch society and culture (Saris and Visser, 2005, p. 217). Another contribution in this issue mentions that *De Gids* up until the Second World War mirrored the Dutch scientific developments, and was not European oriented (Theunissen, 2005, p. 219). This can be explained by the fact that the early twentieth century can be seen as the Second Dutch Golden Age, which was characterised by the many scientific developments by Dutch scientists, which is signified by the many Nobel prize laureates.[68]

It can thus be summarised that, certainly in its early days (after 1848), publications in *De Gids* include themes from both the sciences and the arts. Subjects related to the Humanities were, on the other hand, severely influenced by the new way of thinking that was introduced by the sciences, as is illustrated by Krop (1994). These two fields existed side-by-side, and attempts were made to unite the two fields. Especially the religious debates around modern theology got a place in *De Gids*. Scientific contributions in the early twentieth century were mainly from the field of physics. It is, however, not clear how this new way of thinking interacted with disciplines from the arts, and if it even did: did the introduction of a new science/discipline, for instance, cause a drop in contributions on other topics?

These are just a few examples of the tensions between various disciplines that co-exist in *De Gids*. The science versus religion is an interesting example, but so is the more general sciences versus arts. To shed light on this distribution from a quantitative perspective, I will therefore try to assign disciplinary labels (e.g. physics, literary studies, politics)[69] to the topics from the topic model created. These labels can, in turn, be attributed to either arts or (natural) sciences. By subsequently comparing the number of articles with contributions from one of these areas on a temporal scale, it becomes clear how the academic landscape has changed over time, how the introduction of the natural sciences proceeded, and whether these two fields interacted. The results can be interpreted in the light of existing intuitive and individual statements about the distribution of alpha and beta disciplines at the end of the nineteenth and the beginning of the twentieth century in *De Gids* and may shed new light on the knowledge of the history of science in the Netherlands.[70]

[68] See for a list of laureates from this time period: `https://www.knaw.nl/en/about-us/academy-history/1902-de-tweede-gouden-eeuw`.

[69] One could think of these categories as fields or studies that could exist and be taught at a university or could have their own lemma in a dictionary or in an encyclopaedia. In this thesis, this is done by using the categories as they are known nowadays, using their twenty-first-century definition. To some, this could seem problematic, since this would mean that I am perhaps attaching labels to (topics that are made up from) documents that were written in a time at which these disciplines had a less clear shape or had not yet been formed (i.e that it did not exist with that name at e.g. a university, abiding the above-stated definition). Considering this case study from a new-historicist view is, in my opinion, more constructive and more fruitful, and better suits the technique that is used.

[70] For example, in the case of Darwinism, identifying such a theme in a topic would be very easy, but it would then be very difficult to also distinguish the article's polarisation: whether it is in favour or against the idea of Darwinism. This is not within the scope of this thesis.

Chapter 3

Method

This thesis will implement a standard LDA approach. Following the recommendations from the previous chapter and the words of caution from, among others, Underwood (2014) (Section 2.5.4), the LDA-algorithm is used fully unsupervised, as it was designed by Blei, Ng, and Jordan (2003) and later improved by Hoffman, Bach, and Blei (2010). Thereafter, a pipeline for modifying the corpus, and creating and running the topic model onto this corpus is constructed. This pipeline pre-processes a slice of the corpus according to a requested time span, prepares the documents for analysis, and builds a topic model that includes topic-word and document-topic distributions. Following this, results can be created that shed light on different slices of *De Gids*.

Statistics about the corpus and its condition are described below (see Section 3.1.2), after which a section is devoted to the problems that were observed while working with the corpus (Section 3.1.3). Despite it being digital, the corpus cannot directly be used in the algorithm and this specific LDA implementation as is. Some pre-processing steps are still required in order for it to be analysed by the algorithm and to receive an optimal result. These steps include cleaning the corpus of unwanted artefacts and the removal of stop words, while following the recommendations of some studies described in the previous chapter. For these pre-processing steps, see Section 3.2.

The specific implementation of the LDA algorithm that is used in the subsequent step of constructing and training a model comes from the popular Python software package and tool suite Gensim (Rehurek and Sojka, 2010). Results from the model are handled in a Python environment and are serialised (i.e. saved to a file) where needed for future inspection and reproducibility. The models, as well as the code that is written for this pipeline, are publicly available in a GitLab repository.[1] This includes the wrapper for handling the format in which the corpus is stored. This also means that anyone is invited to use, modify and build upon this code in his or her own experiments. This should be relatively easy and especially useful for documents that are stored in the FoLiA format (Van Gompel and Reynaert, 2013), as

[1] See: `https://gitlab.com/LvanWissen/degids`

is the case with the corpus used in this thesis, if one wants to take advantage of the coded tags and linguistic information in this annotation format. The repository can be found at: `https://gitlab.com/LvanWissen/degids`.

Finally, the experimental settings for each of the three experiments are described at the end of this chapter (Section 3.6). These descriptions include a demarcation of the corpus and an explanation about why this specific time frame and corresponding volumes are interesting and important for the experiment. This complements the historical background that was given in Chapter 2.7. Each of these descriptions includes a section on its evaluation, the best way to present its results (e.g. through word clouds, graphs or other), and the expected output. Results of the experiments are then given in their respective paragraphs in Chapter 4.

3.1 Corpus of *De Gids* (1837-1936)

3.1.1 Statistics

The corpus that is used in this thesis is consists of a total of 10.624 publications of *De Gids*, from the timespan 1837 (starting year) - 1936. This is a subset (the first 100 volumes) of the entire corpus that is made available by the KB.[2] This number of documents includes all separate articles (e.g. editorial notes, news, letters, opinions, serials), as they appear online, in unique entries in the DBNL (Van Stipriaan, 2009). An overview of the number of contributions per year is given in Figure 3.1 and Table A.1 in the appendix in which a steep decline in articles per volume can be observed in the first 20 years of the magazine, to a number of 52 articles in 1855 and 1856. The number of contributions then gradually rises to a peak of 177 in 1910, after which it remains around 130 until 1936. The average length of these articles also fluctuates, but appears to be correlated to the number of articles that are published in one year. The most comprehensive volumes were published from the 1860s onwards, until the 1880s. A complete table with corpus statistics can be found in Appendix A. This table also shows the total number of pages per year, as well as the number of words per year, together with information on the length of the shortest and longest article. This should give an indication of the size of an issue of *De Gids*. In total, this corpus of 100 years of *De Gids* includes 63,814 pages that embody 69,430,134 words[3] in a total of 10,624 articles.

[2] The corpus is available online up to the 2005 volume in a parsed computer-readable format, but, as the magazine is still published, its volumes are still being added to the DBNL. More recent articles can be requested from the website of *De Gids* (`https://de-gids.nl/`), which means that, in theory, every article that has ever been published by *De Gids* could be viewed, but copyrights might be a limiting factor.

[3] The term "word" in the field of linguistics is ambiguous, but is most often synonymous with the term "token." Indicated here and passim, "word" can be seen as a unit of characters separated by a space. Punctuation marks are not counted, but stop words are.

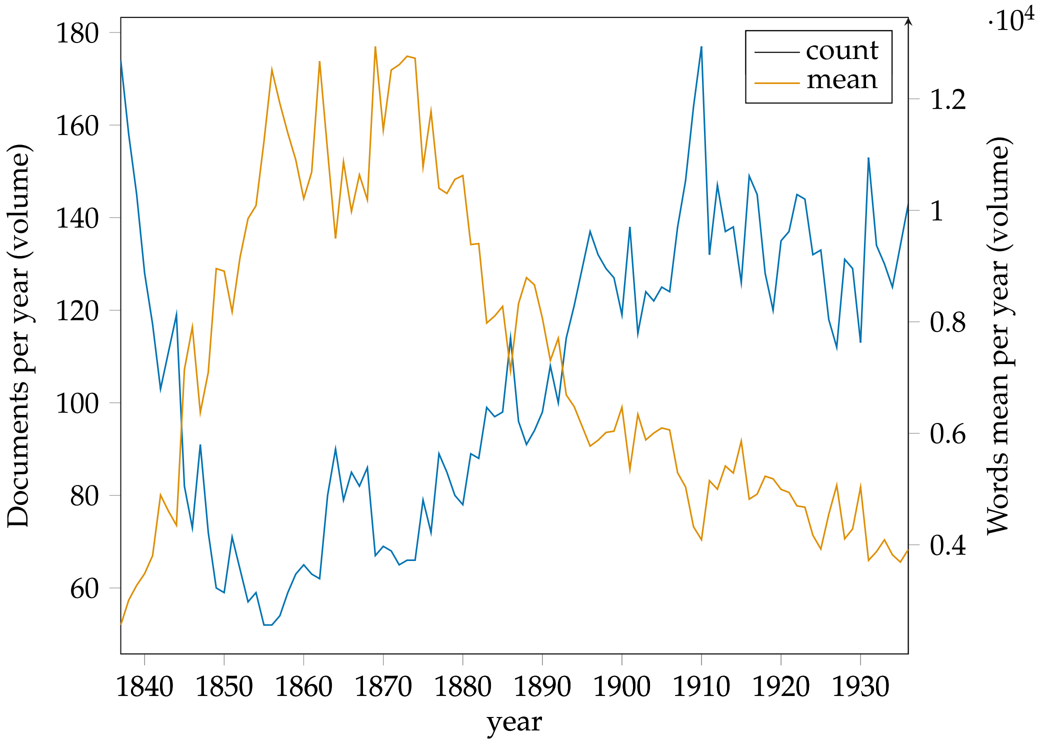


FIGURE 3.1: Number of articles and their average length in word count per volume from the first 100 volumes from *De Gids* (1837-1936).

The difference in length could have had various reasons, ranging from a paper shortage to a lack of contributions that were found suitable by the editors of *De Gids*. What is more, the type of document affects the size: a poem takes relatively more space than a part in a feuilleton. Table A.1 shows that the magazine from 1918 onwards had drastically shrunk in size. This is also seen in in the number of words, but not as much in the number of articles. The articles do, however, become shorter, as indicated by the average number of pages per document.

De Gids and its authors

For 8,406 of these publications, one or multiple authors are listed in the DBNL. The rest was published anonymously or lacks authorship information because this is not annotated: it is not always the case that the author is indicated in the metadata when a contribution is signed with a full name, for example when it concerns a lesser known or occasional author, or when the article is underwritten with initials, such as *B.v.d.B.* which probably refers to *Bakhuizen van den Brink*.[4] The 8,406 publications with an annotated contributor are published by a total of 1,452 unique authors. Colenbrander (476), Jacob Nikolaas van Hall (1814-1918) (346) and poet Hélene Swarth (1859-1941) (169) have by far contributed the most, closely followed

[4] I am leaving the metadata annotations as is. Improving and adding information to the data is not within the scope of this thesis and should be done by the curator of the data.

by Potgieter (167). In reality, the number of contributions by Potgieter to the magazine must have been higher, but due to the magazine's habit of publishing articles anonymously, or due to incomplete metadata, these figures are not higher.

3.1.2 Condition

XML

Every document in the corpus has been processed by Frog (Van den Bosch et al., 2007) and is available in an enriched, machine-readable format: XML. The Frog NLP suite, as implemented in Nederlab (Brugman et al., 2016), has created a FoLiA document (van Gompel and Reynaert, 2013) out of each article in *De Gids*. This FoLiA (Format for Linguistic Annotation) is an XML-based document with several layers that each includes information on the document, its tokens (including PoS information), types (e.g. lemmas), and its entities (e.g. persons, locations, dates). The format is flexible and allows for additions, so that it is easy to incorporate new information into its existing contents. It seems to be the standard for storing Dutch documents and annotations, simply because the number of Dutch documents that are analysed in DH research is not that large, or because of the fact that people are using their own tools.[5]

FoLiA

A shortened example of a FoLiA document can be found in Figure 3.2. Indicated are, from top to bottom: the annotation software that is sequentially used to process a document and add its own information (e.g. token, POS or entity information) to the FoLiA, a text layer with information about the type of element that is described (e.g. `chapter` or `poem` in this case), and, per element, a sentence, token and word layer. The word layer in turn includes detailed POS information, along with the result from a lemmatiser. If applicable, a separate entity layer with entity information (e.g. person, location) is added per sentence. These entities are allowed to span multiple tokens.

metadata

Metadata on the volumes and individual articles is stored separately and is available for each article with information on its title, the page numbers (and thus the relative position of the article in the volume), publication year, volume, and, in most cases, its author with a preferred full name and labels for year of birth and death. The unique identifiers that are used in the metadata refer to the identifiers that are used in the DBNL. The DBNL has categorised the contributions to *De Gids* per year, while the magazine was published on a monthly basis until 1962. It would be hard, but not impossible, to add extra information on the issue number per volume to its metadata, for example by looking at recurring titles that are following a fixed format, or by examining the page numbers of the articles and the size of the magazine. Fortunately, this thesis does not require a granularity of a single volume. Simply

[5] Several other formats exist, such as the format that is used in VU Amsterdams Newsreader project: NAF (NLP Annotation Format, see `https://github.com/newsreader/NAF`). These can include more or less the same kind of information in XML encoded syntax. FoLiA was developed at Radboud University and Tilburg University, and was used in the Nederlab project.

looking at the publications over a year (12 issues) suffices. An example of a metadata entry can be found in Figure 3.3. The type of annotated information should be obvious.

The fact that the corpus is already accessible in a machine-readable and, above all, processed format can be seen as a tremendous advantage. Not only can this format much easier be traversed by a computer when compared to a plain text representation of the articles, but the time-consuming task of correcting OCR-mistakes[6], as well as the annotation of metadata, has already been performed and is therefore not necessary. This means that the corpus quality is relatively good, if not excellent when one considers the fact that it is a historical corpus. There are, however, a few remarks when it comes to the quality of added annotation layers in the FoLiA documents, as will be explained in the next section: 3.1.3.

The FoLiA documents give structure to the contents of an article and allow for specialised lookups. Instead of a representation of the contents of an article as is, as would be featured in a plain text document, the dictionary forms of words are used in the experiments (see section 3.1.3), which is one of the outputs by the Frog toolset. If there were images in the documents, then they are lost in the conversion and processing to FoLiA. The same is true for the original representation on paper, the type area, including typesetting and other markup, and information on which words/-paragraphs were on which pages.[7] This information, except for the division into pages and the inclusion of images, is not available at the DBNL. Luckily, this is not an issue in topic modelling, since the algorithm assumes a Bag of Words representation of the documents in the corpus, which makes representation irrelevant for this technique.

3.1.3 Problems that arise

Topic modelling algorithms benefit from a clear (error-free) homogeneous corpus. The more document-word relation information is available, the better the algorithm is able to uncover a document's hidden structure. It helps if the corpus vocabulary is scaled down and thus generalised. This can be done by removing words that occur frequently in almost every document, since they provide little to no information on word-topic-document relations: if a word is in every document, then the algorithm cannot use this word to distinguish between possible topics. Because the algorithm needs a minimum frequency of two in order to make such a relation, all the hapax legomena can be removed as well. Before the latter is done, it is, in an attempt to save valuable information, worth the effort to generalise the contents of a document as much as possible. Information on word inflexion is thereby of lesser importance.

[6] This was not necessary with *De Gids*, since it was digitised by hand.

[7] This might not seem important when one is working with large bodies of text, but this might be key when other techniques are used or combined with topic modelling, such as visual image recognition or when one wants to analyse concrete poetry, to name an extreme example.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE folia>
<FoLiA xmlns="http://ilk.uvt.nl/folia" xmlns:xlink="http://www.w3.org/1999/xlink" generator="
     libfolia-v0.11" xml:id="_gid001193301_01_0107">
  <metadata type="native">
    <annotations>
      <token-annotation annotator="ucto" annotatortype="auto" datetime="2015-05-23T03:52:37"
           set="tokconfig-nl" />
...
    </annotations>
  </metadata>
  <text xml:id="_gid001193301_01_0107_text">
    <div class="chapter" xml:id="_gid001193301_01_0107_div">
      <head xml:id="_gid001193301_01_0107.TEI.2.text.body.div.head.13625">
        <t>Honger</t>
        <s xml:id="_gid001193301_01_0107.TEI.2.text.body.div.head.13625.s.1">
          <t>Honger</t>
          <w class="WORD" xml:id="_gid001193301_01_0107.TEI.2.text.body.div.head.13625.s.1.w.1">
            <t>Honger</t>
            <pos class="N(soort,ev,basis,zijd,stan)" confidence="0.992385" head="N">
              <feat class="soort" subset="ntype" />
              <feat class="ev" subset="getal" />
              <feat class="basis" subset="graad" />
              <feat class="zijd" subset="genus" />
              <feat class="stan" subset="naamval" />
            </pos>
            <lemma class="honger" />
          </w>
          <entities xml:id="_gid001193301_01_0107.TEI.2.text.body.div.head.13625.s.1.entities.1"
                />
        </s>
      </head>
      <event class="poem" xml:id="_gid001193301_01_0107.TEI.2.text.body.div.lg.13633">
        <t>
          <t-str xml:id="_gid001193301_01_0107.TEI.2.text.body.div.lg.1.13626">
          Een man die een zaag vond</t-str>
...
        </t>
        <s xml:id="_gid001193301_01_0107.TEI.2.text.body.div.lg.13633.s.1">
          <t>Een man die een zaag vond Zaagde een lang, traag brood.</t>
          <w class="WORD" xml:id="_gid001193301_01_0107.TEI.2.text.body.div.lg.13633.s.1.w.1">
            <t>Een</t>
            <pos class="LID(onbep,stan,agr)" confidence="0.981771" head="LID">
              <feat class="onbep" subset="lwtype" />
              <feat class="stan" subset="naamval" />
              <feat class="agr" subset="npagr" />
            </pos>
            <lemma class="een" />
          </w>
...
          <entities xml:id="_gid001193301_01_0107.TEI.2.text.body.div.lg.13633.s.3.entities.1" /
              >
        </s>
      </event>
    </div>
  </text>
</FoLiA>
```

FIGURE 3.2: Example of a FoLiA document in which every single article is saved. In this example, the processed version of document _gid001193301_01_0107 is shown.

```
{
  "NLCore_NLIdentification_versionID": "42ee6cf0-011c-11e4-b0ff-51bcbd7c379f",
  "NLCore_NLIdentification_sourceRef": [
    "_gid001191001_01_0150"
  ],
  "NLCore_NLIdentification_sourceUrl": [
    "http:\/\/www.dbnl.org\/tekst\/_gid001191001_01\/_gid001191001_01_0150.php"
  ],
  "NLDependentTitle_title": "'t Vrindje.",
  "NLDependentTitle_NLPersonRef_personID": [
    "81543e08-011b-11e4-b0ff-51bcbd7c379f"
  ],
  "NLDependentTitle_startPage": 226,
  "NLDependentTitle_endPage": 239,
  "NLTitle_title": "De Gids. Jaargang 74",
  "NLTitle_yearOfPublicationMin": 1910,
  "auteursgegevens": {
    "documents": [
      {
        "NLCore_NLIdentification_nederlabID": "81543e08-011b-11e4-b0ff-51bcbd7c379f",
        "NLPerson_NLPersonName_preferredFullName": "Ina Boudier-Bakker",
        "NLPerson_yearOfBirthLabel": "1875",
        "NLPerson_yearOfDeathLabel": "1966"
      }
    ]
  }
}
```

FIGURE 3.3: Example of the metadata that is available for a single article, in this case, document _gid001191001_01_0150.

Orthography

This means, for example, that word inflexions could be reduced to a single word: its lemma or dictionary form. Ideally, this should also happen for orthographical differences in the corpus. This will not be a serious problem for the smaller experiments that include one or a couple of years of contributions, but it will be for an analysis of the entire corpus, since the notation of certain Dutch words has changed in these 100 years. Unfortunately, the analyser and processing tool in the Nederlab pipeline that created the FoLiA files (Frog) has been trained on modern Dutch (Brugman et al., 2016, p. 1279), which means that it mostly fails to recognise older spelling varieties. Not only are these older varieties not mapped to their newer representations (or vice versa), but it is also impossible to get a lemma for these unknown words, other than a guessed lemma on the basis of syntax rules (e.g. *-en* for plural). Improvements implemented to overcome this issue before analysis in the Nederlab pipeline are mainly aimed at improving the tagging and analysis of seventeenth century or even older Dutch (Tjong Kim Sang et al., 2017). Although the orthography in this corpus of the nineteenth and early-twentieth-century publications can be considered Modern Dutch, there are still differences with Contemporary Dutch.

spelling varieties

Foreign languages

foreign languages

It is the case that contributions to the periodical sometimes contain excerpts and quotes from foreign languages. Especially French and German, but also Greek and Latin are very common in articles in *De Gids*. Because of the way topic modelling algorithms work, these foreign words tend to be grouped together in a cluster, simply because they occur together in the text.[8] This is done by the algorithm on the basis of a single language based characteristic. Word meaning expressed in the clusters, which is a result of the assumption that words that occur together tend to have the same meaning, can therefore not be inferred. This can either be a wanted or an unwanted result of the algorithm, since the clusters of, for instance, Greek or Latin words are perhaps a sign of the occurrence of a "classics" or "juridical" topic. It becomes more difficult when assessing French or German clusters, since it is not immediately clear what topic these clusters indicate in *De Gids*. Because it is very hard to automatically discriminate on the language a word belongs to, and since it might be beneficial for interpreting topics or the theme of a document, the words from foreign languages are not filtered and they are kept in the documents that are presented to the algorithm.

Tokens and lemmas

dictionary format

The document's contents can be further generalised by using lemmas instead of the real textual representations of words, so that words such as *appelen*, *appels*, *appel* are grouped under the same key *appel*. This lemma form can also be seen as the more general dictionary format of a single word. The corpus is already lemmatised by the Frog POS-tagger (Van den Bosch et al., 2007) and it should be noted that this does not always provide a desirable result. Tokens that could be identified as Modern Dutch by the tokeniser, or that contain a certain pattern (e.g. *-en* for plural, *ge-* for a perfect participle) have lemma information with its FoLiA notation. Other words that do not follow such patterns do not. Moreover, the abbreviated *d'* and *l'* in the case of French words and the genitive *'s* in Dutch are still attached to the words in the tokenised document. These affixes can simply be removed by a rule-based filter that is implemented and described under 3.2. The goal of this implementation and use of lemmas is to shrink down the size of the vocabulary of the corpus slice that is used, without the loss of information.

Entities

T.-I. Yang, Torget, and Mihalcea (2011) and Jockers and D. Mimno (2013) used Named Entity Recognition to fine tune the corpus that they used in their model. They either

[8] This phenomenon was also signalled, but disregarded in an analysis by Rhody (2012, pp. 28-29).

```
 1  tweede_kamer (2188)             16  van_den_raad (666)
 2  de_haag (1521)                  17  middellandsche_zee (587)
 3  de_s (1291)                     18  nederlandsch_indie (564)
 4  willem_iii (1277)               19  don_juan (539)
 5  lodewijk_xiv (1086)             20  het_woud (536)
 6  _gravenhage (1061)              21  van_wagner (526)
 7  vereenigde_staat (1029)         22  vereenigde_staten (516)
 8  den_haag (983)                  23  busken_huet (481)
 9  da_costa (960)                  24  van_de_franschen (459)
10  eerste_kamer (907)              25  van_hogendorp (457)
11  de_vries (860)                  26  frederik_hendrik (442)
12  willem_i (790)                  27  karel_v (421)
13  i_blz (786)                     28  dit_tijdschrift (412)
14  victor_hugo (782)               29  hooger_onderwijs (410)
15  de_groot (756)                  30  het_leven (404)
```

FIGURE 3.4: Example of a word list sorted on frequency. It is shown that the first content word appears at index 31. The list consists of the 39 most frequent words in the single article `_gid001183701_01_0002`.

entities

removed entities in their texts based on their class, or added them as single chunks that would survive in the BoW representation. The removed entities were character names from novels that were analysed that would otherwise give false topic information to the model. The chunks were combined with their class (e.g. LOCATION, PERSON) to provide extra interpretative information in the analysis of topics.

In any case, nouns play a large distinctive and interpretative role in finding and defining topics, and so do proper nouns in this corpus. Personal names are often connected with a particular topic. It must be noted that the corpus also contains literary pieces, that contain names such as *Oudijck*, but due to their low document frequency this will not be a problem. It can be expected that much information can be gained when these entities are chunked together to form a single token/word for the algorithm, because multi-word entities would be scattered in a BoW representation: the words *Koning Willem II* will ideally not be scattered, but represented as *Koning_Willem_II*.[9]

The Named Entity Recognition that was used to process the FoLiA files is not perfect and leads to a lot of false entities, some of which can be seen in Figure 3.4. In this list, the 30 most frequent multi-word entities and their frequencies in the entire corpus (1837-1937) are displayed. Most of these entries are desirable, but `de_s`, `i_blz`, and `dit_tijdschrift` at positions 3, 13 and 28 respectively are certainly not.[10] The first experiment is therefore done twice on a slightly different corpus representation. In one, entities are added separately to the documents, because it can be reasoned that entities such as place or personal names provide distinctive topic information. If the entity spans multiple tokens, then the words are glued together with an underscore (_). This results in entities of the sort `tweede_kamer` or `victor_hugo`.

[9] An alternative to this would be to generate n-grams from the corpus. Since entity information is generally available for the documents in the corpus, it has here been decided to use this representation instead.

[10] The entries were most likely triggered by the occurrence of a determiner in front of a noun, influenced by a large part of Dutch surnames (e.g. *De Groot*), or, in the case of the demonstrative pronouns, by the frequency of occurrence.

Using the corpus in which entities are stuck together if they span multiple words is not detrimental for the interpretability of the corpus.[11] In the overall statistics, this measure increases the vocabulary size, but this is not the case for experiment 1, in which this measure actually causes a drop in vocabulary size (see Section 3.6.1). To illustrate whether this vocabulary size drop is a wanted or unwanted implementation, the topic clusters for both corpora are displayed in the first experiment.

3.2 Pre-processing

pre-processing

The quality of the data is, by and large, already sufficient for a topic modelling analysis, but its quality can likely be improved relatively easily by implementing several pre-processing steps. These steps include (partly) solving the orthographical variation in the corpus and filtering unwanted words. To boost the quality of the input data in the model even more, the data is transformed or normalised further to construct an as homogeneously as possible textual corpus, free from corpus artefacts (e.g. punctuation such as ellipses, or hyphenated words due to page breaks). The data is manipulated or altered by different pre-processing steps, which are listed below in their order of execution.

The code in the repository allows for an on the fly implementation of these steps (i.e. processing the files when they are requested), but for the sake of speed, the documents are requested once from the collection of FoLiA documents, pre-processed when loaded, and thereafter serialised into a different, more agile format. This format only includes the words that are input into the topic modelling algorithm (i.e. no words that only appear once in the corpus) and is therefore less suitable for close reading after the topic model has produced its results. The stripped corpus that is created can always be traced back to the original FoLiA documents if this is necessary for this particular purpose.

In Table 3.1 it is indicated to which extent the corpus is affected and altered by the pre-processing steps. The number of words and the vocabulary size is given before and after implementing post filtering measurements (see Section 3.3 for the settings). It seems that the vocabulary size is reduced the most in the `lemmas_stopwords` corpus. Replacing words for their entity representation logically lowers the total number of words in the corpus, but this does not lead to a smaller vocabulary. The difference of more than 200,000 items is quite large and can be explained by the number of faulty or inconsistent entities coming from the FoLiA files, and the fact that one entity can be addressed differently.[12] Especially the vocabulary sizes of the filtered corpora are trimmed down a lot to numbers a bit higher than 160 thousand words.

[11] This follows from experimenting with two models, generated over (i) a corpus with only filtered lemmas, and (ii) a corpus in which the entities are also glued together.

[12] This happens when, for example, instead of a full name, a person is addressed with only his or her surname. In theory, this issue could be solved by implementing co-reference resolution.

TABLE 3.1: Corpus statistics based on the pre-processing steps and the selection criteria. Punctuation is removed and the corpus is converted to lowercase for all four corpora. The numbers are words.

	Before filtering		After filtering*	
Type	Words	Vocabulary	Words	Vocabulary
corpus	69,430,134	1,003,618	41,539,031	200,546
lemmas	69,430,134	918,280	37,153,771	168,784
lemmas_stopwords‡	25,960,437	817,728	24,956,331	161,774
lemmas_stopwords_entities‡	25,469,042	1,059,930	24,173,947	164,724

* Chosen is for `min_dfs=2`, `max_dfs=0.9` and `min_tfs=5`. See filtering (3.2.5).

‡ And every other implementation enabled.

The filters that are used also filter out stopwords to a large extent, which can be seen in the relatively small decrease of vocabulary sizes and corpus sizes when compared to the unfiltered corpora.

3.2.1 Reading the files

The FoLiA files from the requested years are read, after which the program returns a list of all the tokens in the FoLiA documents. The program only returns the words that do not include punctuation. Depending on the setting of the module, the word lemmas are returned instead of the textual word representation from the file.

Adding entities to the result

If the program is requested to also return entities from the FoLiA, instead of just words, then a list of entities is created when the file is read. In the sequence of words that would normally be returned, the corresponding words of an entity are replaced by a single string of tokens from that entity. In the case of an entity that consists of only one word, really nothing has changed. However, if the entity spans multiple words, then they are chunked together, separated by an underscore (_). The entity words are also returned as lemmas, even if they are chunked together. This might cause the occurrence of strange entities in the returned text (e.g. *vereenigde_staat*), but this is not detrimental to the interpretation.

3.2.2 Altering the documents

Before the words from the document that is returned by the above step are processed and counted, the words altered on-the-fly are following several rules:

Conversion to lowercase

The entire word list that is returned is converted to lowercase. This prevents capitalised, non-lemmatised words from ending up in the corpus dictionary and forestalls a possible increase of vocabulary size: *Koning* and *koning* are not the same for the topic model.

Further removal of punctuation

If punctuation is still attached to words returned from the tokeniser, it is removed in this step. This happens, for example, with *art.* and *d.i.*.

Removing diacritics

Diacritics on words are removed to possibly overcome spelling differences that could be observed when analysing multiple volumes. This should not have a negative effect on the interpretability.

Removing hyphens

Hyphens are removed from words for the same reason as above. This also corrects the orthography of words that were possibly split up by line breaks.

Removing apostrophes

It turns out that the tokeniser that was used to create the FoLiA files did not separate prefixes and suffixes of the type *d'* and *'s* from the root of the words, and did not remove single apostrophes at the end of words. They are removed, since they have no function in interpretation. This also helps to generalise the corpus' dictionary and should prevent an increase in the size of the corpus' vocabulary.

3.2.3 Counting and creating a dictionary

Every word that comes, whether or not altered, out of the previous steps is incorporated in a dictionary that converts the word to a unique id, stores information about the number of documents that the word is in, and includes term frequency for this word. These statistics help in formulating the post selection criteria.

3.2.4 Dictionary based removal

Previous chapters discussed “words” in relation to the topic modelling algorithm and the textual representation that is presented to it. In fact, a model takes whatever you present it with into account and incorporates it in its internal numerical representation. This can be much more than “words” alone, since punctuation or numbers can also be included. This definition of “word” is convenient for the reader but differs slightly from how it is used by the model. To translate the internal numerical representation of the model into a representation that can be read by the interpreter, the model makes use of a dictionary that translates token-ids into tokens. This means that, if word ids are removed from the corpus’ dictionary, then they will not be analysed by the topic model. Thus, words can be removed from the corpus by eliminating their id in the dictionary. This is done twice:

Removing non-alphabetical words

If words were passed by the FoLiA document that were not entirely alphabetical (e.g. *18de*), then they are removed in this step. This is also done for any punctuation as token that is left behind.

Removing stopwords

A list of stopwords is used to trim the number of words that occur in nearly every document, such as function words (determiners, prepositions, auxiliary verbs) and other unwanted, relatively frequent words that provide no information when taken out of context. These words carry no meaning and encountering them in created word clusters by the topic model is therefore unwanted.

The stopwords are removed by deleting every word from a stopword list from the dictionary by id. The stopword list is based on the standard Dutch stopword list that is available in the Python `nltk` package (Bird, Klein, and Loper, 2009), complemented by corpus related terms and historical variants of these terms (e.g. *den* and *de*). These are added by manually inspecting the top 2,000 most frequent terms in the entire corpus. The stopword list can be downloaded from the repository, but is also included in Appendix B.

3.2.5 Filtering

When the entire corpus is read, analysed and altered by the above-described steps, the following is known for every word: (i) its term frequency, which indicates how many times the word occurs in the entire corpus, and (ii) its document frequency, which indicates in how many documents the word appears. Based on this information, words can be removed from the corpus, since words that, for instance, occur in nearly all documents are not likely to contribute much to the informational value of any topics in the model. This is done by providing a setting for the criteria described below. The numbers that are displayed in the "After filtering" column in Table 3.1 are the result of implementing the measurements below and implement a setting of a minimum document frequency of 2, a maximum document frequency of 0.9, and a minimum term frequency of 5. This is done for the entire corpus.

A minimum term frequency

This is a minimum frequency of words in the corpus. Words that occur once or only a few times have a lower informational value to the algorithm, since LDA generates topics on the basis of co-occurrence across documents. It is, therefore, safe to delete all terms with a frequency of 1.

A minimum document frequency

Words should occur in at least n documents. Otherwise, they are discarded. If a word occurs in one document, then is it possible that only that particular document exhibits a certain topic. It is, however, more interesting to look at topics that are included in multiple documents. This setting depends on the number of documents in the corpus.

A maximum document frequency

The usefulness of including a word in the vocabulary of a corpus declines if the term is less frequent, but also if the term is too frequent. By changing this setting, words are removed from the corpus if their document frequency exceeds the given number: if 0.9 is given, then words that occur in more than 90% of the documents are excluded.

The top n most frequent terms from the term frequency list

This can act as an alternative to the stopwords list. The most frequent terms of a corpus are often non-informative function words (see also Chapter 2.2.1). This might be useful, but there is the risk of deleting frequent content words.

3.2.6 Serialising the corpus

After these steps of altering and filtering, the corpus can be input into the topic modelling implementation of Gensim. However, to speed up the model's iterations over the entire corpus, the corpus is saved to a more compact and agile format: the so-called Matrix Market format.[13] Instead of reading the XML FoLiA files and performing the alterations and filtering on the fly, the corpus is saved in more or less a plain text format that includes only a numerical reference to a word in a vocabulary and a frequency. These results came out of the pre-processing steps and are sufficient for the model. The natural document outlines are preserved in this corpus: a single article in *De Gids* is equal to a single document in the serialised corpus.

3.3 General settings

The written tools and the programs that are required to run the analyses can be accessed and used freely unless otherwise stated. The tools, as well as the experimental settings and files that are necessary to (re)produce the result are available in the repository. The corpus is not included in this repository because its licence does not allow for it, but part of it, if not all, is accessible through the website of the DBNL. If one is interested in using the corpus of *De Gids*, they can contact the Koninklijke Bibliotheek in The Hague.

copyright

3.3.1 Topic Modelling implementation

To apply the LDA algorithm described in Chapter 2.3 on digital data, one needs an LDA implementation that can be run on a computer easily. The pipeline in this thesis is using a library for Python called Gensim[14] (Rehurek and Sojka, 2010), which is a collection of topic modelling tools that includes the LDA implementation that is described by Hoffman, Bach, and Blei (2010) who use a method of Online Learning to train the LDA model.[15] From this toolset, the single core implementation of the LDA model is used, since this has support for automatically setting the priors (see below, section 3.3.1). The Gensim package includes a topic modelling algorithm, but also contains various other useful tools, such as methods to evaluate the state of a model and methods to inspect its contents. It is also being actively developed. Using Gensim instead of e.g. MALLET gives more flexibility in handling and altering the corpus, setting up the experiments, and analysing the results. This way, the corpus

Gensim

[13] See https://math.nist.gov/MatrixMarket/formats.html#MMformat.

[14] https://radimrehurek.com/gensim/index.html

[15] Another known implementation of the LDA algorithm that is quite popular in the Digital Humanities is MALLET. (A. K. McCallum, 2002). Contrary to the learning method in Gensim, MALLET uses Gibbs sampling to improve its model.

```
lda_model = LdaModel(
                corpus=mcorpus,
                num_topics=300,
                id2word=fcorpus.dictionary,
                chunksize=1063,
                passes=20,
                alpha='auto',
                eta=None,
                eval_every=500,
                iterations=500,
                random_state=2018,
                update_every=10
            )
```

FIGURE 3.5: Calling the topic model with the required parameters.

is processing within the same environment, which is more transparent and allows for easier replication.

My decision to choose the standard LDA implementation instead of an implementation that is aimed at analysing sequential corpora is based on the recommendations mentioned in Section 2.5.4. Coming back to what is described by Wallach, D. M. Mimno, and A. McCallum (2009), I have set the hyperparameters in Gensim to `alpha='auto'` and `eta='None'`. The first initiates an automatic learning of an asymmetric α parameter as prior, which influences the sparsity of the topics created. The height of the alpha parameter determines the number of topics that reside inside a document. The second parameter (earlier referred to as β) influences the topic-word distribution. For this, a symmetric prior is chosen by leaving the default setting of Gensim. This is in line with the findings from Wallach, D. M. Mimno, and A. McCallum (ibid.).

3.3.2 Number of topics

There is no straightforward rule to set the number of topics in a topic model. The number that people tend to use in their research fluctuates from 20 up to 500. It can be expected that a smaller corpus contains fewer topics, so this number should be tailored to the size of the corpus that is used. This number also determines how fine-grained the results ought to be: a higher number of topics possibly makes room for the identification of (smaller) subtopics, but might also introduce more "garbage topics." It is possible to get an idea of the right number of topics for a corpus by using a coherence model that comes with Gensim, although this measure, as is the case with the model's perplexity (see Chapter 2.4), should not be leading when it comes to choosing the right model. Using such a measure is the only aid that is available in choosing the number of topics, without making use of human annotators.

coherence Experimenting with Gensim's built-in method[16] to evaluate the model's coherence

[16] This measurement is based on the paper by Röder, Both, and Hinneburg (2015) and features several steps and a combination of existing coherence measures to come to a value that expresses an improved model's coherence. The authors show that their measures show a high correlation to that of

yields questionable and highly variable results, depending on the initialisation of the corpus. This, in effect, adds more uncertainty to the model and to avoid the risk of cherry-picking a nice looking coherence curve, I have decided not to use such a measure. Instead, I will pick a number of topics based on intuition and manual inspection.

3.3.3 Pipeline

A pipeline is written in order to analyse a specified number of documents from one or multiple years. The data is input to the LDA model in a streaming fashion.

3.4 Inspecting and presenting the data

3.4.1 Lists, tables and visualisations

Representing digital data on paper requires some compromises. Listing all topics with a word cloud would implicate that the appendix of this thesis becomes extremely large. Therefore, the topics are displayed in a slightly less informative manner, namely in the form of a list of 30 terms with the highest probability in that topic. This list is, where possible, complemented with the title and the id of the most distinctive document. In other words, the document in which this topic is most prevalent, which might aid interpretation and gives the opportunity to quickly go to the text for further inspection.

Where applicable, interesting topics are displayed as word clouds, which also show the varying significance of words within the topic cluster: bigger words are more distinctive than smaller ones. The topic numbers are assigned by the program used and reflect no meaning or significant ordering.

3.5 Evaluation

Evaluating non-supervised text-mining approaches is hard because of a lack of training data. It would not be within the scope of this thesis to manually create "topics" from the corpus. The previous Chapter (2.4) already discussed that the outcomes of topic models (i.e. the topics found and the clusterisation of these themes), when set

a human (as a gold standard). The representation of the pre-processed corpus is, at the end of the pipeline, a bag of words one, which constrains the use of several proposed coherence measures that evaluate on a sliding document window. Since the model is, in any case, further evaluated by manual inspection, I have experimented with the reasonable scoring `u_mass` coherence measure, which evaluates on the probabilities of the top words in a topic by looking at the co-occurrence counts of these words in the original training corpus (Röder, Both, and Hinneburg, 2015, p. 4). Frequencies and word location are thus left out of sight.

up right, resemble the interpretation efforts of a human annotator. The evaluation technique that is used is, therefore, a qualitative one for every experiment, which is in line with most of the research discussed under Chapter 2.5 that uses the expertise of the researcher or a historian to assess the quality of the topics and to interpret them.

qualitative evaluation

First, a label is given to all topics in the model where possible. This is done by looking at the top words, the most distinctive words for each topic, and, if the result is unclear, the individual document with a high probability for this topic is looked at as well. For each of the topics, the most distinctive text is given, i.e. the text in which this topic has the highest proportion. By looking at the title of the document or its contents, a "meaning" can also be deduced. If this fails, for example, if the topic does not seem coherent from a human perspective, this is indicated, and the topic label is given a "?."

The results in the form of these topics, their labels, and their most distinctive texts are analysed in the light of the expectations for each experiment. Depending on what might be interesting when considering the historical background of the case studies, the models are analysed further.

3.6 Experiments

The corpus is pre-processed according to the steps above and this outcome is used in the experiments that are described below. In experiment 1 and 2, just a fraction of the corpus is selected for analysis. Experiment 3 uses the full corpus. The general settings that are described above are the same for every corpus unless indicated. This also holds true for the evaluation methods. The first experiment acts more like a sandbox in which several possibilities of analysis are examined. The outcomes of these small experiments and evaluations determine which settings are kept in the following runs.

3.6.1 Experiment 1: Analysing one year of contributions: 1848

In this first, small-scoped experiment, I am analysing the results from just one year of contributions made to *De Gids*. I hand-picked a volume from the magazine based on what I expect that the topic model should find. The year chosen for this experiment is 1848, a year in which the periodical and its contributors engage in discussions on the Dutch Constitutional Reform, which was learnt from close reading the magazine and secondary literature. Discussion on the proposed constitutional changes in The Netherlands had already begun prior to this year (see below), but for practical exploratory and testing purposes, I am only considering the 1848 volume of the

TABLE 3.3: Corpus statistics for 1848

	Before filtering		After filtering	
Type	Words	Vocabulary	Words	Vocabulary
1848	194,240	30,137	174,512	14,109
1848 + entities	189,981	31,869	168,283	13,863

magazine. This means that the topic modelling technique is here used to validate the applicability of the technique and to test it on this corpus.

This selection of just one year in which it is known that "something" happened gives the opportunity to validate the model in a qualitative manner before broadening the scale in the next experiments. Although analysing a small number of documents is not recommended and research shows that analysing a larger corpus of documents proves to be more fruitful, the results which are presented in Chapter 4.1 do, in fact, form a coherent whole. More on why this probably is the case can be read in the respective paragraphs in Chapter 5.1.1.

Statistics

Selecting texts from this year yields a total of 72 documents, for 66 of which an author is listed. A total of 36 unique and known authors has contributed to this year's publications. The total number of words that are in this slice of the corpus is 510,799, and the average publication length is 7,094 words, with a minimum and maximum of 160 and 20,211 words respectively.[17]

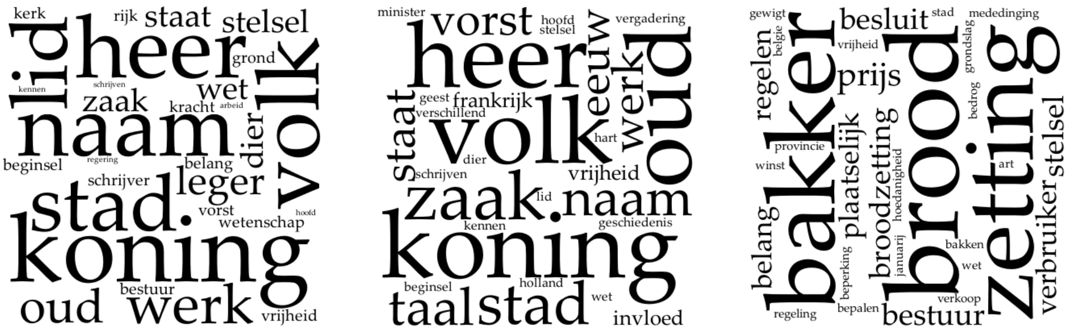
When processed by the pipeline, the corpus has for this year been scaled down to a vocabulary of 14,109 for the normal corpus, and 13,863 for the corpus with entities. This is summarised in table 3.3. Contrary to the overall corpus statistics, the corpus with single term entities has a lower vocabulary size. This is probably a result coming from the filter settings that cause a removal of less frequent entity representations.[18]

Settings

Because of the scale of the corpus for this experiment, there has not been filtered on the basis of a minimum document frequency. The minimum term frequency is set to 2, so that all words that occur once are removed from the corpus. The maximum document percentage is 0.8, which filters out extra (general) words.

[17] This follows from Table A.1

[18] The name *Karel Hendrik Ver Huell* is written in different manners, which all end up as a separate entry in the vocabulary, consequently with a lower term frequency, and thus get filtered out.



(A) Topic 6

(B) Topic 7

(C) Topic 24

FIGURE 3.6: Three topics from a 100 topic model.

Inspection of topic-word clusters coming from a model of 100 topics yields many similar topics, which can be seen in words that are recurring in many of the topics (e.g. *koning* and *frankrijk*).[19] What is more, this number exceeds the number of documents, which is not only uncommon in topic modelling, but also increases the risk of getting one topic that is tailored to a specific document. This can be useful in itself, but this kind of information might also already be read in the title of the document (in this example *De Broodzetting. (_gid001184801_01_0030)*). An example of these topics is displayed in Figure 3.6. In the first two word clouds, quite similar topics are displayed. The third illustrates a topic tailored to an article.

Looking at an amount of 50 already gives better results, but this number might still be too high for this corpus slice, since many similar topics around *koning* and *wetenschap* are displayed. An inspection of 30 topics shows that meaningful clusters are generated by the model, and that most of them can be interpreted. Inspecting an even lower number yields several topics in which two themes appear to be present (the so-called "chimaera topic"), but this does eliminate the topics in the corpus that are less salient.

Therefore, in Appendix C, the topics for the two corpora are given: (i) 30 topics for the filtered corpus, and (ii) 30 topics for the corpus in which entities are presented as single terms (when they consist of multiple words). These results are further discussed in Chapter 4.1.

3.6.2 Experiment 2: Analysing contributions from a small timespan (13 years)

Where the previous experiment consisted of just 72 articles coming from a single volume of the periodical, this next experiment involves a corpus size that comes close to being incomprehensible by a (traditional) researcher. Instead of investigating one

[19] The results are not included in this thesis, but can, of course, be viewed in the appropriate folder in the repository.

TABLE 3.5: Corpus statistics for 1894-1906

	Before filtering		After filtering	
Corpus years	Words	Vocabulary	Words	Vocabulary
1894-1906	3,584,823	276,751	2,998,698	26,018

year, 13 volumes are now examined, which include a total of 1,642 documents, from the years 1894-1906 (inclusive).

As with the first experiment, it is known that "something of relevance" happened in these years on the national and the European stage, although it is not immediately clear if this is also reflected in publications made in *De Gids*. As this "something" I have chosen to examine the Dreyfus affair. To check whether this event, which is described below, is present in the periodical, a single topic model is trained over these 13 years, without providing it with any temporal information: the articles are here treated as from the same source, and since it is expected that there are little to no differences in writing, and that this time period is too short to notice meaning/topic drift, this can be done without facing any problems with regard to chronology. This experiment is different from the previous one in the sense that it is not immediately clear if the historical event itself or its effects and influence can be seen directly in the magazine. It is known that the Dreyfus affair took place in the time span in which the search is conducted, and, based on this, one can speculate that this had its impact on the contributions to *De Gids*. This surmise is backed up by the findings that are described under "historical background" (2.7.2). This means that the topic modelling technique as is here applied will be used as a search mechanism: as an aid to the researcher to find interesting facets of the corpus, its contributions, and its contributors, without being certain about whether something can be found, which is contrary to the outline of the first experiment.

Statistics

Selecting this time period yields a total of 1,642 documents, 1,249 of which an author is known for. A total of 313 unique and known authors has contributed to the publications from these years. The total number of words that are in this slice of the corpus is 8,382,490, and the average publication length is 6,018 words, with a minimum and maximum of 107 and 39,867 words respectively.[20] When processed by the pipeline, the corpus has, for this year, been scaled down to a vocabulary of 26,018 unique entries. This is summarised in table 3.5.

[20] This follows from summarising the values for the respective years in Table A.1

Settings

The minimum document frequency for this corpus slice is set to 5, the maximum document frequency to 0.75, and the minimum term frequency to 10. This yields a corpus with a vocabulary of approximately 26k words. This number is still high compared to the first experiment, but when looking at the relative reduction of the corpus size, this number can be justified for this number of documents.[21]

The number of topics should certainly be higher than the 30 from the previous experiment, but this does not scale with the number of documents in the corpus. In the previous experiment, it was already shown that there is a lot of repetition in themes in the documents in the magazine, as, for example, every issue of *De Gids* includes a *Staatkundig overzigt* section. The expectation is that themes recur in several issues. Therefore, the ideal number of topics should be higher than 30, but most likely lower than 150.

Inspecting the word clusters from a 50, 100 and 150 topic model shows that 150 topics is, in any case, a too high amount, since the model yields several identical topics in the lower proportions. This affirms the expected number of topics for this body of texts and is also reflected in a high perplexity statistic that is calculated by the topic model in an evaluation step.[22] Inspecting the results from a 50 topic model gives a more evenly distributed result, also when looking at the lower proportions in the results. It is surprising to see that this models smallest topic includes terms related to the Dreyfus affair. To see if these clusters might be too dense (e.g. if one topic actually includes two themes), two models of 75 and 100 topics have been inspected, but such a theme is lacking in these topics. Finally, a 60 topic model is inspected, but this topic misses this (in this case wanted) topic as well. The analysis is therefore done with a 50 topic model.

Results on this experiment with the scope of 13 years can be found in Chapter 4.2.

3.6.3 Experiment 3: Detecting changes in a time span of 100 years of contributions.

In the experiments above, the technique is put to the test in analysing a comprehensible corpus size. To demonstrate how topic modelling can be used to its full extent, or to show alternative applications of the technique, it is worthwhile to analyse the entire corpus of *De Gids* that is used in this thesis. This obviously has consequences for

[21] When, for example, looking at the type:token ratio (i.e. the ratio between the unique number of words and the actual amount of words in the documents), this corpus has relatively fewer features than the corpus from the previous experiment: 0.014 compared to 0.009.

[22] A measure that shows how well the model handles and predicts "unseen" documents. But, as was said earlier, this measure should not solely be used to pick a "right" number of topics, but serves in this case as an indicator of a possible too high number of topics, which is acknowledged by manual inspection.

TABLE 3.7: Corpus statistics for 1837-1936

	Before filtering		After filtering	
Type	Words	Vocabulary	Words	Vocabulary
1837-1936	25,384,138	1.059.928	22.674.403	73.963

the type of questions that could be formulated. Instead of focussing on a particular subject as is done in the first two experiments, this time span allow for more abstract or general questions, for example within the field of the history of the sciences and the formation of the sciences and academic disciplines.

Statistics

Selecting on the full first 100 volumes of the corpus yields a total of 10,624 documents, 8,406 of which have one or multiple authors listed in the metadata. A total of 1,452 unique and known authors has contributed to this time periods publications. The total number of words that are in this slice of the corpus is 69,430,134, and the average publication length is 6,535 words, with a minimum and maximum of 12 and 70,491 words respectively.[23]

The same settings as for the previous experiment are used, except for a tightened filter. Setting the minimum document frequency to 5, the maximum document frequency to 0.75, and the minimum term frequency to 15, results in a corpus that has a vocabulary size of 73,963. This is also summarised in Table 3.7.

Settings

The previous experiment has shown how ambiguous or not clear-cut topics could be. A large extent of that corpus slice had no clear theme but can be attributed to a topic of "narratives." It is expected that a similar number of topics in the topic model for this experiment is devoted to narratives as well.[24] Perhaps the hardest part of this experiment is to select the right number of topics for the topic model. For this experiment it is undesirable that the model comes up with topics devoted to a single or a very low number of documents, both for practical reasons (for reasons related to time, as this most likely implies a high number of topics), as for answering the (sub) research question. Ideally, the number of topics is low, reflecting only the words that belong to topics that can be identified as "disciplines" (see historical background in Chapter 2.7.4), but that is not how the model, nor the identification

[23] This follows from summarising the values for the respective years in Table A.1

[24] It might even be the case, depending on the structure of the texts, that the same topics arise in the bigger corpus since the same documents as in the first and second experiment are analysed.

of themes works: some topics contain irrelevant uncoherent information or are uninterpretable, others overlap multiple subjects. This was also shown in the results of the first two experiments. It becomes clear from inspecting various topic model sizes that a higher number of topics is necessary in order to overcome the fact that specific topics[25] are buried in more general topics[26]. For this, the interpretation strategy of this case study is different from the previous two, in the sense that the label that is given to the topic cluster is less detailed.

Since this experiment involves modelling a larger corpus than in the previous case studies, and since this corpus is influenced by a temporal factor with regard to spelling and themes, the settings of the topic model slightly differ. Because the topic modelling technique that is implemented in Gensim is not used as a "true" online model[27], and the number of documents is still comprehensible for the computer, the model's EM-step is calculated over the entire corpus.[28] The model's batch-size is therefore equal to the number of documents: 10.264. The documents are analysed in chunks of 1063. Apart from the lemmatisation that was done in the FoLiA documents, no other regularisation is applied: it is hoped that orthographical differences are levelled out by the corpus' internal structure.

Inspection on models of 75 and 125 topics shows that these numbers are too low for this corpus, since multiple themes are present in many topics, which means that the process of interpreting and annotating is not straight-forward. Scaling up the number to 200 yields better results, but some topics are still clearly showing multiple themes. Scaling the number up further to 300 topics gives many topics that are interpretable. To illustrate the difference between the topic distributions in the 200 and 300 model, in Figure 3.7 the same topic number[29] is included, coming from both models. It is shown that the topic consists of two themes in the 200 model. The theme "chemistry" is not present in the topic of the 300 model, but has been included in a different topic by the model, ideally only reflecting the "chemistry" theme.

[25] Clearly identifiable topics, such as "physics" or "theology."

[26] Topics that contain more general words that are used in many documents, for instance, the topics that could be classified as "narratives."

[27] There is no "new" data fed to the model, since the corpus is fully available from its start.

[28] This means that the modifiers that are discussed in (Hoffman, Bach, and Blei, 2010) (i.e. τ and κ have less of a function here. Adding more documents to the model later to attempt to improve it makes these usable since the "online" method as implemented in the Gensim package puts more weight on the documents that are presented to it earlier in the process, but this would only be a problem with very diverse corpora. I expect that running the estimation over the entire document set solves this issue.

[29] Many of the topic numbers and their most distinctive words run parallel in both models.

Topic 126: pasteur ongeval overtreding inspecteur cijfer belastingschuldige criminaliteit gisting aangifte klasse groep verordening plegen patroon feit vakvereeniging klagen inspectie duitsche pct rangschikken aanwijzing veroordeelingen motor aanslaan dier rijwielwet bevolking spontanea inkomen
Proportion: 0.0002508737
Document: *Kleine criminaliteit.*
Document id: _gid001191601_01_0056 **Probability:** 0.92808664

Topic 126: overtreding belastingschuldige geuns inkomen criminaliteit klasse cijfer plegen verordening opbrengst aanslag groep anna jj bevolking badeloch aanslaan feit belastbaar pct rangschikken aanwijzing hester heffing veroordeelingen sphinx sterk kohier aangeslagenen aangifte
Proportion: 0.0002072901
Document: *Kleine criminaliteit.*
Document id: _gid001191601_01_0056 **Probability:** 0.99908745

FIGURE 3.7: Differences in topic creation in two topic models. The topic from the 200 topic model (above) includes both the theme of jurisdiction and economics, and chemistry. The chemistry theme is not included in the 300 topic model (below). The second cluster from the bigger topic model logically has a lower proportion in the corpus.

Chapter 4

Results

In this chapter, results are described and findings are analysed for all three experiments that were outlined in the previous chapter. The topics that are generated for every experiment are interpreted and the result is given below. Furthermore, the most interesting or remarkable findings are given and presented. For the sake of clarity and readability, some results, graphs and figures are moved to an appendix.

4.1 Experiment 1: 1848

The section below describes the results from a topic modelling analysis on a corpus of a single volume of *De Gids*. For this analysis, a model of 30 topics is trained.

4.1.1 Overview

When looking at the corpus with entities in Appendix C.2, it immediately becomes clear that one particular topic stands out in the coverage of the corpus, namely topic 18. Almost 28% of the corpus consists of the words represented in this topic, of which just a fraction is displayed in Figure 4.2 and the appendix. The theme of this topic, as can be distilled from this representation, but also from its most representative document[1], can be described as "politics," "monarchy," or "government." It not only mentions inland affairs, but it also points to the European stage, mentioning England, Germany, and France in its contents. Further on, it seems that this topic mostly consists of *Staatkundig overzigt* articles, as is illustrated in Figure 4.1 that lists all documents in which this topic has a probability of $> .80$. This also means that the contents of the *Staatkundig overzigt* section of the magazine is consistent throughout its issues in this volume, since all twelve articles are shown in this overview.[2] The other topics are shown less prominently in the corpus, but still include a topic (15)

Staatkundig overzigt

[1] s.n., "Staatkundig overzigt," *De Gids* (12) 1848 779-803. <_gid001184801_01_0040>

[2] This statistic is given by asking for the topic proportion in each document in which the probability for that topic to be present in that document is bigger or equal to 80%. This probability follows from the topic distribution for the document that is expressed in θ (see Chapter 2.3).

1. *Staatkundig overzigt.* (40)
2. *Staatkundig overzigt.* (18)
3. *Staatkundig overzigt.* (12)
4. *Staatkundig overzigt.* (50)
5. *Staatkundig overzigt.* (55)
6. *Staatkundig overzigt.* (46)
7. *Westminster en St. Pauls.* (03)
8. *Staatkundig overzigt.* (34)
9. *Staatkundig overzigt.* (06)
10. *Staatkundig overzigt.* (65)
11. *Staatkundig overzigt.* (60)
12. *Staatkundig overzigt.* (71)
13. *Te Delft, in October 1847.* (05)
14. *Uit Bohemen. een blad uit mijn dagboek.* (68)
15. *Staatkundig overzigt.* (26)
16. *De Republiek der Vereenigde Provinciën in het tijdperk van haren bloei.* (15)
17. *Herinneringen uit Zwaben en Franken. (1846.)* (31)

FIGURE 4.1: Documents in which the probability of topic 18 is > .80. Document id abbreviated (prefix: `_gid001184801_01_00`).

that could be classified as talking about "social themes" and "taxes," with a proportion of 9%, and a topic (14) about "army," "defence"," "colonies," and "press," with a proportion of 7%.

liberalism

Given both the national and international historical developments at the time, it is not surprising that themes around governmental and social issues appear in the topics that are predominantly built up from the *Staatkundig overzigt* articles, but this high proportion confirms that the theme is significantly represented in this volume. The most prominent article[3] of topic 15 also discusses these themes and can perhaps already be interpreted as a theme on "liberalism", or at least a theme that criticises socialism (Aerts, 1997, pp. 186-187). This is the same for the third most distinctive article (p=.79), which discusses the ideas of another French economist.[4] The second most distinctive article (p=.80)[5] attached to this topic also talks about economy, but mentions the constitutional reform explicitly. It is due to the "latent" topic characteristic that this word does not appear in the word cluster in a high(er) position.

4.1.2 Constitutional reforms

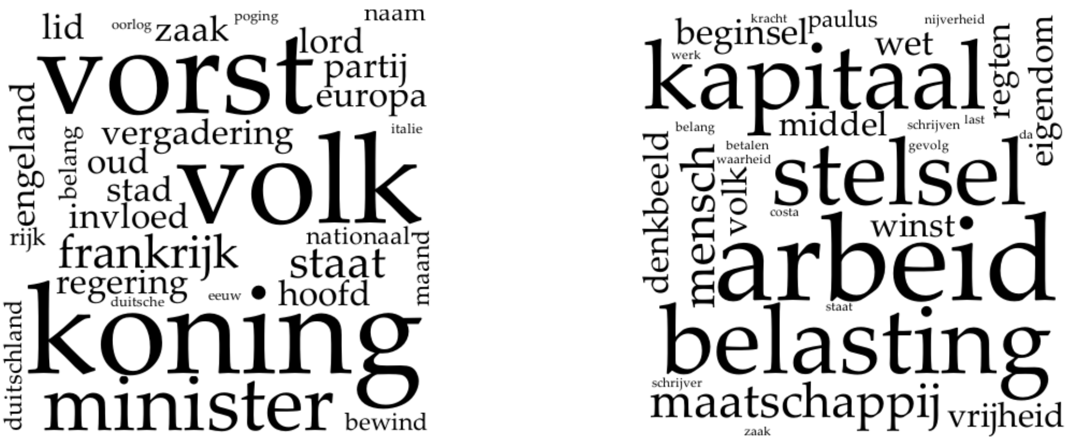
politics

So far, it looks as if the two most prominent topics mention themes that can very well be connected to the reforms of 1848 in The Netherlands and abroad. They capture themes on forms of government, politics, social issues and economy. This can all be deduced from the word clusters that represent these topics. Interpreting the other 28 topics is also rather successful, either by looking at the 30 words with the highest probability of belonging to the topic or by looking at the most descriptive document. Except for the topic with the lowest proportion in the corpus, it is possible to label all of them. The result of this action is given in Table 4.1, in which some keywords are given for each topic. The keywords that carry an ‡ are the result of an interpretation with the history of this time period and the contents of this volume of *De Gids* in

[3] Boer, Willem Richard, 'Frédéric Bastiat en de Socialisten.', *De Gids* (12) 1848 278-319. <_gid001184801_01_0053>

[4] Vissering, S., "Michel Chevaler over Kapitaal en Arbeid," *De Gids* (12) 1848 712-738. <_gid001184801_01_0037>

[5] Rovère van Breugel, J.J., "De Grondwets-herziening en ons Belastingstelsel.," *De Gids* (12) 1848 320-362. <_gid001184801_01_0054>



(A) Topic 18

(B) Topic 15

FIGURE 4.2: Topic 18 and topic 15 from the 1848 corpus in which entities are represented as single terms.

mind. The topic indicated with an ? is not interpretable, since it contains too distinct terms and because it appears to include several themes from other topics: perhaps it still captures themes related to this issue, but a clear-cut label cannot be given to the word cluster; it contains a bit of everything.

Topic 18, as well as topics 15 and 21 that point to themes around the constitutional change and politics in The Netherlands, make up a total of 40% of the volume, but this includes the themes of finance, social themes, and those around the (historical) events in France. These percentages can be compared to the number that is mentioned in Aerts (1997, p. 181), who writes that a quarter of the 1848 volume is devoted to politics.[6] It is not entirely clear how this number was established and what other themes were present in this volume according to Aerts (ibid.). This is the only statistic that is mentioned in relation to this volume.

With this model and its results it is impossible to say the volume solely centres on the constitutional change in itself, and it might be unrealistic to expect that such a topic is so self-contained that it covers only this specific subject. The reality is much more complex and themes are connected to one another. Topic 16, for instance, mainly covers the (natural) sciences but is also present in a document on (a new form of) elections (p=.53). This article[7] responds to the proposed constitutional reform in The Netherlands and the reforms abroad, and is also made up from topic 18 (p=.47), which was linked to every publication on politics. This could be attributed to the fact that the authors in *De Gids* often referred to science and scientific methods to illustrate their points, for instance in the case of the above-mentioned article on socialism

[6] Aerts (1997, p. 181 footnote 1) mentions that he starts counting from 1848 by document and its size (presumably the number of pages) in the magazine. The percentage above comes from the size of the topic in the magazine, which is the result of the sum of the topic proportion in all documents, which is given by the percentage of words from the topic that are covered in the document.

[7] De Clercq, Gerrit, "Twee Nieuwe Bijdragen tot het vóór en tegen der regtstreeksche verkiezingen. " *De Gids* (12) 1848 764-778. <_gid001184801_01_0039>

TABLE 4.1: Interpretation of the 30 topics from experiment 1. This is based on the word clusters in section C.2.

Topic	Interpretation (themes)
0	literature, bible, travel
1^*	emotions, literature
2	synod, church, religion
3^*	social themes, liberalism‡, government, monarchy, France, politics
4	Spain, literature, history
5^*	astronomy, science, physics, Germany, travel
6	science, astronomy, travel
7	science, health, medicine, legal health
8^*	bible, religion, geography
9	Spain, literature
10	colonies, Java, orient
11	Christianity, religion, cultivation
12^*	railways, finance
13	French monarchy, citizens, government, literature
14	army, defense, colonies, press
15	social themes, taxes, finance, liberalism‡, constitution‡
16	science, elections, research, philosophy, religion
17	geography, bible
18	politics, monarchy, government, constitution‡
19	education, Latin, science, physicians
20	religion, history, Middle-East
21	France, Dutch relation to France, government, politics, constitution‡, revolts
22	religion, modern theology‡, bible (Ephesians)
23	monarchy, internal affairs, emotion, Indies, colonies, poetry, art
24	literature, bible, preach, trade
25	prison, sentences, legal issues, social themes
26	biology, animals, anatomy, science
27	art forms, music
28	Spain, literature
29	?

* Topic proportion in corpus $< .001$.
‡ Labelled by connecting to events and other factors of which is known that they are existent in the periodical.
? Topic not interpretable or not clear.

TABLE 4.2: Topic probability distribution for three small documents in the 1848 corpus. Five largest topics given, probabilities are floored. Document id abbreviated (prefix: `_gid001184801_01_00`).

Twee Gedichten van Piet Bogcheljoen (22)		*Koning en volk* (23)		*De Weduwe van Orleans* (32)	
topic	probability	topic	probability	topic	probability
23	0.99750	13	0.99360	13	0.99730
18	0.00030	18	0.00079	18	0.00030
11	0.00019	11	0.00039	11	0.00019
15	0.00019	15	0.00039	15	0.00019
25	0.00009	25	0.00030	25	0.00009

(Aerts, 1997, p. 187).

4.1.3 Small documents

It is interesting to look at how the model handles shorter and more poetic documents in the corpus. In Appendix C.2, topics 23 and 13 are represented by the most distinctive documents, namely *Twee Gedichten van Piet Bogcheljoen*[8] and *De weduwe van Orleans*[9] respectively, which are both (a collection of) poems. The size of such documents affects the topic proportion scores dramatically, which renders the small, distinct topics invisible and captures these documents in a single topic. Both of these documents are responding to, at that time, recent history, mentioning the events around the dethroning of French King Louis-Philippe in the case of *De weduwe van Orleans.*, and the promised reforms by King Willem II in the case of *Twee Gedichten van Piet Bogcheljoen*. Both documents use words that refer to these events. The topic model has attributed a separate topic to each of the contributions, but by zooming out on the topic probability, it becomes clear that some bigger topics also make a slight appearance.[10] Due to the low dimensionality of this data, is it easier to inspect the shared topic statistics for these documents than generating a visualisation in which the interrelatedness of the documents is shown (e.g. t-SNE, a technique to visualise high dimensional datasets).

small document size

As is shown in Table 4.2, *De Weduwe van Orleans* has the same topic distribution for the first five most probable topics as *Koning en Volk*[11], which was already mentioned in the historical background of this time period (see Chapter 2.7.1). This is an even stronger indication that their contents show similarities, although the numbers after the first topic in the distribution are very small.[12] For *Twee Gedichten* only the first

[8] Heije, Jan Pieter, "Twee Gedichten van Piet Bogcheljoen.," *De Gids* (12) 1848 474-477. <_gid001184801_01_0023>

[9] De Génestet, P.A., "De weduwe van Orleans.," *De Gids* (12) 1848 649-652. <_gid001184801_01_0033>

[10] This is one of the effects of a small document size, as it automatically leads to a low topic probability.

[11] Van den Bergh, S.J., "Koning en volk.," *De Gids* (12) 1848 478-479. <_gid001184801_01_0024>

[12] But higher than the long tail, though this is also influenced by the proportions of these topics in the entire corpus.

topic is different: this document basically has its own topic, which it shares with an entry in the *Bibliographisch album*[13], which is about an Indonesian literary yearbook (p=.81). The other topics are shared with the other two documents.

This distribution, although influenced by the overall proportion of the topics included in the corpus, shows the similarities between these three documents on a micro level, looking beyond the first topic in the distribution. Numbers for other documents often include the same large topics, but in a different order and with even lower probabilities. How similar these documents are, is captured by the model by listing two of them in the same topic. "Interpreting " the contents of the *Twee Gedichten* this way appears to be too complex for the algorithm.

4.1.4 Other themes

Other themes in this corpus include the Dutch Indies (topic 14 and 10), social rights (topic 25), and religion-related themes (topic 2, 8, 11, 17, 20 and 22) that all cover separate aspects. Themes on science (topic 6, 7, 16, and 19) mainly refer to the publications on (reviews on) astronomy and physics, but also on health and social themes (parish relief). The Spanish themes are in a separate category as topics but are hard to find without looking at the most distinctive document (topic 4 and 28). Furthermore, it is noteworthy that the name of Isaäc da Costa appears in several topics (15 and 19) in different forms[14], as is the case with religious terms. This says something about his involvement in these religious themes, as is shown in two reviews[15] in this volume.

4.1.5 Differences in pre-processing the corpus

The examples above came from the corpus in which entities are represented as single terms. A comparison of the topics of both corpora in Appendix C shows that the topic model of this result is better at distributing the topics across the documents: one can observe a more evenly distribution of topics. The topic-word clusters are more or less the same in both corpora, but their proportions differ. The second largest topic in the first corpus (topic 24) appears to be totally absent in the second. This is most likely related to the fact that this topic contains words (names, prefixes) that have been filtered out in the corpus with entities. The second corpus thus shows a higher level of generalisability, and this is a wanted effect of the process in which

[13] s.n., "Bibliographisch album.," *De Gids* (12) 388-392. <_gid001184801_01_0056>

[14] He appears more often in the corpus in which entities are not glued together. The total frequency of his surname in the corpus for this volume is 154.

[15] Von Baumhauer, T.K.M., "Mr. da Costa op het gebied der Godgeleerdheid.," *De Gids* (12) 393-429 <_gid001184801_01_0057> and Potgieter, E.J., "Hollandsche Politieke Poëzij.," *De Gids* (12) 739-763. <_gid001184801_01_0038>

entities are treated in the pre-processing steps. For the next experiments, I will work with this extra step.

4.2 Experiment 2: 1894-1906

The results from the previous experiment have shown that the technique of topic modelling is able to highlight the most significant themes in a relatively small corpus. The model was able to capture the foremost theme of politics and give an overview of less salient themes such as colonialism and science. Where possible, this was compared to the findings of Aerts (1997), or to the content of informative articles from the periodical. In other words, the previous experiment has shown that the technique works on this corpus and that it renders useful results.

This second experiment proves to be more complex: not only does this corpus have a larger scale (from 72 to 1,642 documents) but the search goal is more complicated as well. As was described in Chapter 3.6.2, is it unclear whether the theme of the Dreyfus affair, which happened in the period 1894-1906, is discussed at large in *De Gids*.[16] The results given below are therefore connected to the search for this theme or more subtle expressions of and around this theme.

4.2.1 Overview

The number of topic clusters with coherent or interpretable topics is smaller than was expected. The topic model is not able to produce a good fit[17] for models with 100 or more topics and is not able to produce coherent clusters beyond 100 topics, despite the significantly higher number of documents and a more generalised corpus. This means this number is too high for this specific corpus slice. A 50 topic model already includes more dense clusters which by and large cover a single theme. Since the number of topics has logically not scaled proportionally with the number of documents, the topic probabilities that are indicated beside a most distinctive document in Appendix D are lower than in the first experiment. Accordingly, the topic proportions are also lower.

Each of the 50 topics has again been interpreted and labelled. The result of this action is shown in Table 4.3. The two largest topics (35 and 11) that together account for almost 19% of the corpus, are hard to interpret and label. They most likely refer to literary texts or narratives in the 13 volumes of the magazine. This is also

narrative

[16] Simply searching for "Dreyfus" in the periodical in the DBNL renders some results in which Dreyfus is mentioned in passing. It is expected that the model can show in which documents this theme is prominently presented.

[17] The model is not able to distribute the words from the corpus evenly across the given number of topics (e.g. 100). This is likely due to the characteristics of the corpus data: there are not enough distinctive features. Or this is due to the fact that this is inherent to the structure of the corpus.

1. *De doode.* (190101_01_0036)
2. *Een levensdroom.* (189401_01_0031)
3. *De zelfkant der samenleving.* (190101_01_0127)
4. *Geen werk.* (190001_01_0099)
5. *Een arme.* (190301_01_0099)
6. *Vrijage.* (189801_01_0049)
7. *Jan.* (190301_01_0070)
8. *Van de zelfkant der samenleving.* (190301_01_0082)
9. *Een stuk, dat over den kop slaat.* (189601_01_0111)
10. *Een avond.* (189401_01_0005)
11. *De godin die wacht.* (190201_01_0062)
12. *In de gieterij. (Fragment .)* (190401_01_0092)
13. *Waan.* (190401_01_0001)
14. *Wijmpje.* (190501_01_0045)
15. *Sprotje.* (190501_01_0028)
16. *Een grootsch wijf.* (190101_01_0069)
17. *De godin die wacht. (Vervolg.)* (190201_01_0080)

FIGURE 4.3: Documents in which the probability of topic 11 is > .80. Document id abbreviated (prefix: `_gid001`).

illustrated when looking at the documents with these topics as the highest in their topic-document distribution: mainly the *Overzicht der Nederlandsche letteren* and the *Letterkundige kroniek* sections show high probabilities for topic 35. Topic 11 shows many titles of literary contributions (i.e. narratives). Examples of these documents and their titles are given in Figure 4.3.

Other topics are the results of stories (i.e. literature, prose) as well. To show their relatedness, they can be plotted onto a two-dimensional space by positioning them according to their mutual characteristics: similar word distributions. Figure 4.4 displays a two-dimensional plot of a PCoA[18]. Shown is that topics 4, 6, 10, 11, 21, 41, 42 and 48 are grouped together on the horizontal dimension (to the right). This is also largely the case for the largest topic, namely topic 35. Their total proportion in the corpus is almost 34%, measured by the relative size of the topics in every document. This can be explained by the fact that literary contributions of stories and feuilletons (e.g. *De Godin die Wacht*[19] and *De Stille Kracht*[20]) make up a rather large part of the periodical. Rendering this visualisation this way helps identify the themes and nature of the topics that at first glance seem uninterpretable and vague. The model has built separate topics to cover the differences in theme in each of these narratives. For example, topic 11 contains familial terms, while topic 42 contains terms related to country life. Topic 10 and 41 most likely also covers narratives, whilst including many words related to emotion and senses.

themes in narrative

Texts related to science are clustered in topics 1, 18, 22, 34, 39, and 45, and make up 6% of the corpus. Most of them cover themes related to health. For instance, topics 18 and 34 include words that refer to diseases as well as chemistry, and directly mention Louis Pasteur in the most significant terms. Topics 1 and 45 are biology and botany related, and topic 22 mentions terms allied to exploration (i.a. terms referring to nature, landscape, and the Arctic areas). Topic 39 is more of a philosophical and

[18] Principal Coordinate Analysis: a method to show the relatedness between points (here: topics) in a vector space. Similar topics, in this case, are clustered around the same coordinates in the space.

[19] Published in six parts. First publication: De Wit, Augusta, "De godin die wacht.," *De Gids* (66) 1-52. <_gid001190201_01_0062>

[20] Published in two parts. First publication: Couperus, Louis, "De stille kracht.," *De Gids* (64) 383-478. <_gid001190001_01_0080>

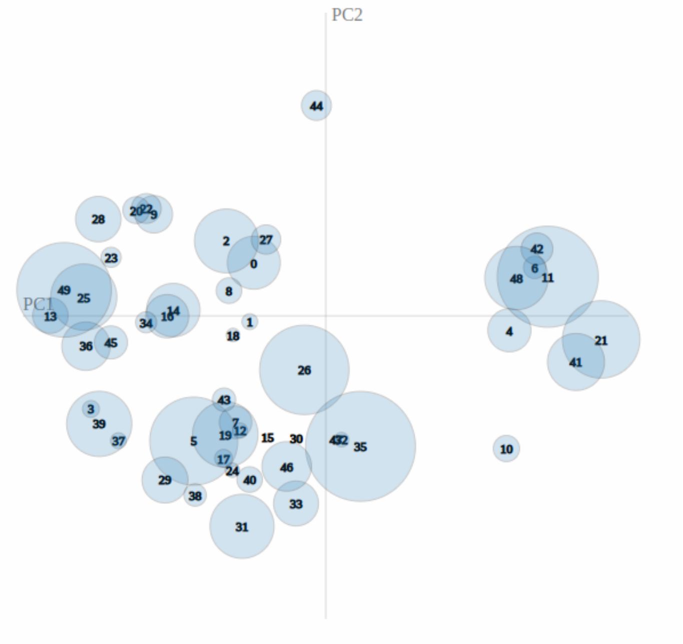


FIGURE 4.4: Plot of the 50 topics from the topic model using MDS (Multi-Dimensional Scaling). Visualisation generated by pyLDAvis (Sievert and Shirley, 2014). The circle's dimensions reflect its relative proportion in the corpus.

methodological order and has *Modern positivisme* as its most distinctive document, which indicates its scope. This topic combines philosophy related terms.

Another large topic is the theme of education (topic 49, 7.6%), Catholicism and religion (topic 26, 6.8%), and (classical) music (topic 5, 6.6%). The latter proportion is influenced by the number of *Muzikaal overzicht* sections in the magazine. This section first appeared in 1894 (Aerts, 1997, p. 492). This is the same for the *Dramatisch overzicht* section that first appeared in the 1881 volume, and discusses theatre plays and drama, which is captured by the model in topic 7 (0.9%). Similar to what was shown in the first experiment, this illustrates the consistency of these thematic sections in *De Gids*.

4.2.2 The Dreyfus affair

The result that is shown in Appendix D is somewhat of a surprise, despite my speculations on the topic. Where I had written in the previous chapters that it might be unrealistic to expect that a topic with "dreyfus" as one of the first terms would appear in the results, this has actually occurred in the model of 50 topics. The same words were also appearing in models that generated more clusters but they were not as clearly grouped as in the 50 topic model. It must be noted that this cluster appeared by coincidence (though in accordance with the data): the same settings were

TABLE 4.3: Interpretation of the 50 topics from experiment 2. This is based on the word clusters in Appendix D.

Topic	Interpretation (themes)
0	diplomacy, revolution, Europe
1	biology, botany, plant culture, science
2	Anglo-Boer war, South-Africa, colonies, diplomacy
3	language, linguistics, evolution theory
4*	narratives
5	music, art, composers, German classical music, arts
6*	narratives
7	drama, arts
8	?
9	church, Middle-Ages
10*	emotion, narratives
11*	family, literature, narratives
12	?
13	Islam, fatherhood, jurisdiction
14	diplomacy, social themes, foreign affairs
15	Dreyfus affair
16	Old-Dutch, older literature, Dutch history, history
17	fables, art criticism
18	plague, diseases, health, science
19	literature, literary studies, letters, correspondences
20	housing, social themes, socialism‡
21*	songs, poetry
22	exploration, travels, science
23	agriculture and farming, libraries
24	religion, drama, philosophy
25	politics, liberalism‡, administration
26	Catholicism, church, religion
27	Indies, colonies, England
28	Indies, colonies, finance, politics, administration
29	literature, language and culture, literary studies, arts
30	United States, health
31	France, poetry, arts
32	literature, publishing, religion
33	literature, naturalism‡, realism‡
34	Louis Pasteur, health, chemistry, science, biology, bacteria
35*	literature, narratives
36	jurisdiction, law, prison, sentences
37	translation, translated literature, publishing, copyright
38	education, Dutch language
39	science, positivism‡, philosophy
40	poetry
41*	emotion, art, expression
42*	country life, narratives
43	dance, arts, electricity?
44	seafaring, marine, military
45	nature, biology, science, botany
46	German, drama
47	?
48*	narratives, naturalism‡, realism‡
49	education, administration, politics

* Narratives, literature or stories.

‡ Labelled by connecting to events and other factors of which is known that they are existent in the periodical.

? Topic not interpretable or not clear.

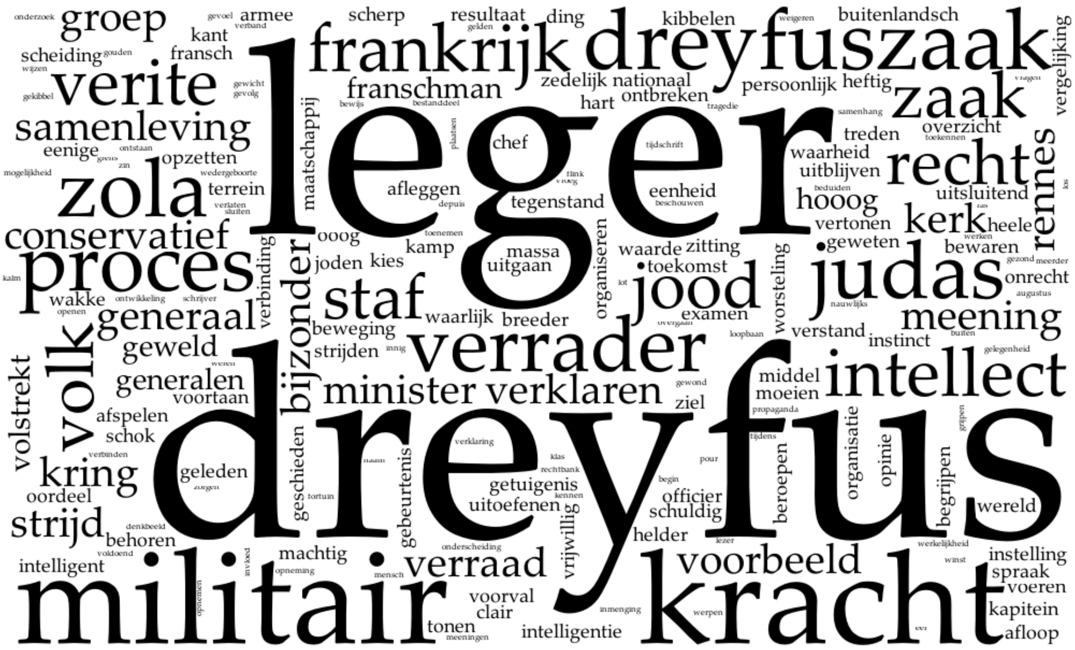


FIGURE 4.5: Wordcloud based on the 500 terms with the highest probability from topic 15 from the second experiment. Shown is the topic around the Dreyfus affair. The size of the words is determined by their probability in the topic.

used in all models, starting with the same seed.[21] This clustering of these words means that all the words that appear in this topic in a high position are somehow related, as follows from their (co-)occurrence in the corpus, as was calculated by the LDA algorithm. The most significant words from this corpus are shown in the word cloud in Figure 4.5.

Nevertheless, this topic's proportion in the corpus is very small (0.02%). This is also illustrated by the number of documents that feature a relatively large probability of this topic in their document-topic distribution. A list of these documents is displayed in Figure 4.6. To obtain this list, the minimum probability was set to > 0.01. This is much lower than the minimum probability that was used in similar other figures of distinctive documents in this thesis, and indicates how small the size of this topic in the documents, as well as in the corpus, is.

The list in Figure 4.6 shows that the Dreyfus Affair is discussed in the *Buitenlandsch overzicht* sections of the magazine: one time in 1895, three times in 1898, and three times in 1899. This section of *De Gids* had started appearing in the magazine in 1893 and featured large "personal notes and reports on foreign events from the political, societal, and cultural field" (Aerts, 1997, p. 530). Moreover, the theme is present one time in the *Letterkundige kroniek* in 1903. Below, each of these articles and their contents are discussed, in order of publishing in the magazine.

[21] The same initialisation is picked for reproducibility and is equal for all experiments and all models, and this outcome is a result of the first creation of this model with this particular corpus.

The *Buitenlandsch overzicht* section

In the first publication that is listed, the author of the *Buitenlandsch overzicht* section, W.G.C. Byvanck (1848-1925)[22] discusses the position of Joseph Chamberlain in the 1895 United Kingdom general election. This article can be read as an ode to radicalism (or radical liberalism), or at least in favour of Chamberlain. Despite the very low probability of this topic in the document (2.0%), it is still interesting to show how this document is listed in this overview, which is due to the usage of the same terms that are prominent in the other listed documents. The author describes the situation around Chamberlain with the same discourse that is used in the other articles, using terms such as *radicaal*, *dissenter*, *vrijheid*, *macht*, *overlooper*, *vijand*, and *Judas*. Dreyfus is not mentioned, so the listing of this article should be considered an error.

The second article, published in January 1898, puts Emile Zola in perspective to French heroes and heroines (i.a. Voltaire and Jeanne d'Arc) and reflects on his, at the time, recently published *J'accuse*, which signalled his interference in the matter of Dreyfus. The author praises Zola for his activism and recapitulates the entire case up to the intervention of the intellectuals of which Zola was one. It is clear from the wording in the article that the author supports Zola and therefore Dreyfus: he mentions "the army's honour," "anger against Jews and Protestants," and "conservatism" as the main reasons for the Dreyfus delusion. The fact that he introduced his overview by making use of a reference to the Catholic church and Jeanne d'Arc might not be without a reason since the image of Jeanne d'Arc was used during the period of the Dreyfus affair (and after) as a propaganda motif for anti-Semites, symbolising a true French virtue (Kilgore, 2008).[23]

The next document[24] appears in the next issue of *De Gids* and shows that Byvanck has become a true Dreyfusard. He discusses political events in England and Austria, signs of war from America and Asia, and mentions the tension in South-Africa, but says these shrink to insignificance when compared to the "derailment" that is shown in the Dreyfus affair in France. He pessimistically mentions the trial against Zola (7-23 February 1898) and the effect of one of the defendant's speeches. Dreyfus is called *Judas*, and his ethnicity (Jewish) and birthplace (Alsace) are all factors that, together with the "humiliation of 1870" (a reference to the French-German war), that is France's *Majuba hill* (a reference to the decisive battle in the First Boer-War in 1881), cause the derailment of the "fable" that this has become. Again, this illustrates how well read the Dutch *De Gids* public must have been, not only regarding this theme but regarding general (recent) history as well.

[22] But indicated below the article in the tradition of the magazine as "B*".

[23] To which Frenchman, probably from the fifteenth century, the author refers when he mentions '*die ouderwetsche theoloog van voor vijf eeuwen*' (p. 381) is still unclear to me. It might be Simon de Cramaud (1345-1423).

[24] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (62) 581-584. <_gid001189801_01_0028>

In September 1898 the next publication[25] appeared, which follows the course of events as described in Chapter 2.7.2. Byvanck extends his metaphor of the derailed fable to this piece, in which he situates France in a drama that was derived from "a novel featuring spies and crooks." It is interesting to note that he does not explicitly mention Dreyfus or the other actors in the affair, but the document is still signalled by the topic model. Words such as *verrader*, *spionnen*, and *onschuld* may be the cause of this, in his brief passage.

It takes a while before the next piece appears[26], but Dreyfus had not been forgotten. In July 1899 Byvanck writes in his section about the first The Hague Convention of 1899, at which the superstates of the nineteenth century were present. When discussing France's position on the world stage, he mentions the position of the army in France, and this does not go without mentioning the case of Dreyfus, which he sees as a "danger to the army's position" and a "threat to the republic." In a very brief summary he uses the terms *militaire quaestie*, *burgerlijke rechten*, and *antisemitisme*.

The final two articles by Byvanck are published in August[27] and September[28] 1899. In the first article, which is entirely devoted to the Dreyfus case,[29] Byvanck reflects on the second court-martial in Rennes. He combines his earlier metaphor of the fable with Dreyfus's nickname when he mentions the two institutions that triggered the spread of this *Judaslegende*: the army and the church. What is more, the imprisonment of Picquart and the exile of Zola are mentioned by Byvanck, along with other figures, such as Zola's lawyer Labori, and one of the defendants in Rennes, Freystaetter. He goes on to discuss the separation between the intellectuals and dreyfusards on the one hand, and conservative society on the other, and gives his opinion on the role that Dreyfus's *Joodschap* and *Joodsche nationaliteit* have played in the affair. Dreyfus is referred to as *gevangene van het Duivelseiland*, and is "reformed" into a Dreyfusard himself: one that wants the best for France.

After Dreyfus had been convicted for a second time and after he was pardoned in September 1899, Byvanck, in his final article, expresses his marvel at a failed "rebirth of the French spirit" when he reflects on the outcome of the process in Rennes. He mentions Dreyfus once, and criticises the French government, after which he goes on about "a more important matter" in the South-African republic.

The *Letterkundige kroniek* section

There is one article that stands out in the foreign section of *De Gids*, and that is the discussion of Zola's posthumously published *Vérité* (1903) in the *Letterkundige*

[25] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (62) 167-170. <_gid001189801_01_0104>
[26] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (63) 368-371. <_gid001189901_01_0080>
[27] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (63) 562-566. <_gid001189901_01_0092>
[28] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (63) 197-200. <_gid001189901_01_0103>
[29] This is also reflected in the 99.7% probability in the model.

99.7%	*Buitenlandsch overzicht.* (189901_01_0092)
10.6%	*Buitenlandsch overzicht.* (189901_01_0080)
8.0%	*Buitenlandsch overzicht.* (189801_01_0019)
6.2%	*Buitenlandsch overzicht.* (189801_01_0028)
3.7%	*Letterkundige kroniek.* (190301_01_0043)
2.0%	*Buitenlandsch overzicht.* (189501_01_0098)
1.7%	*Buitenlandsch overzicht.* (189801_01_0104)
1.5%	*Buitenlandsch overzicht.* (189901_01_0103)

FIGURE 4.6: Documents in which the probability of topic 15 is > .01, here sorted by percentage. Document id abbreviated (prefix: `_gid001`). The volume of the publication year can be distilled from the first four numbers from the document id.

kroniek by an unnamed author in 1903.[30] In his review of the book, which Zola based on the events in the Dreyfus affair, the author views the book as a faint copy of the Dreyfus case. He gives an extensive summary in which he is able to mirror every passage in the book to events from the case. Zola is also compared to Couperus, in whom he sees Zolas match when it comes to the description of a child character. The main point the review makes is about the position of the (Catholic) church, both in the book and in France in the nineteenth and twentieth century. The author is critical and negative when it comes to the parallels to the case of Dreyfus in *Vérité*:

> Voor zoover het proces, in *Vérité* verhaald, een transpositie of een parallel wil zijn van het Dreyfus-proces, dat zoovelen onzer maanden, jaren lang in spanning heeft gehouden, moet gezegd worden, dat het werkelijk proces aangrijpender was, dieper indruk heeft nagelaten dan dit verdichte, al was de verdichter ook een man van zoo groote kracht als Emile Zola. (p. 166)

The author indicates what impact the communications around the affair have had: for many months, the affair kept him (and the reader) in suspense.

Not included articles

To illustrate how well the model functions in serving relevant documents, it is also relevant to discuss the articles from *De Gids* that were not picked up by the algorithm but do have Dreyfus as a (main) theme. These are both found by simply searching for "Dreyfus" in the DBNL, and by reading every *Buitenlandsch overzicht* from the 1894-1906 time period. Not included in the list of documents (Figure 4.6) that do contain Dreyfus as a topic with a relatively high probability, is the first article in *De Gids* in which Dreyfus is more or less explicitly mentioned and is also a *Buitenlandsch overzicht* from September 1896.[31] This article mentions Dreyfus in passing while describing events and conflicts in the world at that time:

[30] This must have been J.N. van Hall (Aerts, 1987c, p. 74). s.n., "Letterkundige kroniek.," *De Gids* (67) 161-167. <_gid001190301_01_0043>

[31] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (60) 185-188. <_gid001189601_01_0110>

> Zelfs heel ver weg in de leêge stilte van Zuidelijke zeeën hooren we de wanhoopszucht van den uitgebannene der samenleving, Judas Dreyfus, zooals hij genoemd wordt, alsof al wat verdrietig, wat krenkend, wat schrijnend was, zijn stem moest voegen bij de blazende disharmonie. (p. 185)

While this is the first mention of his name, neither the author of this section, nor any other author in the magazine explains what has happened to the "social outlaw in the southern seas." He does not even mention his first name but calls him by the term that was used in, among others, a political anti-Semitic cartoon: Judas Dreyfus (Chanteclair, 1894). This not only presupposes that the readers of *De Gids* most likely already knew the course of events from other media (see Chapter 2.7.3), but the use of this nickname further suggests that the reader must have been very well informed indeed.

Another missed article[32] is a very subtle attack on Felix Fauré in the *Buitenlandsch overzicht* of September 1898. In addition, an article that was excluded is one[33] by Byvanck from 1903, which reflects on the funeral of Zola, and which refers to the case of Dreyfus as "his [Zola's] act," and as a "turning point in French history," since this event has sharpened France's resilience against "militarism and clericalism." All three of these articles feature topic 0 (diplomacy, revolution, Europe) with a high probability (respectively 66%, 42% and 52%). The first article also includes topic 2 ((Boer) war and diplomacy, 20%) and topic 11 (family and narratives, 9%). The second article includes topic 26 (Catholicism and religion, 17%) and topic 35 (literature and narratives, 17%). The third article has a high probability for topic 26 (Catholicism and religion, 18%). This distribution correctly captures the contents of the articles.

Finally, several other articles that include the term "Dreyfus" are present in *De Gids* in this time period. Some of them go more in depth than others and mention Dreyfus in relation to the state of France[34] and the world stage[35]. The others mention Dreyfus in passing to give an example of an event that was of interest, for instance in two fuilletons: in Louis Couperus's *De Stille Kracht*[36] and in a chapter of Bas Veth's *Het leven in Nederlandsch-Indië*[37].

[32] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (62) 535-538. <_gid001189801_01_0127>

[33] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (67) 369-374. <_gid001190301_01_0117>

[34] Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (62) 584-587 <_gid001190101_01_0034> and Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (69) 198-202 <_gid001190501_01_0010> and Kops, W.P., "La jeunesse dorée.," *De Gids* (64) 271-301 <_gid001190001_01_0073>

[35] Van Hamel, Anton Gerard, "Paul Kruger. ," *De Gids* (68) 358-366 <_gid001190401_01_0075>

[36] Couperus, Louis, "De stille kracht.," *De Gids* (64) 1-112 <_gid001190001_01_0091>

[37] Veth, Bas, "Op de boot naar Batavia.," *De Gids* (64) 261-289 <_gid001190001_01_0041>

Relation to other themes

How topic 15 correlates with other topics is hard to tell. Its co-occurrence ($p > .01$) with other topics is limited to the documents that are listed in Figure 4.6. Ignoring the documents in which this topic plays a marginal role, the topic frequently occurs together with topics 0 (diplomacy, Europe), 2 ((Boer) war and diplomacy), 7 (drama and arts), 11 (family and narratives), 14 (diplomacy, social themes and foreign affairs), 25 (politics, liberalism and administration), 26 (Catholicism and religion), 35 (literature and narratives), and 46 (German and drama).[38] These numbers are, however, solely based on the documents discussed above. Lifting the constraint on minimum probability renders topic 15 invisible.

After inspecting the articles that were not captured by the topic model, it appears that topic 26 (Catholicism and religion) is attributed to two of the articles. This theme can be placed in perspective to the Dreyfus affair. Although this topic has a large proportion in the corpus (6.8%), its mentioning in the *Buitenlandsch overzicht* sections in the periodical should also be considered: this is not a common place for this theme. This fact therefore supports the findings that this theme often occurs together with the Dreyfus theme, along with themes related to diplomacy and administration.

4.2.3 Other themes

What was also happening in this time period, as was already briefly indicated in Chapter 2.7.3 and in the discussion of the articles above, is the Second Anglo-Boer War, which lasted from 1899-1902 and received much attention in The Netherlands (Bossenbroek, 2012). Articles on this theme also appear in *De Gids*, and words related to this theme are captured and clustered in topic 2 (proportion 3.5%), although this topic is also used for articles that talk about “war” in general.

What was mentioned earlier in the Theory section, is now shown in the topic distribution of this corpus in topic 46: languages tend to cluster together so that, in this case, German words are clustered in one topic because of their similar co-occurrence in the documents. This topic is related to texts on drama.

4.3 Experiment 3: 1837-1936

Whereas the previous experiments showed a combination of distant and close reading strategies to zoom in on a single theme in a relatively small corpus, the third experiment centres on a distant reading approach of the first 100 years of *De Gids*, the full corpus that is used in this thesis. The 300 topics in the generated topic model

[38] Shown by calculating the correlation between topics in the overall document-topic distribution.

are each interpreted in the light of the search goals of this case study (see Chapter 2.7.4). This means that it was attempted to annotate each topic, if possible, with a disciplinary label or with a theme, such as “narrative” or “language.” The results of this interpretation are shown in Table 4.4, in which the (disciplinary) labels that were given to each of the topics in the annotation procedure are grouped by their academic division: Arts, Sciences, or Social sciences. Other themes and categories are listed under Other.

4.3.1 Overview

The number of topics per discipline already give a global idea of the variety of each discipline in the corpus.[39] To provide insight in the proportion of each discipline or theme and each division, the proportion in the entire corpus is displayed in the first column of Table 4.4.[40] In line with expectations, History, Literary studies, and Political sciences make up a large number of the 300 topics, which is also reflected in their respective proportions: 7.24%, 13.83% and 16.99%. Topics that refer to literary contributions, parts of novels, poetry, or other forms of prose and storytelling have been categorised as “narrative.” This category makes up a reasonable amount of the corpus: 23.6%.

Immediately noticeable is that the total proportion of the Sciences is less a tenth of that of the Arts. Just 2,97% of the total corpus contributions can be classified as belonging to studies under this subdivision. The language-related themes under Other are almost certainly influenced by similar orthography between Dutch and foreign languages.[41] These percentages are calculated in the same manner as the proportion of the topics in the previous experiments. These calculations are done for the entire corpus of 100 volumes. To show how these proportions fluctuate during the first 100 years of *De Gids*, several comparisons between disciplines and academic divisions are made in the sections below.

[39] It is possible that a discipline consists of several smaller topics, which each might represent a different theme within this field.

[40] It must be noted that, due to the fact that some topics are annotated as belonging to multiple disciplines (e.g. both Theology and Literary studies in the case of topic 0, see Appendix E), the sum of the proportions is higher than 100%.

[41] For example, it seems unlikely that 2.47% of the corpus is written in French. This could be explained by the fact that highly frequent (function) words in foreign languages, which have not been filtered in the pre-processing, are also used in Dutch. An example is the French *dans* [in] and the Dutch *dans* [dance]. It should also be noted that the label is given by inspecting the most distinctive words in these topics (but with other, harder to define topics, individual - more representative - documents have been analysed as well). Other words, which the model sees as somehow related to French, German, or English words, affect this measure.

TABLE 4.4: Interpretation of the 300 topics from experiment 3. This is based on the word clusters in Appendix E. The disciplinary labels are grouped in categories Arts, Sciences, Social sciences, and Other. In the first column of the table, the proportion of the field and the discipline/theme in the full corpus is displayed as a percentage, but the total is higher than 100. This is due to the fact that some topics are shared between disciplines or fields.

Proportion	Discipline or theme	Topic numbers
34.26%	Arts	
1.26%	Art History	36, 127, 289, 290
1.38%	Classical Studies	140, 150, 159, 245, 248, 275, 278
7.24%	History	3, 5, 17, 18, 20, 29, 31, 34, 40, 58, 61, 81, 84, 109, 131, 134, 142, 143, 154, 156, 164, 168, 170, 174, 177, 178, 179, 197, 200, 224, 228, 234, 242, 246, 247, 249, 285
1.05%	Linguistics	13, 79, 102, 193, 241
13.83%	Literary Studies	0, 7, 15, 26, 28, 38, 39, 43, 49, 51, 64, 69, 79, 80, 86, 93, 100, 110, 114, 117, 120, 121, 125, 133, 135, 140, 147, 161, 166, 171, 194, 200, 206, 207, 215, 223, 226, 227, 230, 238, 253, 265, 268, 269, 270, 273, 275, 280, 281, 283, 286, 294, 296
0.07%	Musicology	55
4.12%	Philosophy	56, 78, 107, 190, 236
5.7%	Theology	0, 9, 27, 56, 74, 81, 87, 108, 115, 118, 148, 165, 180, 212, 256
2.97%	Sciences	
0.71%	Biology	4, 41, 85, 233
0.4%	Chemistry	60, 72, 137
0.49%	Geology	23, 53, 146, 170
0.8%	Health	2, 105, 162, 267
0.16%	Mathematics	112, 132
0.88%	Physics	75, 90, 105, 113, 119, 137, 267
30.59%	Social sciences	
1.55%	Anthropology	4, 10, 16, 19, 35, 53, 70, 89, 128, 237, 256, 277
0.1%	Criminology	126, 184
3.67%	Economy	30, 88, 91, 155, 175, 189, 205, 208, 299
3.1%	Geography	10, 19, 22, 77, 158, 170, 232, 242, 277, 297
3.77%	Law	30, 98, 122, 131, 162, 189, 288, 293
2.41%	Pedagogy	70, 172, 188, 195, 209, 217, 219, 272
16.99%	Political Sciences	8, 11, 12, 21, 28, 32, 42, 45, 50, 52, 53, 67, 69, 79, 83, 88, 95, 104, 111, 124, 129, 141, 144, 151, 152, 167, 199, 201, 202, 205, 209, 221, 231, 235, 239, 240, 266, 274, 284, 292, 297
0.6%	Psychology	213, 262
2.84%	Sociology	39, 41, 52, 65, 126, 130, 149, 175, 182, 217, 229, 254, 259, 260, 266
33.48%	Other	
1.06%	?	1, 14, 25, 59, 68, 123, 176, 210, 251, 279
3.94%	Art*	37, 73, 85, 191, 203, 216, 250, 252, 291
23.6%	Narrative∗	6, 24, 33, 37, 44, 46, 47, 48, 57, 62, 63, 69, 71, 76, 82, 92, 94, 96, 97, 99, 101, 103, 106, 116, 136, 138, 145, 153, 157, 160, 163, 169, 173, 181, 183, 185, 187, 192, 196, 198, 204, 211, 214, 218, 220, 222, 225, 243, 244, 257, 261, 263, 264, 276, 282, 287, 295, 298, 299
1.14%	Language: English	54, 66
2.47%	Language: French	54, 255, 258
1.29%	Language: German	139, 271
0.37%	Language: Old-Dutch	186

* Contributions on painting, music and other art forms.

∗ Narratives, literature or stories.

? Topic not interpretable or not clear.

4.3.2 Looking over time

To see how disciplines develop over time in *De Gids*, the proportion of all topics is measured for each volume of the periodical.[42] Per year, the sum of all the topic proportions of all the topics in a label from Table 4.4 makes up the proportion of this respective discipline/theme in that year. For instance, the total proportion in a volume of the discipline "Chemistry" is calculated by adding up the proportions of topic 60, 72 and 137. The 100 values are then plotted in a graph in which the x-axis represents time (the years 1837-1936) and the y-axis represents proportion in the volume (from 0.00-1.00).[43]

Several variants of plots are given and discussed in relation to the proportion of themes in the magazine. These include a comparison between (i) topics from the Arts and Sciences (Figure 4.7), (ii) topics from theology, philosophy, and sciences (Figure 4.8), (iii) topics from arts, sciences, social sciences, and other topics (Figure 4.9), and (iv) an individual graph for each of the 30 annotated disciplines and themes. The latter is moved to Appendix E.2.

A second presentation is given in Appendix E.3 in which, for each of these 30 disciplines/themes, a scatter plot is drawn in which a polynomial regression trend line is plotted. Each point in the space of these plots represents a document that has a probability higher than .01 for one of the topics from Table 4.4 that make up a discipline. If a document contains multiple topics belonging to the same discipline, the probabilities are added up. Histograms on the side of the graphs indicate the frequencies of each value on the y-axis and x-axis. Contrary to the proportion plots, these plots weight every document from the corpus in the same way: lengthier documents do not boost the score in a year. Instead, these plots give insight into the number of documents that are devoted to a certain topic over time, and show how strongly a document is connected to a certain discipline. Their trendlines show how the probability of a discipline in a volume varies over the years. If relevant for a low amount of data, which is the case for infrequent topics, confidence intervals are shown. For aesthetics and convenience, disciplines or themes within the same academic field or group have been given the same colour.

[42] The granularity of this "search" is thus divided by year. In theory, this could be further refined to a division by issue, but since the metadata that come from the DBNL do not allow for easy issue extraction, and since a single issue might be too small for this type of analysis (i.e. just 6 documents at once (12 issues per year), or 18 if individual parts (4 per year) are analysed), it was decided to use the natural boundary of a single year, or one volume.

[43] This is presented as a line graph for convenience, but could have been represented as a bar chart or a scatter plot as well. This should and does not imply a smooth transition in disciplinary proportion over time. However, for viewing convenience, and to correct the granularity of a single volume, the triangular window of 5 years of a proportion value is plotted. Please mind the varying scale of the y-axis, as numbers can get very small for infrequent topics.

4.3.3 Arts and Sciences

In this first comparison, the proportion of topics related to Science is compared to the proportion of topics related to the Arts, in accordance with the labels in Table 4.4. The graph in Figure 4.7 shows two lines that represent the proportion of Science and Arts related contributions in the periodical in its first 100 volumes. Contrary to expectations following from Chapter 2.7.4, the proportion of Science related matter is rather consistent throughout the years.[44] This could be explained by the fact that there had always been some attention for the Sciences in *De Gids*, either coming from (specialist) reviews in its early years, articles especially meant for the general public, or from commemorating articles. Although the proportion of the Sciences stays equal during these years, the number of documents that feature a topic from the Sciences increases from 1860 onwards, as is shown in the horizontal histogram in Figure E.6. This rise should be attributed to a larger number of articles with a low probability for a Sciences related topic. Nevertheless, the same trend is shown with contributions from the other fields: The number of documents with topics related to either of these fields roughly increases over the years. This could either mean that the total number of documents per year increases, which is indeed the case (as shown in Figure 3.1), and/or that the contributions to *De Gids* become less specific: a document might feature several themes coming from different fields. In this case, an increased number of articles related to the Sciences does not necessarily mean that there is also increased attention for this field since the number of documents on other fields rises as well.

Two small spikes are shown around 1907 and 1922. The graph in Figure E.2 shows that it is mainly Biology and Geology that are responsible for these spikes, as well as Physics in 1922. The latter might be related to the death of former editor J.P. Kuenen in this year. He was commemorated by, amongst others, Hendrik Lorentz[45].[46] No special reason other than a lengthy article[47] on Darwin and his *Origin of Species* explain the rise in 1907, apart from the popular-scientific contributions by i.a. Van 't Hoff and Lorentz in this time period (Aerts, 1997, p. 454).

There is more to say about the other line in the figure that represents the proportion of contributions from the Arts in the magazine. It shows a decline in documents related to the Arts from 1890 onwards. This is the same time period in which the attention for art and music was rising, as is shown by the respective plots in Figure E.7 and E.9. This is also in line with Aerts (1987a, p. 90), who writes about increased interest in painting, music, and drama. It appears, by inspecting other graphs from

[44] The proportion is small. The scale, but also the scope of this case study make it hard to zoom in on small changes. More detail however, is given in the next graph (Figure 4.8).

[45] Lorentz, H.A., "Kuenen als natuurkundige .," *De Gids* (86) 1922 208-215.<_gid001192201_01_0127>

[46] This was also the year in which it was 100 years since Louis Pasteur was born, which was celebrated in the periodical by: Van Loghem, J.J., "Louis Pasteur. 1822-1922.," *De Gids* (86) 1922 323-353. <_gid001192201_01_0136>

[47] Van Bemmelen, J.F., "Verdwenen dieren.," *De Gids* (71) 1907 450-473.<_gid001190701_01_0061>

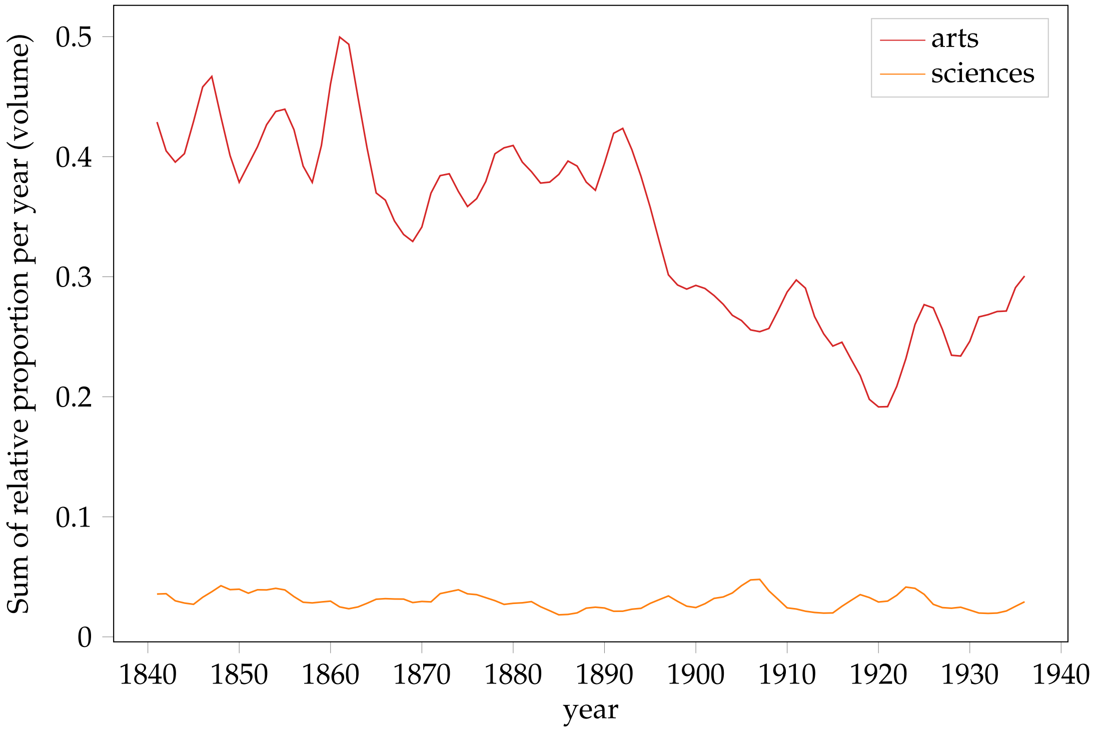


FIGURE 4.7: Relative proportion of disciplines belonging to Arts and Sciences per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

the figures in the appendix, that this decline is mostly caused by the two largest disciplines within this field: history and literary studies. It is hard to find a particular cause of this. Changes of the periodical, as well as editorial changes are likely to have had some part in it, for instance the change of editors that came with the resignation of jurist W. van der Vlugt (1853-1928), who claimed that the magazine did not pay attention to societal and political developments (Aerts, 1987a, p. 92). It is indeed this period in which the proportion of political contributions has dropped below 15% in the magazine for half a century (see Figure E.5).

Also notable is the left bar in the horizontal histogram (and thereby also the density of the points in the space to the left side) in the Arts graph in Figure E.6, which shows there are many contributions in this time period of the first ten years of the periodical. This can largely be attributed to the high number of publications related to literary studies in these years (see the corresponding plot in Figure E.10),[48] which, in turn, should be attributed to Potgieter and Bakhuizen van den Brink in this time period, when the magazine had yet to be considered "general and cultural." It would take until 1847, the time when Gerrit de Clercq proposed a transformation of the magazine into a real general cultural one, before the magazine was also giving attention to foreign policy and sciences, instead of reviews and fine literature (see Chapter 2.6.2). This transformation can be seen in the steep decline of the proportion

[48] Art History, Classical Studies, History, Linguistics, and Theology show similar trends.

of narrative related publications (Figure 4.9) and the surge of contributions related to social sciences. It is mainly Political Sciences that is at the root of this (Figure E.11, taking up 20-25% of the magazine's content in this time period.

A claim that was made by Saris and Visser (2005), which was already mentioned in Section 2.7.4, is that the number of Science-related publications in the twentieth century is as much as 8%. Although the data in this experiment stops in 1936, the graph in Figure 4.7 shows that the proportion of the Sciences is slightly higher in the twentieth century compared to earlier years, but does not exceed 5%. Other research should show how the proportion of the Sciences varies in the course of the twentieth century, but based on this figure it can perhaps be stated that the percentage is close to the number given by Saris and Visser (ibid.), and both their and this result show that the proportion of the Sciences in *De Gids* is smaller than other fields. This is furthermore consistent throughout the years.

4.3.4 Science and religion

The previous section discussed the proportion and prevalence of two fields in *De Gids*: the Arts and Sciences. In line with the description in Chapter 2.7.4 it can be said that the introduction of (new) scientific methods is visible in *De Gids*, among others in the formation of Modern Theology. To further shed light on how the magazine deals with the introduction of the (natural) sciences and their methods in relation to the more traditional fields, Figure 4.8 below further divides the Arts into Philosophy and Theology.

As stated in Chapter 2.7.4, the editors of *De Gids* decided to minimalise the number of publications on religion after 1862.[49] This is clearly shown in Figure 4.8, as this is the point at which the line that represents the proportion of Theology steeply declines.[50] It continues to decline until the 1930s when less than 2% of the magazine is devoted to this discipline. This means that, over the course of 70 years, the proportion of Theology related publications has shrunk to below that of publications on the Sciences, coming from a proportion of almost 18%. The decline of discussions on this topic in the 1860s should be placed in the context of an increasing philosophical liberation (Aerts, 1987e, p. 58), which coincides with the open decline of the Modern Theology movement in The Netherlands and thus *De Gids*.[51]

This is the time in which Conrad Busken Huet, a by then former clergyman, was asked by Potgieter to join *De Gids'* board of editors. When he was appointed in 1863

[49] Aerts (1987f, p. 45) writes that this decision had little to no effect. He mentions 1862 again in Aerts (1997, p. 252), but does not write about any lack of effect. In any case, the result from my analysis shows that something did change with publications on Theology after this year: i.e. there was some effect.

[50] The scatter plot in E.11 shows a similar trend of Theology related documents: a decrease in publications that include Theology as a theme, and these publications also feature less Theology, as the average probability of this discipline in the documents drops as well.

[51] This follows from Chapter 2.6.2 and Buitenwerf-van der Molen (2007, p. 13).

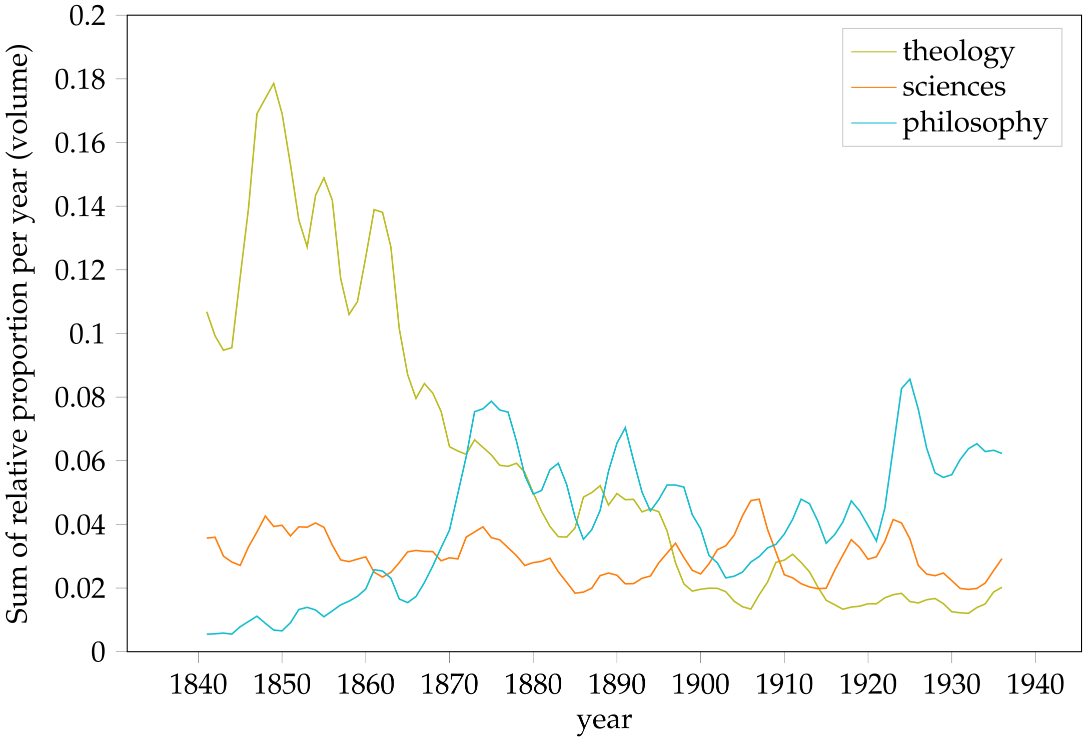


FIGURE 4.8: Relative proportion of Theology, Sciences, and Philosophy per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

he even got his own section *Kronijk en Kritiek* to write about literature (Aerts, 1987f, p. 47), but, interestingly, this did not lead to an increase in the proportion of Literary Studies (see Figure E.2). The fact that Busken Huet, together with Potgieter and Van Limburg Brouwer quit the board of editors in 1865 might have influenced this (Aerts, 1987e, p. 45). It would take until the 1870s before this proportion started to increase again when Charles Boissevain (1842-1927) was appointed as literary editor (Aerts, 1987b, p. 61). However, the proportion of fine literature did increase around this year.

The graph in Figure 4.8 shows that, from 1962 onwards, the proportion of Philosophy related articles took a surge, whereas the line in the figure shows a strong negative correlation with the proportion of Theology from the 1860s. By this time, positivism got a foothold in *De Gids*, which was instigated by, among others, scientist and clergyman Allard Pierson (1831-1896) (Aerts, 1997, p. 252), which might be captured by this increased proportion of Philosophy. A closer inspection of the combined terms from the topics related to Philosophy[52] shows that this discipline includes Theology related terms as well (e.g. *devotie*, *god*, *scheppen*), but more prominently features terms such as *bewustzijn*, *mensch*, and *denkbeeld*.[53] Further assigning this discipline to a particular movement, trend, or school by looking at the top terms of the aggregated topics is quite difficult.

The fact that the proportion of Theology founders from the 1860s onwards should be sought in editorial decisions of the board of *De Gids*, as well as in societal factors, such as an overall decreased attention for (Modern) Theology. Figure 4.8 (and the individual plots) shows that the decline in interest in Theology makes room for the growth (in the proportion) of Philosophy related articles. These two disciplines show the strongest and largest (negative) correlation in the 1860-1880 period. To further identify the exact themes that are discussed in these Philosophy topics, a close reading of the individual articles is needed, since the combined top terms for the respective topics are not sufficiently informative.

4.3.5 A disciplinary overview

The next Figure 4.9 shows a similar graph that contains the three academic fields Arts, Sciences and Social Sciences, as well as the proportion of art and narrative, and the category "other." Scatter plot versions are given at the start of Section E.3. Individual plots are given in the Appendix at Sections E.2 and E.3.

In Chapter 2.6.3, it was mentioned that (Aerts, 1987d, pp. 120-121) claimed that in the early twentieth century 40% of the magazine was devoted to literary contributions, and 30% to arts, sciences and society related matter. Figure 4.9 shows an increasing

[52] Topics 56, 78, 107, 190, and 236.

[53] And, for this discipline, very distinctive names: *Thomas van Kempen* and his *De imitatione Christi* (15th century).

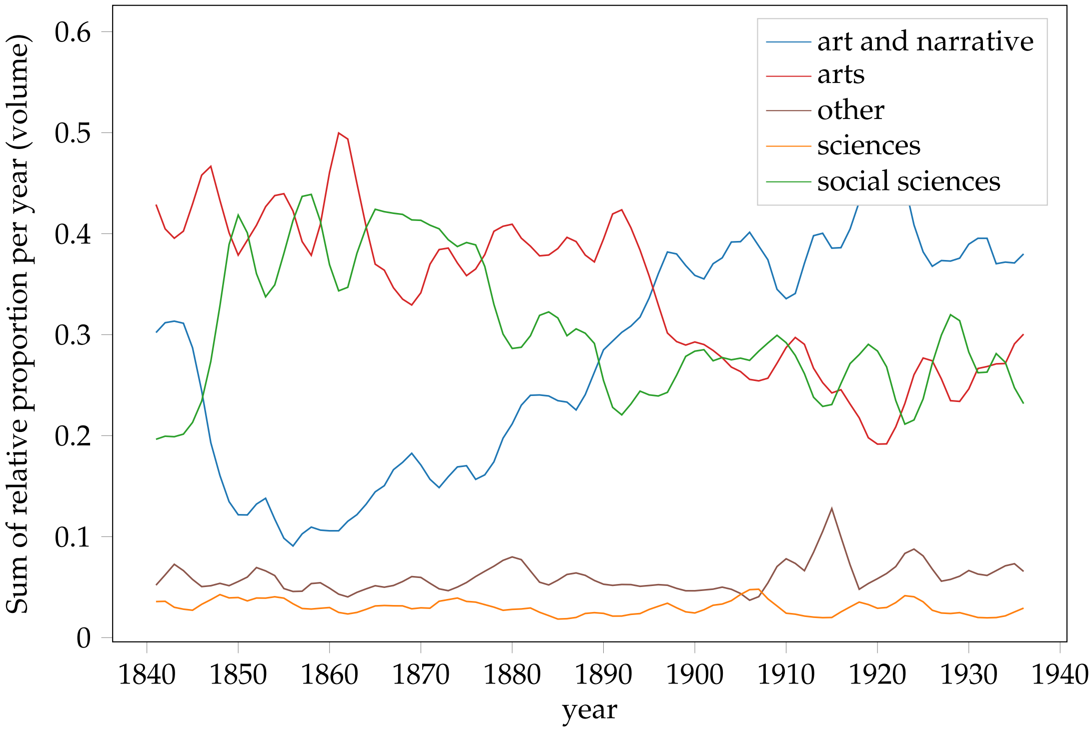


FIGURE 4.9: Relative proportion of art and narrative, Arts, Sciences, Social Sciences and other per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

trend for the proportion of contributions featuring art and narrative. By 1890, the proportion of art and narrative is just over 30%, and this continues to increase until the 1920s when almost half of the magazine's content can be characterised as art or a literary contribution. This development then halts at just below 40% until 1936. This percentage is fully in line with the manual counting of Aerts (1987d). In Figure 4.9, it can also be read that the proportion of art and fine literature increases at the cost of the Arts and the Social Sciences, as the proportions of these fields decline when the proportion of art and narrative increases.

The Social Sciences line in Figure 4.9 primarily consists of the proportion of Political Studies related matter, as this is the discipline that has a nearly constant proportion of more than 10% in the magazine. The emergence of this discipline in *De Gids* that is shown in the first 10 years likely has its foundation in the liberalist agenda of the magazine and its contributors. In Chapter 2.6.2 it was already stated that the magazine propagated discussion and a freedom of speech from 1848 onward, and was one of the driving forces of the liberalist movement. Contributions in this area can be seen in Figure 4.9 in the period 1848-1875, and can more closely be inspected in the individual plots of Political Sciences and Sociology in Figure E.5. The high proportion of Political Sciences can be attributed to this particular involvement of *De Gids*, and the high proportion of Sociology in this time period is likely related to the discussion around poverty relief, universal suffrage, and other social themes in this time period (Aerts, 1987f, pp. 35-36; Aerts, 1997, pp. 340-341). However, these topics slowly disappear after 1875, the time in which the (old, Thorbeckian) liberals cleared the road for a younger liberal movement, or when there was discord amongst the liberals connected to *De Gids* (Aerts, 1997, p. 317). The Political Sciences line (Figure E.5) reaches a proportional low around 1890. Just as with the Sociology proportion, can this be interpreted as a stagnation of the representation of (old) liberalist views in the magazine, and a transformation into a new liberalist, more democratic view (ibid., p. 345).

When looking more closely at the plots of other individual disciplines, it can be seen that the largest disciplines in the magazine are History, Literary Studies, Political Sciences, and Theology. In general, the proportion of any of those disciplines decreases over the course of the 100 years in the periodical, although History shows a rapid rise at the end of the timespan, but it is hard to relate this to a particular cause. Some of the other disciplines show an interesting pattern or trend throughout the 100 years that were analysed as well. For instance, from the horizontal marginal plot of the Biology figure in E.7, it can be deduced that the number of documents that feature a Biology related topic largely increases from 1860 onwards, which is visible in two spikes in 1900-1920 that can be observed in the proportion plot in Figure E.1. Aerts (ibid., p. 454) writes that the interest in scientific fields other than Health and Biology fades away in the thirty years after 1865. This is reflected in the graphs about Chemistry, Mathematics, and Physics, but not in that of Geology (Figure E.1ff.). Biology and Health became more popular. However, it must be noted that the percentages and the differences over time are very small for these disciplines: there often is a change of less than 1%, or lower in the case of Mathematics.

Criminology, Economy, Law, and Psychology show similar trends of an increasing attention for each of these disciplines.[54] Economy even makes up 7% of the magazine in the 1870s and 8% of the magazine in the 1930s. The former percentage is related to writing on the *landrentekwestie* (Land Securities System or Cultivation System issue) in the Dutch Indies, which has been criticised by *De Gids* since 1848, mainly based on the ethical implications of these systems (ibid., p. 433). The Cultivation System did not comply with the liberalist ideology of amongst other P.J. Veth.[55] Eventually, the *Suikerwet* and the *Agrarische wet* of 1870 lead to an increased liberalisation of the Dutch Indies and put an end to this system (ibid., p. 430). These measurements and the developments in the Dutch Indies were discussed in *De Gids*, by, among others, lawyer W.K. van Dedem (1839-1895).[56]

[54] A similar trend of decreasing attention is shown for the discipline of Geography, which is placed under Social Sciences, but this might be due to its interconnectedness with Theology, for instance in articles such as: Veth, P.J., "Bijbelsche Aardrijkskunde." *De Gids* (12) 1848 121-176 <_gid001184801_01_0007>.

[55] Veth, P.J., "De Cultuur-wet." *De Gids* (30) 1866 65-122 <_gid001186601_01_0025>

[56] His articles pop up when filtering on the most distinctive documents for the Economy discipline: Van Dedem, W.K., "De laatste regeling der landrente op Java." *De Gids* (37) 1873 103-126. <_gid001187301_01_0039> and Van Dedem, W.K., "De regeling der landrente op Java." *De Gids* (39) 1875 31-101. <_gid001187501_01_0062>.

The 8% in the 1930s can be linked to the economic situation of The Netherlands and its neighbouring countries in this time period, as well as the global issue of the Great Depression of the 1930s. Texts with high probabilities for Economy in this time period relate to economic aspects of the crisis of the thirties, as well as political and military tensions in Europe.[57] Both of these time periods seem to be connected to the proportion of Law-related contributions, as the graph (Figure E.3) shows a similar trend in the 1870s and 1930s. Also, this discipline features the same documents with high probabilities. Not only the Dutch Indies are criticised in Law related matters, but injustices in other (non-Dutch) colonies are as well (Aerts, 1997, p. 425). Contributions around 1930 are about the same developments in Europe.[58]

It is difficult to relate the rise of the proportion of the disciplines mentioned above to the disciplinarisation (i.e. disciplines tend to specialise into individual fields and become more specific over the years) discussed in Chapter 2.7.4. The landscape of disciplines in the magazine has become less diverse at the end of the first 100 years, as it is mainly fine literature and Arts that are included in its contents. Small rises in disciplines such as Criminology, Economy, Law, Pedagogy, and Psychology (all Social Sciences) might suggest that there is more attention for such studies in society, but the examples above of Economy and Law show a relation with events on the world stage as well. At least it shows how popular these themes and disciplines were in the first 100 years of *De Gids*. It has become clear that some of these disciplines are related and show similar trends over the years, or make room for other fields. It is hard to tell at which point a discipline really became established using this method, either in *De Gids* or in society.[59] Nevertheless, fluctuations in proportion or frequency of a discipline can be directly related to major events in The Netherlands or abroad. Findings of, among others, Aerts (ibid.) or primary articles from the magazine confirm this claim. Further research into the propagation of disciplines should deal with a larger corpus since *De Gids* only shows a small excerpt of the interest in Arts, Sciences, Social Sciences, and other themes, as there are magazines from its time period that focus exclusively on a particular discipline or field. This is, however, closely connected to the transformation of *De Gids* into a cultural literary magazine over the years.

[57] For instance mentioned in Plate, A., "Politiek en maatschappij." *De Gids* (93) 1929 6-37 <_gid001192901_01_0039> and Van de Woestijne, W.J., "De positieve bestrijding van het oorlogsgevaar" *De Gids* (95) 1931 218-249 <_gid001193101_01_0135>.

[58] For instance, developments in politics and the rise of fascism: Maas Geesteranus, H.G.J., "Het Fascisme en het Italiaansche strafrecht" *De Gids* (94) 1930 268-276 <_gid001193001_01_0047>.

[59] This is partly related to this method of topic modelling: every topic occurs in every document, just in varying probabilities.

Chapter 5

Discussion

This thesis had the following two goals:

1. Shedding new light on the corpus of *De Gids* by using an unsupervised digital tool;
2. Asserting the applicability and usability of topic modelling for humanities oriented cultural-historic research.

The focus of the research questions in this thesis was closely related to the methodology used. From its start, this study has centred on the use of one tool in particular to analyse the Dutch general cultural periodical *De Gids*, instead of reasoning from a point of problem solving and question answering using (any) digital tools as aid. It has been specifically decided to investigate whether the technique of topic modelling could still be considered a useful one in Digital Humanities' research, in which this thesis can be situated. By formulating three very diverse case studies that each focus on a different time period and aspect of the magazine, not only an answer to this question has been given, but the case studies have also revealed new information from the magazine, or have confirmed existing claims about the periodical's contents.

The results, as they are presented in the previous chapter and its appendices, provide an answer to the sub-questions as formulated in Chapter 1. By starting small and gradually incrementing the complexity of the method and the research questions, it has become clear how the technique of topic modelling, and specifically LDA, can be applied in research in various ways, but this method has also highlighted its pitfalls. The first two experiments show how the technique can be used as a potential powerful search tool, and exhibit how an interpreter might deal with both known and unknown events in the topic modelling results of a small or modestly sized corpus. The first experiment has shown that the technique is able to capture and highlight events that are present in the corpus. The second experiment applied the same methodology to a bigger and therefore more complex and uncertain corpus. The third and final experiment showcases the application of topic modelling on a macro scale, by taking a first step in giving a distributional overview of the disciplinary contents of *De Gids* throughout its volumes. The usefulness of the results varies per experiment and are discussed below, together with the usability of the method and its limitations.

5.1 Evaluation

5.1.1 Experiment 1

The corpus that was used for the first case study can be considered small and still comprehensible for "traditional" research. The added value of using a technique such as topic modelling on such a small corpus is that it can give insight in the corpus very quickly (given that the output is generated easily). The findings of this first experiment have, in practice, turned out to be obsolete due to the preliminary research that was done in Chapter 2.7.1 (historical background), but this was part of the subquestion that was used in this first experiment. It has been demonstrated that the outcomes of a topic model on this volume of *De Gids* can be traced and interpreted in the light of the historical background of 1848. It is the case, as is illustrated by the documents that are presented underneath the clusters in section C.2, that looking at the title of a contribution to *De Gids* often already gives the same insight as the topic clusters would give a human reader in this corpus slice, although this is only the case for the smaller topics in this part. The titles of texts in *De Gids* are not always informative (e.g. in the case of the *Bibliografisch album*, the subject can be anything that is discussed in a review).

A conclusion for the first experiment is that topic modelling on such a small corpus works, but should only be used for very specific goals. This could be to discover topical co-occurrence or spotting the topical proportions in the corpus: something that could be done by hand[1], but can be done more easily using this method. In that sense, training a model on a modestly sized corpus can be compared to applying the technique to a single (chunked) novel or text, something that is not recommended in related work (see below). The fact that one particular topic that has a proportion of 28% in the volume stands out, and immediately gives insight into the main theme of this specific volume: politics. However, it does not yet reveal the requested theme of the constitutional reform of 1848 in itself. The model captures recurring sections of the magazine (e.g. the *Staatkundig overzigt* sections), which not only shows the internal consistency of these parts of the periodical, but also proves the validity of the model's outcome: when documents and topics are grouped, this is meaningful information. Other applications of the model's outcome have shown that it can be used to find similar documents, which was done with three literary contributions

[1] This was done by Aerts (1997) who occasionally gives percentages for themes in his corpus of *De Gids*.

(related to the main theme of politics) in this particular volume, but the result is not very strong and proves to be little to no aid in interpreting the contents of the articles.

The results from this first case study show to what extent the magazine was involved with political matters in 1848. By comparing these results to the historical background of *De Gids* and the affiliation of its editors, this comes as no surprise. It might have been insightful to analyse the years before and after 1848 in the exact same manner, although this was largely done on a zoomed out scale in the final case study. The results affirm how closely *De Gids* was connected to politics and (at that time period) liberalism.

The results also confirm that the technique of topic modelling is working for this corpus, and shows that the technique can even be applied to smaller corpus sizes, but also that one has to stay in close contact with the text's contents to obtain useful results. Immediately starting to close read and interpret the texts might be faster than setting up a topic modelling environment and interpreting its results. If the burden of setting up, testing and fine-tuning the result can be removed, then I would certainly recommend looking at the results of a topic model at a glance. In this particular example, it would show that this volume is mainly about politics (which can serve as a starting point for further research), even if one does not have background information on its year of publishing.

5.1.2 Experiment 2

With regards to the method, the second experiment mainly shows how capricious topic models can be, and that there is also a bit of luck involved in getting the "right" topics in your model for analysis. At least, this is the case with the topic modelling implementation that was used in this thesis (traditional unsupervised LDA) in which no prior belief of the words in the topic clusters was given to the model to find a particular "Dreyfus topic."[2] To a large extent, this is related to the difficulty of finding the "right" number of topics for a model, as the number of model sizes that were inspected for all experiments illustrate. Techniques to automate this do not work well enough.

The best model was found by examining various model sizes, and had to contain leads to manifestations of themes in *De Gids* around the Dreyfus affair from 1894-1906 in France. The model that is used in the analysis actually contains a topic entirely devoted to this theme, and this outcome can be considered remarkable, given

[2] In theory, the belief that there must be a topic around "Dreyfus" could have been given to a different partly-supervised implementation of the algorithm (such as the technique used in Templeton et al. (2011)), also as a search mechanism. This might be quicker for this type of search goal but increases its complexity.

the nature of topic models and the fact that there was little to no information on this theme available in secondary literature.[3]

Again, this method can be interchanged with close reading all articles from *De Gids* from this time period, but the benefit of setting up a topic model is that the relevant articles are already grouped according to their theme. The model provides an extra layer of (automatic) interpretation of the result (by clustering similar words), and is thereby able to find and cluster similar texts. This way, texts can be found that do not include the term "Dreyfus" but merely allude to the French officer and the events in France. The model guides the interpreter/reader in finding the right position in the corpus and the right texts. Through plain reasoning, one can, of course, also search for subordinate words, such as *verraad* or *officier*, but I would argue that this method is easier.[4] In any case, the topic is distinctive and demarcated enough to be spotted by the model as a distinct topic. The fact that one of the documents[5] explicitly discusses this theme might be the cause of this. However, there were also a few documents about the Dreyfus theme that were ignored by the model, even though they featured "Dreyfus" as a term. They were published in the same section of the magazine as the documents that were identified by the model as belonging to the topic. Its results should, therefore, be seen as an aid in searching a source, and not as definitive method to reveal all occurrences of a theme.

The model also showed that it is able to differentiate between literary contributions (e.g. narratives) in *De Gids*, and other articles on politics, administration, or scientific reviews. This method can be used to quickly assert the percentage of a certain category in the corpus, as was done in the first experiment with the politics related matter. The proportion that is presented in this thesis reflects the real proportion of the words in the corpus, and does not account for changes in document size.[6]

Considering the representation of the Dreyfus affair in *De Gids*, this case study has given valuable insight into how the affair was followed in The Netherlands in general and in *De Gids* in particular. The few mentions of the affair are often very subtle and suggest that the magazine's reader was well informed about it. The results show the position the magazine took in the early days of the affair and its compassion with

[3] Moreover, a similar Dreyfus related topic is included in the final experiment. Topic 235 in this case study includes *joden*, *militair*, *officier* and *leger* as most distinctive themes, indicating that this theme, or the theme of military terms and the mentioning of *jood* and *joden* stands alone as very distinctive in the corpus. This is probably one of the reasons this theme was generated by the 50 topic model in this case study.

[4] Again, that is if the model produces such a result. The result that was presented in this experiment around the Dreyfus theme seems extraordinary. At the same time, this shows one of the topic models strengths.

[5] I.e. Byvanck, W.G.C., "Buitenlandsch overzicht.," *De Gids* (63) 562-566. <_gid001189901_01_0092>

[6] This is one method of counting. Others would be to simply count the number of documents related to a certain topic, although the problem of multi-topicality is also present. Another method would be to count the number of pages in an article, but counting the number of words would be a more accurate measure.

both Dreyfus and Zola.[7] These findings add to the scarce literature on the representation of the affair in The Netherlands. The fact that *De Gids* appeared on a monthly basis during this time period might have something to do with the lack of truly informative articles on the Dreyfus case. The contributions as discussed in the previous chapter are of a reflective or opinionating nature, and not so much of an informing one: there were other channels and media through which one received the news, as was illustrated in the historical background in Chapter 2.7.2. The related themes that were listed at the end of Chapter 4.2 reflect the findings of the close reading of the articles listed. The evidence on which these co-relations are based is, however, small (and were thus comprehensible enough to close-read). Still, this gives an example of yet another method in which the results from a topic model can be used, although I would not claim this can substitute a close-reading approach. The result is simply not strong enough or is too blurry.

5.1.3 Experiment 3

The final case study challenged the technique to show its strengths in analysing the full corpus of the first 100 years of *De Gids*. To give a thematic overview of the contents in the magazine over time, a temporal dimension was added to this analysis, so that trends could be identified and interpreted. Specifically, themes were abstracted to the level of (academic) disciplines, to comply with the nature of this general cultural periodical. Annotating topic clusters in the result of this case study meant interpreting a topic as belonging to a certain (academic) discipline or other major theme. This made the annotation procedure more complex than was the case in the first two experiments. Present-day categories and labels were used in this process, and if there was any doubt about the category of a topic during the annotation process, either (i) both categories were given to the topic as a label, or (ii) this topic was omitted in the results and labelled "?." Allowing this multidisciplinarity for a topic put less stress on abstracting from (specific) theme/topic to general label or field, and lowered the chance of losing information and cherry-picking.

Plotting the discipline proportion and the total number of documents belonging to a specific discipline in a line graph or a scatter plot rendered some interesting findings. Proportion numbers that were given resemble the numbers that were given by i.a. Aerts (1997), and claims that were made in Aerts et al. (1987) and Aerts (1997) were confirmed by statistics from the outcomes of the 300 topic model. Results could often directly be linked to decisions that were made by the board of editors of *De Gids*, as was, for instance, shown in the drop in the proportion of the Theology discipline in 1862. They could also occassionally be linked to other events and related documents. This level of abstraction on disciplinary labels hardly allowed for inspection

[7] This is, in particular, true for the author of most of these contributions, W.G.C. Byvanck.

on individual themes/topics but turned out to be transparent enough to go back to the documents that caused a rise or fall in the proportion of a certain discipline.

The most significant results from this case study come in the form of two trends, visible in the decreasing proportion of Theology (and an increasing proportion of Psychology) related documents, and the increasing proportion of art and narrative in the magazine. Especially the fact that a by an editor proposed change to the magazine was so directly reflected in a declining line in the graph of Theology, suggests that the model can be put to excellent use for this type of analysis. This and other notable changes, and their connection to events and decisions from the magazine's history further underline this. Other results are, where possible, interpreted in the light of the magazine's background or claims from i.a. Aerts et al. (1987) and Aerts (1997). This shows that, to some extent, the technique can be used as a verification device to corroborate existing claims.

This final case study can be situated amongst studies by Newman and Block (2006), Nelson (2010), and Riddell (2014), but differs with regard to the type of corpus used: the contents of *De Gids* are much richer and more diverse than one would encounter when analysing a newspaper or academic journals. This study especially drew inspiration from the work by Riddell (2014) when it came to analysing and presenting the results considering its search goal. This goal was to situate changes in topical proportion in the light of disciplinarisation and to give an overview of the disciplinary distribution of *De Gids* over time. This type of research could even be considered a competitor or another approach to the famous "culturomics" research by Michel et al. (2010). It possibly bridges the gap between manually coming up with a list of predefined words on the one hand, and automatically retrieving synonyms and topic related terms on the other hand (e.g. in research on trends using word embeddings (Hamilton, Leskovec, and Dan Jurafsky, 2016)).

The conclusions that were drawn in this third case study are mostly about the proportion of a discipline in *De Gids*, and not directly about an overall trend in society: a decrease seen in the proportion of a discipline in *De Gids* might, for instance, indicate that possible contributions from this discipline have been moved to a (more specialised) periodical, or that the focus of *De Gids* has changed. One should be hesitant in drawing such inferences. Nevertheless, the results from *De Gids* give some insight into what happened during the time period that was analysed, and might form the starting point for research into larger societal trends.

5.2 Validity

It should be noted that the models that are trained on this corpus are specific to this corpus alone. It contains the words that are used in the texts from *De Gids*, which means that the model is less viable to be used on a set of other texts with presumably a (slightly) different vocabulary. Although the text is partly lemmatised (and therefore partly normalised), it still remains trained on the nineteenth- and early twentieth-century *De Gids*, depending on which model one inspects. However, the corpus is of such a nature that the models can be seen as general enough to be applied to an analysis of similar or comparable magazines from this time period, such as *De Nieuwe Gids* or *Vaderlandse Letteroefeningen*.[8]

Experimenting with the corpus and the models' sizes shows the importance of correctly tuning the model. In practice, this comes down to adjusting the correct setting of the expected number of topics and determining the batch size for the EM-step (for setting local and global maxima). Setting the model's hyperparameters is done by the technique used (the user-friendly Gensim package (Rehurek and Sojka, 2010). The number of training iterations was set to a number at which the model's perplexity was barely changing anymore. Settings vary per case study and can be seen in the code in the repository. Fixed model seeds have been picked in favour of reproducibility.

This research satisfied the general requirements for a successful topic modelling approach as described by, amongst others, Tang et al. (2014) and Brett (2012), although one could argue that the necessity of a high enough number of documents was flaunted in the first experiment. Nevertheless, these results were coherent as well. This is probably due to the document sizes of articles from *De Gids*, which might be longer than is the case for typical documents that are used in topic modelling studies. The number of topics has stayed relatively low, both for practical reasons, and for producing informative results. This, together with the fact that *De Gids* has proven to be a suitable corpus for this kind of modelling, makes the technique (and specifically the Python implementation) more resilient than was first expected. Visualisations, such as ordinary lists, word clouds, and various types of graphs, were picked because they best display an aspect of the results as produced by the model. This, on the one hand, resembled the examples by Brett (ibid.) and Jockers (2013), but adds new types of visualisations that were not used in these studies as well. These visualisations aided the interpretation of the results, and, where needed, the documents were analysed to not solely rely on the top terms presented by the model (cf. Schmidt (2012b)).

This study was different from existing research into topic modelling on periodicals in the sense that the corpus that was used contained a wider range of text types (e.g. narratives, poetry, reviews, informative pieces), which were mostly identified on a meta level, while other research focusses on newspapers or scientific/academic journals. It is because of the unique position that the highly influential periodical *De Gids* had in Dutch culture that this has proven to be an ideal corpus for Digital

[8] If one wants to use the models from this thesis in his or her own experiments, then one can download a serialised version (i.e. an already trained and saved version) of each model from the GitLab repository. See: `https://gitlab.com/LvanWissen/degids`

Humanities oriented research, simply because it, to a large extent, covers the full range of (scientific) interest in The Netherlands from its start in 1837 until the Second World War. Using this periodical, I have tried to give an example of using topic modelling on a Dutch corpus by using relatively easy to use and, above all, accessible tools that are becoming more common in the world of the Digital Humanities (and specifically the history and literature branch of this field).

5.2.1 Historical continuity

I have taken the words of caution of i.a. Underwood (2014), Goldstone and Underwood (2014), and Schmidt (2012b) into account with regard to working with topic models and their assumption of historical continuity. The model implementation that was used did not have knowledge about the year of publishing of the article that was fed to it. The temporal dimension for the analysis was added by requesting topics and probabilities for individual years. Results show that the model is perfectly capable of working with diachronic data, which affirms the claim by Ted Underwood (2014) that vanilla LDA models can be used in diachronic research as well. In the third case study, the fact that one discipline was allowed to contain multiple topics largely corrected for any change in terms (i.e. topic drift) and, for instance, orthographical changes, which are complicating factors in dealing with diachronic corpora. What is more, topics were allowed to appear in multiple disciplines, which made the labelling decisions less strict. This can be one strategy to deal with temporal corpora using the traditional LDA implementation. How these decisions and these results relate to findings from techniques specifically for this usage (e.g. Topics over Time (Wang and A. McCallum, 2006)), could be the subject of a different study.

5.2.2 Limitations

When working with smaller corpora, there is the risk that a topic is tailored to a single document, which happens a few times in the examples given for experiment 1. This same risk exists if the corpus size is scaled up, for instance when a document really stands out on the aspect of theme and contents, which, in turn, might be valuable and useful information in itself. The result really depends on the corpus' contents and the search granularity that is needed for the analysis: changing the requested number of topics from a low to high number implicates a different meaning of the idea of a "theme." Unfortunately, setting this number should be done through trial and error, inspecting several model's sizes and picking the one that shows little to no overlap in theme per topic. Picking the right number of topics is thus not straight-forward. Although coherence measures can aid the human interpreter, this was not considered helpful in deciding on the "right" number for the experiments in this thesis.

This research had the benefit of using a corpus that is in crisp condition and does not suffer from OCR errors. However, it does suffer in the area of language processing, which is shown in incorrect lemmatisation, entity recognition, and coping with (historical) orthographical variation. Furthermore, metadata on the magazine, its volumes, parts, and issues are not always correct or complete. Both the publishing style of the magazine (i.e. anonymously), as well as its size in the DBNL are causes of this.[9]

It is inherent to the model that labels such as *grondwetswijziging*) or "political studies" will (likely) not appear in the topic and have to be inferred. To get the most out of the results of a topic model, one therefore has to become an expert on the corpus or the corpus' time period. Knowing its historical background helps in interpreting the distribution of word clusters that are rendered by a model. And while analysing these results, one should never lose sight of the original texts to contextualise the findings by the model. The technique is no silver bullet for text-analytical research, but merely assists a scholar in his or her analysis.

[9] Perhaps an increased (automatic, e.g. through an API) accessibility of the data in i.a. the DBNL will improve these corpora over time, either via search projects or through crowd-sourcing. This can at least be done with parts of the corpus that are not prone to copyright issues.

Chapter 6

Conclusion

The aim of this research was to investigate whether topic modelling could have a place in modern day literary-historical research. It has given an answer to this question, by showing three concrete examples in which the technique has been used to its full capabilities in excavating (parts of) the eminent general cultural periodical *De Gids*. The three case studies have shown that the real benefit of using the technique exists by applying it to largescale corpora, for instance in a manner that was shown in the final experiment. However, the first two case studies demonstrated the validity of the model's results when it comes to capturing the magazine's contents. The model was able to show the magazine's involvement in politics and liberalism in 1848 and was able to highlight other themes and text types in the period of the Dreyfus affair. For further analysis into the representation of this affair in *De Gids*, the model was used as a search tool that highlighted relevant words and their positions in the corpus and documents.

The final case study showed a result that could not have been presented without the help of the model. Its automatic and unsupervised outcome (a list of 300 themes) was interpreted, abstracted, and annotated with a disciplinary thematic label. After this, for each of the periodical's volumes, statistics on the proportion and frequency of these labels were generated by the model. This made it possible to visualise fluctuations in the distribution of (academic) disciplines and fields in the magazine over time. Remarkable changes in proportion or frequency were checked against findings in the comprehensive and commendable work by Remieg Aerts (1997), and were verified by looking at the articles from the magazine. Transitions in the presented graphs could almost always be linked to the periodical's editorial decisions or environmental factors.

Although the technique is not at all new, it still shows some real potential for use on large corpora that lack metadata or are not yet extensively described, since it offers a relatively easy way for a (Digital) Humanities' scholar to get a grip on a corpus and reveal its underlying thematic structure. To come to a conclusion before going on to a set of recommendations and suggestions for further research, I will reflect on the several modes in which topic modelling has been used in research in the Humanities, as was given by Roland (2016).

This thesis has shown how the technique can give a scholar an entry point to a body of texts, how it can complement a "traditional" close reading approach, and how relatively easy it is to render statistics for a very high number of documents. Furthermore, it has (i) given insight in topic modelling as a tool to guide the interpreter, which was shown in the first, but foremost in the second case study, which revealed relevant documents for the search goal of finding more information around the Dreyfus affair in *De Gids*. It has (ii) revealed "social conditions" of the texts in the corpus, by looking at thematic distributions in one, thirteen, and 100 volumes of *De Gids* respectively. It thereby has situated the periodical in its own time period and has given an overview of its contents in full. And finally, (iii) it has looked at a macro scale to disciplinary trends in *De Gids*, going back and forth through the magazine, zooming in on relevant results, interpreting them and explaining these by connecting them to external factors. These three case studies can serve as exemplary for DH research in Dutch studies. Until the time that other (non-clusterisation) techniques arise that are able to work with unlabelled data, topic modelling is the best technique for this type of research.

6.1 Recommendations

Despite several warnings in the literature on topic modelling in general or in the Digital Humanities, the technique seems to be more robust than was thought initially. Even when a low number of documents is used, as was showcased in the first experiment, the model is still able to produce coherent topics from this corpus. The result obviously becomes more robust when more documents are used, but an initial low number should not be a sole reason to avoid topic modelling as a technique. The real quality of the word clusters created is dependent on the internal structure of the corpus and the length of the articles (and the available time).

The experiments described in this thesis illustrate that there are several ways of applying the technique in DH research. How a topic model can and should be used very much depends on what one expects to find or wants to investigate: It can be used as verification device to verify or falsify (existing) hypotheses, or as a tool to get to know a corpus and its contents. Flexibility in applying the tools and handling the results is key in these cases. Luckily, more and more software is being (openly) developed and published these days. It is implied that one needs to know how to use those tools, but it certainly helps if one also knows how to modify them to fit the data and goals.

Furthermore, it is a given and known fact that most time in DH research is going to the editing and pre-processing of a corpus. This was also the case with the already parsed FoLiA files that make up the corpus of *De Gids* that is used in this thesis. The biggest advantage of this was that the lemmatised version of the corpus could be used, but this could also have been done by using separate tools, possibly more fit to perform this task on this (historical) corpus. The added token layer was not used, and the quality of the entity layer was questionable. Still, in theory, using this standard toolset for linguistic annotation should be beneficial for the ease of processing corpora digitally. If the tokenisation, lemmatisation, and NER in the FoLiA toolset improve, then the quality of the topic modelling results would consequently also be boosted. Training it on Dutch from the nineteenth century time period might be a starting point for this, as it seems that this era is neglected in favour of training on contemporary Dutch or historical seventeenth-century Dutch.

6.2 Further research

First, I find it remarkable that so little literature can be found on what the Dreyfus affair brought about in The Netherlands in the 1994-1906 time period or after, especially given the number of articles that were published in the media during this time span (see Chapter 2.7.3). The documents from *De Gids* that were discussed in Chapter 4.2 suggest that their reader, in any case, was well informed about the affair. Investigating the scarce publications that were made by renowned authors and media, and other side publications that could be found in the DBNL or Delpher might shed more light onto what the affair meant for The Netherlands, its politics, and its values.

Second, another approach for studying this rich corpus could be to look at the contributions that were made by E.J. Potgieter, the leading force in the early days of *De Gids*. It would then, for example, be interesting to see if his contributions significantly differ from publications by authors in the same time period, for instance by looking at thematic content, or authorial style. The same can be done for other authors that put their mark on the magazine.

Third, although the basic "vanilla" implementation of the LDA algorithm in the Gensim package was used in this thesis, the models overall did present relatively coherent and interpretable topics. To see how some of the custom flavours of the technique might further improve the technique, it could be interesting to further assess this corpus of *De Gids* with extended versions of LDA (not suffering from historical interpretative constraints, see Chapter 2.5.4), or even with a combination with a word2vec implementation that combines the strength of distributional semantics with the unsupervised topic modelling approach.

Fourth, the final case study of this thesis explored the territory of the History of Science, and more particularly the history of (academic) disciplines. It looked at this development in disciplines from the perspective of the single cultural periodical *De*

Gids. To extend and enlarge its scope into the creation and formation, and maybe even evaluation of disciplines, another compelling corpus would be the Winkler Prins Encyclopaedia (1870-1882) (and French and German counterparts), although this would first require a better digitised corpus for a start.

Finally, it should be concluded that no technique can beat a human interpreter with regard to quality and inference, but as long as time is limited and corpora keep becoming bigger, using an unsupervised approach such as topic modelling can make the life of a Digital Humanities scholar much more interesting. The fascinating contents of *De Gids* are so diverse, that it can serve a purpose in any research, not only within the Humanities.

Bibliography

Abbott, Andrew (2010). *Chaos of disciplines*. University of Chicago Press.

Aerts, Remieg A. M. (1987a). "'Als Epicurische goden': 1890-1915". In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 90–111. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0007.php.

– (1987b). "'De fakkel der verlichting': 1865-1880". In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 54–71. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0005.php.

– (1987c). "'De vloek der impotentie: 1880-1890". In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 72–89. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0006.php.

– (1987d). "'Onze grijze tempeltjes': 1915-1938". In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 112–142. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0008.php.

– (1987e). "Busken Huet en het beeld van De Gids". In: *De Gids* 150. URL: https://www.dbnl.org/tekst/_gid001198701_01/_gid001198701_01_0009.php.

– (1987f). "Een liberaal tijdschrift: 1844-1865". In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 40–53. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0004.php.

– (1987g). "Voor en tijdens de bezetting: 1938-1945". In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 112–142. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0009.php.

– (1997). *De letterheren: liberale cultuur van de negentiende eeuw: het tijdschrift De Gids*. Amsterdam: Meulenhoff. URL: http://www.dbnl.org/tekst/aert010lett01_01/.

– (2012). "De Gids in de ruimte van zijn tijd". In: *De Gids* 7. URL: https://de-gids.nl/2012/no7/de-gids-in-de-ruimte-van-zijn-tijd.

– (2018). *Thorbecke wil het: biografie van een staatsman*. Amsterdam: Prometheus.

Aerts, Remieg A. M. et al. (1987). *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Den Haag: Meulenhoff. URL: https://repository.ubn.ru.nl/bitstream/handle/2066/104606/104606.pdf.

Bayertz, Kurt (1985). "Spreading the Spirit of Science". In: *Expository Science: Forms and Functions of Popularisation*. Springer, pp. 209–227. URL: https://link.springer.com/chapter/10.1007/978-94-009-5239-3_11.

Bel, Jacqueline H. C. (2015). *Bloed en rozen: Geschiedenis van de Nederlandse literatuur 1900-1945*. Amsterdam: Bert Bakker.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Blei, David M. (Apr. 2012a). "Probabilistic Topic Models". In: *Commun. ACM* 55.4, pp. 77–84. URL: http://doi.acm.org/10.1145/2133806.2133826.

– (2012b). "Topic Modeling and Digital Humanities". In: *Journal of Digital Humanities* 2.1. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/.

Blei, David M. and John D. Lafferty (2006). "Dynamic topic models". In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 113–120. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.2783&rep=rep1&type=pdf.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022. URL: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Blevins, Cameron (2010). *Topic modeling Martha Ballard's diary*. URL: http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/.

Boas, Henriëtte (Jan. 1998). "Emile Zola en de Dreyfus-affaire een eeuw later". In: *Trouw*. URL: https://www.trouw.nl/home/emile-zola-en-de-dreyfus-affaire-een-eeuw-later~a406a47a/.

Bossenbroek, Martin (2012). *De boerenoorlog*. Amsterdam: Singel Uitgeverijen.

Brett, Megan R. (2012). "Topic Modeling: a Basic Introduction". In: *Journal of Digital Humanities* 2.1. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/.

Brugman, Hennie et al. (2016). "Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pp. 1277-12*. Ed. by N. Calzolari et al. Portoroz, Slovenia, pp. 1277–1281. URL: http://repository.ubn.ru.nl/bitstream/handle/2066/162489/162489.pdf.

Büch, Boudewijn (1983). "Dreyfus in pamfletten uitgevochten?" In: *Maatstaf* 31.10/11, pp. 5–12. URL: https://www.dbnl.org/tekst/_maa003198301_01/_maa003198301_01_0116.php.

Buitenwerf-van der Molen, Mirjam Fokeline (2007). *God van vooruitgang: de popularisering van het modern-theologische gedachtegoed in Nederland (1857-1880)*. Hilversum:

Uitgeverij Verloren. URL: https://openaccess.leidenuniv.nl/handle/1887/11453.

Calis, Piet (1987). "'Het standpunt van de liberaliteit': 1945-1987". In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 156–203. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0010.php.

Chang, Jonathan et al. (2009). "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Advances in Neural Information Processing Systems 22*. Ed. by Y. Bengio et al. Curran Associates, Inc., pp. 288–296. URL: http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf.

Chanteclair (Nov. 1894). *A propos de Judas Dreyfus*. URL: http://exhibits.library.duke.edu/items/show/20955.

Deerwester, Scott et al. (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* 41.6, p. 391.

Fleming, Paul (2017). "Tragedy, for Example: Distant Reading and Exemplary Reading (Moretti)". In: *New Literary History* 48.3, pp. 437–455. URL: https://doi.org/10.1353/nlh.2017.0021.

Fyfe, Aileen (2008). "Science and Religion in Popular Publishing in 19th-Century Britain". In: *Clashes of knowledge*. Springer, pp. 121–132. URL: https://link.springer.com/chapter/10.1007/978-1-4020-5555-3_6.

Goldstone, Andrew and Ted Underwood (2012). "What can topic models of PMLA teach us about the history of literary scholarship". In: *Journal of Digital Humanities* 2.1, pp. 39–48. URL: http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/.

– (2014). "The quiet transformations of literary studies: What thirteen thousand scholars could tell us". In: *New Literary History* 45.3, pp. 359–384. URL: https://muse.jhu.edu/article/558875/pdf.

Graham, Shawn, Ian Milligan, and Scott Weingart (2013). "On Topic Modeling". In: *The Historian's Macroscope: Big Digital History - working title*. Open Draft Version. Imperial College Press, London, pp. 76–88. URL: http://www.themacroscope.org/.

Graham, Shawn, Scott Weingart, and Ian Milligan (2012). *Getting started with topic modeling and MALLET*. URL: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet.

Hall, David, Daniel Jurafsky, and Christopher D. Manning (2008). "Studying the history of ideas using topic models". In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, pp. 363–371. URL: https://dl.acm.org/citation.cfm?id=1613763.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky (2016). "Diachronic word embeddings reveal statistical laws of semantic change". In: *arXiv preprint arXiv:1605.09096*. URL: http://aclweb.org/anthology/P16-1141.

Heinrich, Gregor (2008). “Parameter estimation for text analysis”. In: *Technical Note*. URL: http://www.arbylon.net/publications/text-est.pdf.

Hoffman, Matthew D., Francis R. Bach, and David M. Blei (2010). “Online learning for latent dirichlet allocation”. In: *advances in neural information processing systems*, pp. 856–864. URL: http://papers.nips.cc/paper/3902-online-learning-for-latentdirichlet-allocation.

Hofmann, Thomas (1999). “Probabilistic latent semantic analysis”. In: *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 289–296. URL: http://www.iro.umontreal.ca/~nie/IFT6255/Hofmann-UAI99.pdf.

Huijnen, Pim and Juliette Lonij (Jan. 2016). *From keyword search to concept mining - Keyword Generator as a tool for the study of historical news media*. Accessed 2 August 2018. URL: http://blog.kbresearch.nl/2016/01/11/from-keyword-search-to-concept-mining/.

Jacobi, Carina, Wouter van Atteveldt, and Kasper Welbers (2016). “Quantitative analysis of large amounts of journalistic texts using topic modelling”. In: *Digital Journalism* 4.1, pp. 89–106. URL: https://www.tandfonline.com/doi/abs/10.1080/21670811.2015.1093271.

Jacobi, Tineke and Joke Relleke (1987). “Een ’echt Kritiesch Tijdschrift’: 1837-1843”. In: *De Gids sinds 1837. De geschiedenis van een algemeen-cultureel en literair tijdschrift*. Ed. by Remieg A. M. Aerts et al. Amsterdam: Meulenhoff, pp. 8–27. URL: https://www.dbnl.org/tekst/aert010gids02_01/aert010gids02_01_0003.php.

Jautze, Kim, Andreas van Cranenburgh, and Corina Koolen (2016). “Topic modeling literary quality”. In: *Digital Humanities 2016: Conference Abstracts*, pp. 233–237. URL: http://andreasvc.github.io/dh2016.pdf.

Jockers, Matthew L. (Sept. 2011). *The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors*. Accessed 25 January 2018. URL: http://www.matthewjockers.net/2011/09/29/the-lda-buffet-is-now-open-or-latent-dirichlet-allocation-for-english-majors/.

– (2013). *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Jockers, Matthew L. and David Mimno (2013). “Significant themes in 19th-century literature”. In: *Poetics* 41.6, pp. 750–769.

Jurafsky, Daniel and James H. Martin (2017). *Speech and Language Processing*. Third Edition draft. URL: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Kestemont, Mike (2012). “Het gewicht van de auteur”. PhD thesis. Antwerpen: Universiteit Antwerpen, Faculteit Letteren en Wijsbegeerte, Departementen Taal- en Letterkunde.

Kilgore, Jennifer (2008). “Joan of Arc as Propaganda Motif from the Dreyfus Affair to the Second World War”. In: *Revue LISA/LISA e-journal. Littératures, Histoire des Idées, Images, Sociétés du Monde Anglophone–Literature, History of Ideas, Images and Societies of the English-speaking World* 6.1, pp. 279–296. URL: https://journals.openedition.org/lisa/519.

Krop, Henri (1994). "Natuurwetenschap en theologie in de negentiende eeuw. De filosofische achtergrond van de moderne theologie". In: *Theoretische geschiedenis* 21, pp. 16–31. URL: https://www.dbnl.org/tekst/krop001natu01_01/krop001natu01_01_0001.php.

Lauscher, Anne et al. (2016). "Entities as topic labels: combining entity linking and labeled lda to improve topic interpretability and evaluability". In: *IJCol-Italian journal of computational linguistics* 2.2, pp. 67–88. URL: https://ub-madoc.bib.uni-mannheim.de/41843/1/4-lauscher_et_al.pdf.

Liu, Lin et al. (2016). "An overview of topic modeling and its current applications in bioinformatics". In: *SpringerPlus* 5.1, p. 1608. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5028368/.

Long, Hoyt and Richard Jean So (2016). "Literary pattern recognition: Modernism between close reading and machine learning". In: *Critical Inquiry* 42.2, pp. 235–267. URL: https://lucian.uchicago.edu/blogs/literarynetworks/files/2015/12/LONG_SO_CI.pdf.

Lonij, Juliette and Joris van Eijnatten (2016). *Frame Generator*. URL: http://lab.kb.nl/tool/frame-generator.

McAuliffe, Jon D. and David M. Blei (2008). "Supervised topic models". In: *Advances in neural information processing systems*, pp. 121–128. URL: https://arxiv.org/pdf/1003.0783.pdf.

McCallum, Andrew Kachites (2002). "MALLET: A Machine Learning for Language Toolkit". http://mallet.cs.umass.edu.

Meeks, Elijah and Scott Weingart (2012). "The digital humanities contribution to topic modeling". In: *Journal of Digital Humanities* 2.1, pp. 1–6. URL: http://journalofdigitalhumanities org/2-1/dh-contribution-to-topic-modeling/.

Michel, Jean-Baptiste et al. (2010). "Quantitative analysis of culture using millions of digitized books". In: *Science* 331 (6014), pp. 176–182. URL: http://science.sciencemag.org/content/331/6014/176.

Mimno, David (s.d.). *Topic modeling bibliography*. Accessed 4 April 2018. URL: https://mimno.infosci.cornell.edu/topics.html.

Mitterand, Henri (2002). *Zola. L'honneur: 1893-1902*. Vol. 3. Paris: Fayard.

Moody, Christopher E (2016). "Mixing dirichlet topic models and word embeddings to make lda2vec". In: *arXiv preprint arXiv:1605.02019*. URL: https://arxiv.org/pdf/1605.02019.pdf.

Moretti, Franco (2000). "Conjectures on world literature". In: *New left review*, pp. 54–68. URL: https://newleftreview.org/II/1/franco-moretti-conjectures-on-world-literature.

– (2013). *Distant reading*. Verso Books.

Nelson, Robert K. (2010). *Mining the dispatch*. URL: http://dsl.richmond.edu/dispatch/pages/home.

Newman, David J. and Sharon Block (2006). “Probabilistic topic decomposition of an eighteenth-century American newspaper”. In: *Journal of the Association for Information Science and Technology* 57.6, pp. 753–767. URL: http://www.ics.uci.edu/~newman/pubs/JASIST_Newman.pdf.

Peperkamp, Ben (2004). “"Mannekens in de maan" van Nicolaas Beets. Over the moon hoax (1835-1836) en de publieke waardering van de sterrenkunde in de negentiende eeuw”. In: *Nederlandse Letterkunde* 9.2, pp. 101–41. URL: https://www.uu.nl/wetfilos/wetfil05/literatuur/peperkamp2004.pdf.

Read, Piers Paul (2012). *The Dreyfus Affair: The Scandal That Tore France in Two*. New York: Bloomsbury Publishing USA.

Rehurek, Radim and Petr Sojka (May 2010). “Software Framework for Topic Modelling with Large Corpora”. English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, pp. 45–50. URL: http://is.muni.cz/publication/884893/en.

Rhody, Lisa (2012). “Topic modeling and figurative language”. In: *Journal of Digital Humanities* 2.1, pp. 19–35. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/.

Riddell, Allen Beye (2014). “How to read 22,198 journal articles: Studying the history of German studies with topic models”. In: *Distant Readings: Topologies of German culture in the long nineteenth century*, pp. 91–114. URL: https://www.ariddell.org/static/wustl-german-journals-unrevised-proof.pdf.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi (2016). “A model of text for experimentation in the social sciences”. In: *Journal of the American Statistical Association* 111.515, pp. 988–1003. URL: https://www.tandfonline.com/doi/full/10.1080/01621459.2016.1141684.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). “Exploring the space of topic coherence measures”. In: *Proceedings of the eighth ACM international conference on Web search and data mining*. ACM, pp. 399–408. URL: http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf.

Roland, Teddy (2016). *"Topic Modeling: What Humanists Actually Do With It." A Guest Post by Teddy Roland, University of California, Berkeley*. URL: http://digitalhumanities.berkeley.edu/blog/16/07/14/topic-modeling-what-humanists-actually-do-it-guest-post-teddy-roland-university.

Rosen-Zvi, Michal et al. (2004). “The author-topic model for authors and documents”. In: *Proceedings of the 20th conference on Uncertainty in artificial intelligence*. AUAI Press, pp. 487–494. URL: https://arxiv.org/ftp/arxiv/papers/1207/1207.4169.pdf.

Russell, Colin A (2003). “The conflict of science and religion”. In: *The History of Science and Religion in the Western Tradition*. Routledge, pp. 37–42.

Russell, Stuart J. and Peter Norvig (2010). *Artificial intelligence: a modern approach*. 3rd. Upper Saddle River, NJ, USA: Prentice Hall Press, Pearson Education.

Saris, Frans W. and Rob Visser (2005). *Bèta's onder de letterheren*. Vol. 168. Amsterdam: Meulenhoff, pp. 217–348. URL: https://www.dbnl.org/tekst/_gid001200501_01/_gid001200501_01_0025.php.

Schmidt, Benjamin M. (Nov. 2012a). *When you have a MALLET, everything looks like a nail*. Accessed 7 June 2018. URL: http://sappingattention.blogspot.com/2012/11/when-you-have-mallet-everything-looks.html.

– (2012b). "Words alone: Dismantling Topic Models in the Humanities". In: *Journal of Digital Humanities* 2.1. URL: http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/.

Shalit, Uri, Daphna Weinshall, and Gal Chechik (2013). "Modeling musical influence with topic models". In: *International Conference on Machine Learning*, pp. 244–252. URL: http://proceedings.mlr.press/v28/shalit13.pdf.

Sievert, Carson and Kenneth Shirley (2014). "LDAvis: A method for visualizing and interpreting topics". In: *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70. URL: http://www.aclweb.org/anthology/W14-3110.

Snow, Charles Percy (1961). "The two cultures and the scientific revolution: the Rede lecture 1959". In: URL: http://sciencepolicy.colorado.edu/students/envs_5110/snow_1959.pdf.

Stutje, Jan Willem (2014). "Antisemitisme onder Nederlandse socialisten in het fin de siècle". In: *BMGN-Low Countries Historical Review* 129.3, pp. 4–26. URL: https://dspace.library.uu.nl/bitstream/handle/1874/298801/9736-19914-1-PB.pdf?sequence=2.

Tang, Jian et al. (2014). "Understanding the limiting factors of topic modeling via posterior contraction analysis". In: *International Conference on Machine Learning*, pp. 190–198.

Velde, Henk te (2013). "'Van grondwet tot grondwet: oefenen met parlement, partij en schaalvergroting, 1848-1917". In: *Een land van kleine gebaren: een politieke geschiedenis van Nederland 1780-2012*. Ed. by Piet de Rooy. Amsterdam: Boom, pp. 109–194.

Templeton, Clay et al. (2011). "Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus". In: *Chicago Colloquium on Digital Humanities and Computer Science*. Chicago, IL. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.432.3598&rep=rep1&type=pdf.

Theaterencyclopedie contributors (2018). *Anton van Sprinkhuysen*. Accessed: 24 August 2018. URL: https://theaterencyclopedie.nl/wiki/Anton_van_Sprinkhuysen.

Theunissen, Bert (2005). "Gids en Galápagos". In: *De Gids* 168, pp. 219–223. URL: https://www.dbnl.org/tekst/_gid001200501_01/_gid001200501_01_0026.php.

Tjong Kim Sang, Erik et al. (2017). "The CLIN27 Shared Task: Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation". In: *Computational Linguistics in the Netherlands Journal* 7, pp. 53–64. URL: https://dspace.library.uu.nl/handle/1874/360876.

Underwood, Ted (Apr. 2012a). *Topic modeling made just simple enough*. Accessed 25 January 2018. URL: https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/.

– (Nov. 2012b). *Visualizing topic models*. Accessed 30 July 2018. URL: https://tedunderwood.com/2012/11/11/visualizing-topic-models/.

– (2014). "Theorizing research practices we forgot to theorize twenty years ago". In: *Representations* 127.1, pp. 64–72. URL: http://rep.ucpress.edu/content/127/1/64.

Berg, Willem van den and Piet Couttenier (2009). *Alles is taal geworden: Geschiedenis van de Nederlandse literatuur, 1800-1900*. Amsterdam: Bert Bakker.

Bosch, Antal van den et al. (2007). "An efficient memory-based morphosyntactic tagger and parser for Dutch". In: *LOT Occasional Series* 7, pp. 191–206. URL: https://ilk.uvt.nl/downloads/pub/papers/tadpole-final.pdf.

Zwaan, Janneke M. van der, Maarten Marx, and Jaap Kamps (2016). "Validating Cross-Perspective Topic Modeling for Extracting Political Parties' Positions from Parliamentary Proceedings." In: *ECAI*, pp. 28–36. URL: https://e.humanities.uva.nl/publications/2016/zwaa_vali16.pdf.

Atteveldt, Wouter van et al. (2014). *LDA models topics... But what are 'topics'*. URL: http://vanatteveldt.com/wp-content/uploads/2014_vanatteveldt_glasgowbigdata_topics.pdf.

Berkel, Klaas van (1998). "Citaten uit het boek der natuur". In: *Opstellen over Nederlandse wetenschapsgeschiedenis, Amsterdam*, pp. 53–54. URL: https://www.dbnl.org/tekst/berk003cita01_01/.

Berkel, Klaas van, Albert van Helden, and Lodewijk C. Palm (1999). *The History of Science in the Netherlands: Survey, Themes and Reference*. Brill.

Cranenburgh, Andreas van and Rens Bod (2017). "A Data-Oriented Model of Literary Language". In: *CoRR* abs/1701.03329. arXiv: 1701.03329. URL: http://arxiv.org/abs/1701.03329.

Gompel, Maarten van and Martin Reynaert (Dec. 2013). "FoLiA: A practical XML format for linguistic annotation - a descriptive and comparative study". In: *Computational Linguistics in the Netherlands Journal* 3, pp. 63–81. ISSN: 2211-4009. URL: http://www.clinjournal.org/sites/clinjournal.org/files/05-vanGompel-Reynaert-CLIN2013.pdf.

Stipriaan, René van (2009). "At the Speed of Writing, and Beyond: A Short History of the Digital Library of Dutch Literature (DBNL)". In: *Logos* 20.1, pp. 74–78.

Waling, Geerten (2016). *1848 - Clubkoorts en revolutie: democratische experimenten in Parijs en Berlijn*. Nijmegen: Vantilt.

Wallach, Hanna M, David M Mimno, and Andrew McCallum (2009). "Rethinking LDA: Why priors matter". In: *Advances in neural information processing systems*, pp. 1973–1981.

Wang, Xuerui and Andrew McCallum (2006). "Topics over time: a non-Markov continuous-time model of topical trends". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 424–433. URL: https://people.cs.umass.edu/~mccallum/papers/tot-kdd06.pdf.

Weingart, Scott (Nov. 2011). *Topic Modeling and Network Analysis*. Accessed 25 January 2018. URL: http://www.scottbot.net/HIAL/index.html@p=221.html.

– (July 2012). *Topic Modeling for Humanists: A Guided Tour*. Accessed 25 July 2018. URL: http://www.scottbot.net/HIAL/index.html@p=19113.html.

Wesseling, Hendrik Lodewijk (1987). *Vele ideeën over Frankrijk: opstellen over geschiedenis en cultuur*. Amsterdam: Bert Bakker.

Wittek, Peter and Walter Ravenek (2011). "Supporting the Exploration of a Corpus of 17th-Century Scholarly Correspondences by Topic Modeling." In: *SDH 2011 Supporting Digital Humanities: Answering the unaskable*. University of Copenhagen. URL: https://www.clarin.nl/sites/default/files/sdh2011-wittek-ravenek.pdf.

Yang, Min et al. (2018). "A Topic Drift Model for authorship attribution". In: *Neurocomputing* 273, pp. 133–140. URL: https://www.sciencedirect.com/science/article/pii/S0925231217313759.

Yang, Tze-I, Andrew J. Torget, and Rada Mihalcea (2011). In: *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics, pp. 96–104. URL: https://aclanthology.info/pdf/W/W11/W11-1513.pdf.

Zhai, ChengXiang and Sean Massung (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York, NY, USA: Association for Computing Machinery and Morgan Claypool. ISBN: 978-1-97000-117-4.

Zola, Émile (2001). *De waarheid rukt op: Alle teksten over de Dreyfus-affaire*. Vertaald door Henny van Schaik. 's Hertogenbosch: Uitgeverij Voltaire.

Appendix A

Corpus statistics

TABLE A.1: Corpus statistics for the first 100 volumes from *De Gids* (1837-1936), described per year. Given are the number of documents in a volume, the total number of pages per volume and the average number of pages per article, and metrics for the number of words in these documents. The number of words is the number of tokens excluding punctuation as returned by the tokeniser.

	documents	pages		words					
	count	sum	mean	sum	mean	median	min	max	std
1837	174	641	4	445,643	2,561	2,408	69	8,490	1,747
1838	158	673	4	475,201	3,008	2,947	135	11,722	2,131
1839	145	613	4	474,478	3,272	2,621	12	14,385	2,637
1840	128	637	5	445,806	3,483	3,122	59	18,718	2,771
1841	117	657	6	445,096	3,804	3,124	13	12,583	2,966
1842	103	680	7	503,909	4,892	4,203	159	15,846	3,578
1843	111	726	7	510,911	4,603	3,580	234	22,706	3,816
1844	119	746	6	517,694	4,350	3,795	100	16,299	3,101
1845	82	998	12	585,718	7,143	5,872	132	70,491	8,168
1846	73	904	12	577,011	7,904	7,813	153	20,161	4,873
1847	91	1,029	11	580,002	6,374	4,953	159	32,753	5,708
1848	72	812	11	510,799	7,094	5,982	160	20,211	4,920
1849	60	804	13	537,252	8,954	8,278	1,309	25,148	4,783
1850	59	796	13	525,639	8,909	7,770	142	21,128	4,756
1851	71	804	11	580,757	8,180	7,424	422	19,239	4,683
1852	64	836	13	586,333	9,161	7,871	249	22,798	5,772
1853	57	796	14	561,578	9,852	8,640	2,350	22,224	4,875
1854	59	868	15	594,909	10,083	9,597	679	28,642	6,990
1855	52	816	16	584,126	11,233	10,670	661	24,426	5,587
1856	52	888	17	650,914	12,518	12,940	399	22,147	5,331
1857	54	880	16	643,524	11,917	11,460	281	28,640	6,323
1858	59	988	17	671,764	11,386	12,077	204	26,666	7,304
1859	63	909	14	686,993	10,905	9,557	105	28,745	7,831
1860	65	900	14	663,682	10,210	8,508	189	26,794	6,442
1861	63	973	15	673,172	10,685	10,319	728	23,040	5,842
1862	62	1,024	17	785,846	12,675	11,512	2,869	28,512	6,450
1863	80	592	7	885,106	11,064	10,672	92	27,799	6,151

	documents	pages				words			
	count	sum	mean	sum	mean	median	min	max	std
1864	90	632	7	854,557	9,495	9,499	68	23,083	5,313
1865	79	607	8	858,683	10,869	9,749	785	27,684	5,663
1866	85	616	7	848,365	9,981	9,246	415	23,978	4,921
1867	82	592	7	872,278	10,638	10,658	296	24,101	5,458
1868	86	632	7	876,329	10,190	10,098	644	21,087	4,856
1869	67	606	9	867,042	12,941	12,538	1,302	34,545	7,245
1870	69	576	8	789,376	11,440	10,784	1,116	27,004	5,418
1871	68	620	9	851,163	12,517	11,796	444	40,334	6,738
1872	65	584	9	819,771	12,612	12,228	1,440	26,072	6,021
1873	66	617	9	842,472	12,765	12,063	682	35,164	7,084
1874	66	608	9	840,046	12,728	12,178	849	30,543	6,076
1875	79	640	8	851,558	10,779	10,144	1,468	26,148	6,071
1876	72	616	9	847,549	11,772	10,512	1,286	32,971	6,424
1877	89	657	7	925,153	10,395	10,073	369	32,941	5,878
1878	85	648	8	875,733	10,303	9,305	239	30,464	6,059
1879	80	608	8	844,426	10,555	10,346	200	24,026	5,292
1880	78	600	8	828,666	10,624	10,182	241	27,804	6,158
1881	89	624	7	835,408	9,387	9,168	237	25,752	5,705
1882	88	593	7	827,408	9,402	7,982	109	28,769	6,287
1883	99	605	6	789,848	7,978	7,462	278	21,787	4,990
1884	97	617	6	785,947	8,103	7,195	102	24,387	5,552
1885	98	624	6	811,083	8,276	6,471	136	27,850	6,024
1886	114	597	5	812,065	7,123	5,637	119	21,307	4,636
1887	96	597	6	799,275	8,326	6,622	136	26,772	5,636
1888	91	641	7	800,108	8,792	7,044	146	30,134	6,242
1889	94	588	6	814,364	8,663	7,518	103	28,198	6,380
1890	98	611	6	790,359	8,065	6,326	239	31,106	6,458
1891	108	586	5	789,193	7,307	5,088	275	25,558	6,099
1892	100	592	6	770,566	7,706	5,608	418	31,894	6,401
1893	114	611	5	762,568	6,689	4,645	432	30,350	6,054
1894	121	618	5	783,816	6,478	4,786	215	23,575	5,575
1895	129	590	5	789,956	6,124	4,775	315	34,285	5,780
1896	137	603	4	790,617	5,771	4,898	243	18,962	4,557
1897	132	597	5	775,525	5,875	3,965	159	28,656	5,816
1898	129	608	5	776,041	6,016	4,249	251	32,838	5,979
1899	127	605	5	766,904	6,039	4,447	342	29,959	5,405
1900	119	600	5	769,628	6,467	5,240	124	39,867	6,738
1901	138	590	4	739,050	5,355	4,024	107	27,516	5,051
1902	115	590	5	729,145	6,340	5,702	319	22,249	5,115
1903	124	592	5	729,322	5,882	4,686	265	26,344	5,208
1904	122	612	5	732,486	6,004	3,610	227	24,697	5,580
1905	125	611	5	762,031	6,096	4,323	170	29,772	5,525
1906	124	602	5	751,484	6,060	4,344	187	24,277	5,166
1907	138	617	4	730,224	5,291	4,832	128	14,489	3,906
1908	148	623	4	745,521	5,037	2,850	142	24,328	4,989

	documents	pages		words					
	count	sum	mean	sum	mean	median	min	max	std
1909	164	623	4	708,963	4,323	3,068	85	18,919	3,734
1910	177	600	3	724,182	4,091	2,812	107	14,858	3,331
1911	132	619	5	679,804	5,150	2,510	72	27,453	5,720
1912	147	589	4	734,651	4,998	2,662	79	23,133	5,229
1913	137	597	4	741,729	5,414	2,875	118	31,555	6,128
1914	138	594	4	729,585	5,287	4,558	172	31,549	4,782
1915	126	602	5	738,410	5,860	4,075	90	24,997	5,234
1916	149	587	4	718,009	4,819	2,954	60	24,631	4,948
1917	145	585	4	711,862	4,909	2,744	15	18,790	4,905
1918	128	582	5	669,498	5,230	4,241	78	25,095	5,096
1919	120	523	4	622,019	5,183	3,719	65	21,669	5,056
1920	135	520	4	674,195	4,994	3,481	97	21,181	4,886
1921	137	531	4	676,684	4,939	3,363	84	23,012	4,960
1922	145	530	4	681,201	4,698	3,212	79	22,070	4,420
1923	144	521	4	673,206	4,675	3,576	78	16,614	4,206
1924	132	460	3	550,610	4,171	3,034	92	18,551	4,120
1925	133	459	3	521,984	3,925	3,076	69	18,556	3,600
1926	118	446	4	536,663	4,548	3,372	64	19,491	4,184
1927	112	468	4	567,634	5,068	4,038	99	23,320	4,606
1928	131	439	3	538,042	4,107	3,348	48	16,474	3,794
1929	129	453	4	552,794	4,285	3,176	69	15,139	3,857
1930	113	447	4	568,947	5,035	3,231	66	17,836	4,551
1931	153	435	3	569,203	3,720	2,260	37	15,345	3,458
1932	134	446	3	519,558	3,877	2,970	63	22,874	3,913
1933	130	492	4	531,533	4,089	2,226	26	42,277	5,596
1934	125	411	3	477,588	3,821	2,777	64	17,130	3,884
1935	134	420	3	494,481	3,690	2,624	73	21,042	3,446
1936	143	442	3	560,127	3,917	3,246	43	21,053	3,739
Total	10,624	63,814		69,430,134					

Appendix B

Stop Words

1 n
2 't
3 t
4 a
5 aan
6 aannemen
7 aantal
8 ach
9 achter
10 af
11 al
12 aldaar
13 aldus
14 algemeen
15 alleen
16 allerlei
17 alles
18 als
19 alsof
20 althans
21 altijd
22 alzoo
23 and
24 ander
25 andere
26 anders
27 b
28 bv
29 beginnen
30 behalve
31 behoeven
32 beide
33 ben
34 bepaald
35 bereiken
36 besluiten
37 bestaan
38 betreffen
39 bezitten
40 bij
41 bijna
42 binnen
43 bl
44 bladzijde
45 blijkbaar
46 blijken
47 blijven
48 blz
49 blz.
50 boven
51 bovenal
52 bovendien
53 by
54 c
55 ce
56 di
57 daar
58 daaraan
59 daarbij
60 daardoor
61 daarentegen
62 daarin
63 daarmede
64 daarna
65 daarom
66 daarop
67 daarover
68 daartoe
69 daarvan
70 dadelijk
71 dag
72 dan
73 dat
74 de
75 deel
76 deelen
77 den
78 denken
79 der
80 dergelijk
81 derhalve
82 des
83 deze
84 dezelfde
85 dezer
86 dicht
87 die
88 dienen
89 dit
90 doch
91 doen
92 door
93 drie
94 du
95 duidelijk
96 duizend
97 dus
98 echt
99 echter
100 een
101 eene
102 eenen
103 eener
104 eenig
105 eenigen
106 eenmaal
107 eens
108 eenvoudig
109 eerst
110 eigen
111 eigenlijk
112 eind
113 eindelijk
114 elk
115 elkaar
116 elkander
117 en
118 ende
119 enkel
120 enz
121 er
122 erg
123 ernstig
124 est
125 est
126 et
127 even
128 evenals
129 evenmin
130 evenwel
131 f
132 fijn
133 gaan
134 gaarne
135 gansch
136 ge
137 gebeuren
138 gedurende
139 geen
140 geene
141 geenszins
142 geheel
143 geheele
144 gelijk
145 gemakkelijk
146 genoeg
147 gering
148 geval
149 geven
150 geweest
151 gewoon
152 geworden
153 gij
154 goed
155 groot
156 groote
157 grooten
158 grooter
159 haar
160 had
161 half
162 hand
163 harer
164 heb
165 hebben
166 heeft
167 heel
168 heen
169 hem
170 hen
171 het
172 hetgeen
173 hetwelk
174 hetzelfde
175 hetzij
176 hier

177 hierbij
178 hiermede
179 hierop
180 hiervan
181 hij
182 hoe
183 hoewel
184 hoog
185 hooge
186 hooger
187 houden
188 hun
189 hunne
190 i
191 ieder
192 iemand
193 iets
194 ii
195 iii
196 ik
197 il
198 immer
199 immers
200 in
201 inderdaad
202 indien
203 intusschen
204 is
205 iv
206 ja
207 je
208 jegens
209 juist
210 kan
211 klein
212 komen
213 kon
214 kort
215 krachtig
216 krijgen
217 kunnen
218 la
219 laag
220 laat
221 land
222 lang
223 langs
224 laten
225 le
226 leeren
227 leven
228 liever
229 liggen
230 maar
231 maken
232 me
233 mede
234 mee
235 meer
236 meest
237 meestal
238 men
239 menen
240 menig
241 met
242 mgen
243 midden
244 mij
245 mijn
246 min
247 minder
248 minst
249 mits
250 moet
251 moeten
252 mogelijk
253 mogen
254 mooi
255 my
256 na
257 naar
258 naarmate
259 naast
260 nadat
261 namelijk
262 natuurlijk
263 neen
264 nemen
265 nergens
266 niemand
267 niet
268 niets
269 nimmer
270 noch
271 noemen
272 nog
273 noodig
274 nooit
275 nu
276 o
277 oa
278 of
279 ofschoon
280 om
281 omdat
282 omtrent
283 ondanks
284 onder
285 ongeveer
286 ons
287 onzer
288 oog
289 oogenblik
290 ooit
291 ook
292 op
293 opdat
294 opzicht
295 over
296 overal
297 overig
298 p
299 paar
300 pag
301 par
302 pas
303 per
304 que
305 qui
306 r
307 reeds
308 ruim
309 s
310 schijnen
311 schoone
312 sedert
313 sinds
314 slecht
315 slechts
316 sommig
317 soms
318 soort
319 spoedig
320 ss
321 staan
322 steeds
323 stellen
324 stil
325 te
326 tegen
327 tegenover
328 telkens
329 tenzij
330 terug
331 terwijl
332 tevens
333 thans
334 the
335 tien
336 tijd
337 to
338 toch
339 toe
340 toen
341 tooneel
342 tot
343 totdat
344 treffen
345 trouwens
346 tusschen
347 twee
348 u
349 uit
350 un
351 und
352 une
353 uw
354 v
355 vaak
356 vallen
357 van
358 van_de
359 van_den
360 vandaar
361 veel
362 veelal
363 veeleer
364 ver
365 vervullen
366 vier
367 vijf
368 vinden
369 vol
370 volgen
371 volgend
372 volgens
373 volkomen
374 voor
375 vooral
376 voort
377 vrij
378 waar
379 waaraan
380 waarbij
381 waardoor
382 waarin
383 waarmede
384 waarmee
385 waarom
386 waarop
387 waarover
388 waartoe
389 waaruit
390 waarvan
391 waarvoor
392 wanneer
393 want
394 waren
395 was
396 wat
397 we
398 weder
399 weer
400 weg
401 weinig
402 wel
403 weldra
404 welk
405 wellicht
406 welnu
407 werd
408 weten
409 wezen
410 wie
411 wien
412 wier
413 wij
414 wijze
415 wil
416 willen

417 worden
418 wordt
419 zacht
420 zal
421 ze
422 zeer
423 zeggen
424 zeker
425 zelf
426 zelfs
427 zelve
428 zelven
429 zes
430 zetten
431 zich
432 zichzelf
433 ziedaar
434 zien
435 zij
436 zijn
437 zijner
438 zijns
439 zitten
440 zo
441 zonder
442 zoo
443 zooal
444 zoodanig
445 zoodat
446 zoodra
447 zoolang
448 zooveel
449 zoover
450 zoowel
451 zoozeer
452 zou
453 zulk
454 zullen
455 zwaar
456 zwak

Appendix C

Experiment 1: 30 topics from 1848

The 30 topics below are each represented by 30 words that are coming from two types of corpora: (i) a filtered corpus and (ii) a corpus in which entities (e.g. persons, things) are represented as a single term. If possible, the most distinctive document from the corpus is given for the topic, together with the document id (for easy lookup in the DBNL, and the probability of the topic in the document.[1] The topics are sorted on their proportion in the corpus and this proportion is also given. Both corpus states are given to show the difference in interpretability of the preprocessing step that adds entities as single terms to the documents.

C.1 Filtered corpus (30 topics)

Topic 20: minister volk koning staat vergadering frankrijk zaak lord regering vorst partij lid leger bewind engeland nationaal invloed hoofd europa duitsche parlement fransch republiek beginsel italie troep maand oostenrijk oorlog poging
Proportion: 0.2085466
Document: *Staatkundig overzigt.*
Document id: _gid001184801_01_0071 **Probability:** 0.9980989

Topic 24: heer paulus schrijver naam costa da werk kaart brief god d schrijven abarim stad num moab christus gebergte lezen leger voorkomen israelit voorstelling bewijzen bepalen nebo woestijn punt grond arnon
Proportion: 0.12033258
Document: *M r . da Costa op het gebied der Godgeleerdheid.*
Document id: _gid001184801_01_0057 **Probability:** 0.9998452

Topic 0: wetenschap gesticht krankzinnig observatorium hansen von gotha natuur waarneming stad seeberg sterrekunde les gebouw hertog reis onderzoeking kennen uur zach vraag voeren dier werktuig geest invloed kennis stichten leiden verschillend
Proportion: 0.1000671
Document: *Herinneringen uit Zwaben en Franken. (1846.)*
Document id: _gid001184801_01_0031 **Probability:** 0.99975044

Topic 3: volk vorst koning stad oud hart hoofd eeuw aartspriester flora magt huis geestelijk rijk geld vergeten dood don roepen dier arm liefde rust naam oorlog regt zee schoon ranald kracht
Proportion: 0.07804272

[1] This fails if the probability for a topic for a document is too low (<0.01).

Document: *Westminster en St. Pauls.*
Document id: _gid001184801_01_0003 **Probability:** 0.9995349

Topic 16: lid klassikaal kerk synode vergadering koning commissie provinciaal vertegenwoordigen kerkbestuur bestuur romance gemeente oud benoemen stem hervormen art kerkelijk synodaal predikant kiezen rodrigo regt benoeming graaf vertegenwoordiging vader regl ouderling
Proportion: 0.07255067
Document: *De Vertegenwoordiging der Nederlandsche Hervormde Kerk door de Synode*
Document id: _gid001184801_01_0047 **Probability:** 0.9998704

Topic 4: arbeid stelsel kapitaal brood prijs arm belang vrijheid bestuur gelden zaak volk heer maatschappij kracht regelen beginsel algemeene grondslag regeling staat doel wet leggen middel winst mededinging vraag grond stedelijk
Proportion: 0.06615815
Document: *De buitengewone Armenbedeeling te Amsterdam in 1847.*
Document id: _gid001184801_01_0017 **Probability:** 0.9984963

Topic 23: gevangenis stelsel gevangene opsluiting gevangen eenzaam krankzinnigheid straf arbeid da pennsylvanisch behoefte maatschappij vorm costa invloed kracht zin geschiedenis cel oratorium kunst geest oud beginsel mendelssohn daarenboven voorkomen vrijheid ontwikkeling
Proportion: 0.06172863
Document: *Het Pennsylvanische Gevangenisstelsel.*
Document id: _gid001184801_01_0048 **Probability:** 0.9998869

Topic 14: heer huell frankrijk schimmelpenninck oud fransch lied napoleon keizer republiek eeuw huis werk prins volk schrijven oranje hall hart tractaat holland koning louis hoofd vrede no bekend mei vaderland onderhandeling
Proportion: 0.049112853
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0041 **Probability:** 0.9994741

Topic 2: geregtelijk onderzoek geneeskunde wetenschap geneeskundig regter kennis opvatting krankzinnigheid regtsgeleerde geneeskundige zaak begrip toepassing beschouwen ontwikkeling toestand erkennen onderwerp mensch wet bijzonder uitspraak vraag werk geneeskundigen vorderen bepaling von punt
Proportion: 0.040733352
Document: *Geregtelijke Geneeskunde.*
Document id: _gid001184801_01_0002 **Probability:** 0.9769219

Topic 18: java taal volk javaansch oud sanskrit roth prof vorst geschiedenis egyptisch eiland voorstelling werk rijk godsdienst invloed godheid bekend wereld eeuw begrip philosophie indie indische javanen archipel overlevering saka god
Proportion: 0.03693816
Document: *De oudste bronnen van onze Metaphysische begrippen.*
Document id: _gid001184801_01_0014 **Probability:** 0.9972537

Topic 28: naam heer verkiezing baalbek senden schrijver hermon stad heliopolis libanon egyptisch baalath waarschijnlijk meening voorkomen gebergte gevoelen salomo vreede zon oud djebel baalgad on dal jos palestina noordelijk werk werkelijk
Proportion: 0.030915426
Document: *Bijbelsche Aardrijkskunde.*
Document id: _gid001184801_01_0064 **Probability:** 0.99984574

Topic 9: indie tijdschrift heer besluit kolonie koloniaal genootschap aangelegenheid nederlandsch kennis belang publiek drukpers gouvernement hoevell stuk moederland regering redactie uitgave drukken verslag openbaarheid neerlandsch invloed drukkerij toestand indische beginsel bepaling
Proportion: 0.028658776

Document: *De openbaarheid in koloniale aangelegenheden.*
Document id: _gid001184801_01_0044 **Probability:** 0.81724477

Topic 25: maatschappij arbeid mensch beginsel wet stelsel kapitaal eigendom denkbeeld middel wetgever winst leering regering gevolg maatschappelijk klas staat zaak menschelijk vrijheid volk regelen waarheid nijverheid ontwikkeling vrucht wijsheid natuur arbeiden
Proportion: 0.027567208
Document: *Frédéric Bastiat en de Socialisten.*
Document id: _gid001184801_01_0053 **Probability:** 0.89652526

Topic 11: compagnie schip vaart octrooi usselincx westindische holland vermelden nieuwnederland eiland staten meteren naam statengeneraal koopman westindien amsterdam kust o'callaghan handel rivier hendrik schipper vereenigde block zaak spanje hudson hy heeren
Proportion: 0.022139644
Document: *Bijdragen tot de geschiedenis onzer Kolonisatie in Noord-Amerika.*
Document id: _gid001184801_01_0063 **Probability:** 0.9998262

Topic 13: stand geneeskundig taal leger eed officier onderwijs wet hoogl regering schrijver wetenschap trouw vrijheid latijn geneesheer gevoelen gehoorzaamheid koning stad geneeskundigen studie reden grond volk voorstellen hoogeschool gevoel opstand voeren
Proportion: 0.021764064
Document: *De verbetering van ons Geneeskundig Onderwijs.*
Document id: _gid001184801_01_0067 **Probability:** 0.99966836

Topic 19: belasting inkomst wet last betalen belastingstelsel handel regten inkomen pct accijnsen schatkist consumtie administratie bezwaar grondwetsherziening kapitaal vraag accijns volk klagen welvaart zout fabrijken drukken registratie millioen spoorweg scheepvaart kost
Proportion: 0.012917889
Document: *De Grondwets-herziening en ons Belastingstelsel.*
Document id: _gid001184801_01_0054 **Probability:** 0.37852752

Topic 1: republiek parijs rome bevolking stad koning hoofdstad frankrijk eeuw misschien wereld geschiedenis staat hoofd monarchie zetel magt oud vrijheid volk weelde europa hart staatkundig behoefte magtig taal willem bloed geest
Proportion: 0.009145592
Document: *Twee Gedichten van Piet Bogcheljoen .*
Document id: _gid001184801_01_0023 **Probability:** 0.9974287

Topic 15: stad kastilie vorst aragon koning cortes adel geschiedenis spanje rijk belasting eeuw kastieljaansch burger don letterkunde cronica strijd wet geestelijkheid arabieren mooren volstrekt verklaren strijden middeleeuwen el belang betalen fuero
Proportion: 0.0087781325
Document: *Kastieljaansche Letterkunde in de Middeleeuwen.*
Document id: _gid001184801_01_0010 **Probability:** 0.9747667

Topic 17: heer da costa strijd groningsche groningers brochure dc grond onderscheiden meening dikwijls hoogleeraar berigte christelijk stilzwijgen godgeleerde latijnsch misverstand schrijven partij zoeken denkbeeld waarheid karakter gelden paulus overtuiging antwoord gevoel
Proportion: 0.0025579578
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0019 **Probability:** 0.99851066

Topic 8: elders verhongren hart zegen lijden troost verheugen kennen heeren vrouw brood hemel 'k kaak teug dronkenschap smaken koude hoeden strooien kunst schrijven rijk schenken werk recht getroosten danken strijd lusten
Proportion: 0.0008086911
Document: *Eene droevige Gedachte.*
Document id: _gid001184801_01_0011 **Probability:** 0.9951177

Topic 12: licht uil geest lief nacht wiek zonne almacht kracht grens werken slaan horen wachten sluiten missen genieten vrezen gods morgen bloei ginds verbreken verdwijnen ontspruiten oor ruimte blind zegen beneden
Proportion: 0.00044929926
Document: *Licht.*
Document id: _gid001184801_01_0059 **Probability:** 0.9915465

Topic 6: droppel benoodigd huit befaamd parti mondig stelsel beginsel wet maatschappij arbeid staat volk zaak mensch vrijheid eigendom koning werk stad heer lid middel naam belasting oud denkbeeld dier kapitaal gevolg
Proportion: 4.053785e-05

Topic 21: lid klassikaal vergadering synode commissie bestuur costa zaak heer kerk stem provinciaal kerkbestuur koning da oud paulus volk wet vertegenwoordigen naam schrijver gemeente werk wetenschap stelsel belang bepalen beginsel stad
Proportion: 5.7935663e-06

Topic 7: koning volk heer oud taal zaak vorst naam werk stad eeuw hoofd geschiedenis staat invloed minister vergadering lid frankrijk dier god kennen wet geest schrijven beginsel vrijheid belang verschillend schrijver
Proportion: 5.7911134e-06

Topic 26: volk wetenschap wet invloed schrijven ontwikkeling zaak koning staat heer werk oud naam taal dier grond lid mensch stad god schrijver stelsel vorst geschiedenis kennis kennen geest bepalen von arbeid
Proportion: 5.76822e-06

Topic 5: oud stelsel werk heer volk arbeid staat geschiedenis schrijver naam belang stad koning kracht frankrijk wetenschap middel buiten beginsel kennen wet mensch erkennen zaak grond vorst kapitaal von gevangenis lezen
Proportion: 5.7367247e-06

Topic 10: naam lid heer zaak vergadering bestuur koning stad klassikaal wetenschap commissie schrijven synode staat werk kerk stelsel oud schrijver beginsel frankrijk da bekend wet belang kennis volk invloed regt arbeid
Proportion: 5.7315983e-06

Topic 29: naam heer zaak stad frankrijk wetenschap volk arbeid oud kapitaal kennis wet werk invloed stelsel dier beginsel doel belang vrijheid regering mensch toestand staat vraag ontwikkeling maatschappij von buiten kennen
Proportion: 5.7277753e-06

Topic 27: heer volk naam stelsel oud zaak stad wet invloed koning regering belang beginsel staat grond werk vorst wetenschap geschiedenis gevangenis kennis mensch vorm kracht eeuw middel vrijheid schrijver toestand vraag
Proportion: 5.712666e-06

Topic 22: naam heer oud zaak stad werk leger volk koning eeuw grond von invloed schrijven staat wetenschap vorst lid erkennen frankrijk schrijver geest bekend beginsel vergadering verklaren dier geschiedenis java bijzonder
Proportion: 5.7121747e-06

C.2 Filtered corpus with entities as single terms (30 topics)

Topic 18: koning volk vorst minister frankrijk staat vergadering stad zaak regering partij lord lid engeland hoofd europa oud invloed nationaal naam bewind duitschland belang maand rijk duitsche italie oorlog eeuw poging
Proportion: 0.27818778
Document: *Staatkundig overzigt.*
Document id: _gid001184801_01_0040 **Probability:** 0.9998089

Topic 15: arbeid belasting kapitaal stelsel maatschappij mensch beginsel wet denkbeeld volk vrijheid eigendom middel regten winst paulus nijverheid zaak staat werk kracht costa waarheid da schrijven belang gevolg schrijver last betalen
Proportion: 0.09158891
Document: *Frédéric Bastiat en de Socialisten.*
Document id: _gid001184801_01_0053 **Probability:** 0.97254735

Topic 14: leger holland vesting besluit vijand kolonie koloniaal belang indie schip verdediging kennis eiland aangelegenheid zaak drukpers middel gewest stuk vaart sterkte toestand regering gouvernement java naam gevolg vijandelijk heer moederland
Proportion: 0.071111515
Document: *Bijdragen tot de geschiedenis onzer Kolonisatie in Noord-Amerika.*
Document id: _gid001184801_01_0063 **Probability:** 0.99981487

Topic 25: gevangenis gesticht stelsel krankzinnig brood gevangene krankzinnigheid opsluiting gevangen eenzaam zetting bakker straf prijs behoefte bestuur arbeid pennsylvanisch vrijheid maatschappij verbetering getal wet besluit invloed belang broodzetting cel pct ontstaan
Proportion: 0.0651802
Document: *Het Pennsylvanische Gevangenisstelsel.*
Document id: _gid001184801_01_0048 **Probability:** 0.9998883

Topic 11: god mensch werk geest hart opvoeding ontwikkeling geschiedenis volk gods dier voorstelling denkbeeld school openbaring schoon menschdom schrijver heidenen voorlezing zin waarheid paulus kracht wereld aarde dood christus storm bezigen
Proportion: 0.059871312
Document: *De Geschiedenis van de Opvoeding des Menschdoms en de Openbaring des Bijbels.*
Document id: _gid001184801_01_0058 **Probability:** 0.99982953

Topic 9: koning romance schrijver oud graaf vader rodrigo leger eeuw heer vrijheid officier regering eed gonzalo werk jong regel vreemd vorst trouw volk spanje gedicht stuk vers gehoorzaamheid wet broeder kastilie
Proportion: 0.056085285
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0051 **Probability:** 0.999583

Topic 7: geneeskundig wetenschap onderzoek geregtelijk kennis regter zaak opvatting geneeskundige wet regtsgeleerde krankzinnigheid erkennen vorderen vraag geneeskunde toestand bijzonder geneeskundigen onderwerp beschouwen begrip toepassing ontwikkeling geregtelijk_geneeskunde punt bepaling uitspraak algemeene vorm
Proportion: 0.053440467
Document: *Geregtelijke Geneeskunde.*
Document id: _gid001184801_01_0002 **Probability:** 0.99976957

Topic 17: kaart naam num heer stad woestijn gebergte moab abarim ligging senden israelit xxi berg nebo beek bruyn palestina grens punt arnon werk vermelden zered schrijver deut noorden register dibon pisga
Proportion: 0.039812263

Document: *Bijbelsche Aardrijkskunde.*
Document id: _gid001184801_01_0007 **Probability:** 0.9973802

Topic 21: republiek frankrijk ver_huell parijs keizer fransch napoleon schimmelpenninck prins heer bevolking hall hoofd tractaat sqq rome oranje stad zaak onderhandeling koning mei huis oud commissie natie artikel schadeloosstelling admiraal vaderland
Proportion: 0.034900945
Document: *Karel Hendrik Ver Huell en Rutger Jan Schimmelpenninck.*
Document id: _gid001184801_01_0001 **Probability:** 0.99983585

Topic 6: observatorium hansen gotha wetenschap sterrekunde arm reis von_zach werktuig seeberg waarneming hertog sterrekundig stad commissie gebouw sterrekundige inrigting bekend uur berlijn kennis woning verschillend leiden bijzonder zaak leipzig buitengewoon wetenschappelijk
Proportion: 0.03301221
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0072 **Probability:** 0.9993474

Topic 16: wetenschap verkiezing natuur vraag kennen regtstreeksch geest onderzoeking god voorbeeld schrijver les verschijnsel mensch wijsbegeerte feit oorzaak kennis waarneming verklaren onderzoek invloed ontwikkeling volkswil antwoord heer_kemper tijdperk geschiedenis toepassing stelsel
Proportion: 0.030783772
Document: *De Wetenschap der Natuur, haar doel en de wijze van haar te beoefenen.*
Document id: _gid001184801_01_0042 **Probability:** 0.97885674

Topic 2: lid klassikaal synode vergadering commissie provinciaal kerkbestuur vertegenwoordigen kerk bestuur gemeente benoemen art synodaal stem koning kerkelijk kiezen benoeming hervormen_kerk vertegenwoordiging predikant regl regt alg ouderling vertegenwoordiger reglement kerkeraad afvaardigen
Proportion: 0.028903047
Document: *De Vertegenwoordiging der Nederlandsche Hervormde Kerk door de Synode*
Document id: _gid001184801_01_0047 **Probability:** 0.9998707

Topic 0: naam heer oud baalbek senden lied schrijver stad hermon werk bekend libanon waarschijnlijk heliopolis voorkomen baalath eeuw gebergte salomo egyptisch meening zon gevoelen baalgad jos noordelijk palestina dal werkelijk no
Proportion: 0.028163332
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0041 **Probability:** 0.9994605

Topic 10: java taal javaansch sanskrit volk vorst eiland rijk geschiedenis indie javanen invloed eeuw indische bekend overlevering oud adji_saka betrekking krama archipel werk beschaving uitdrukking ngoko magt maleisch crawfurd vast vreemd
Proportion: 0.024163418
Document: *Iets over het Javaansch en de oudste geschiedenis der Javanen .*
Document id: _gid001184801_01_0043 **Probability:** 0.9998449

Topic 22: brief baur paulus schrijven ef kritiek geschrift schrijver schwegler christus tubinger voorstelling standpunt voorkomen kerk echtheid apostel bewijzen christendom ta efezerbrief werk strauss betrekking gnostisch kennen plaatsen christelijk tes kor
Proportion: 0.02234359
Document: *De echtheid van den Brief aan de Efeziërs bestreden en verdedigd.*
Document id: _gid001184801_01_0062 **Probability:** 0.9997672

Topic 28: aartspriester geestelijk geld alcalde regt regter wetboek non dichter fueros copla y schrijven liefde wet geestelijkheid koning spanje eeuw exceptio advokaten vonnis spaansch eisch vriend rijk advokaat dief fuero del
Proportion: 0.018366382

Document: *Kastieljaansche Letterkunde in de Middeleeuwen.*
Document id: _gid001184801_01_0036 **Probability:** 0.7695369

Topic 20: prof roth egyptisch voorstelling oud godheid wereld volk begrip semitisch godsdienststelsel werk godsdienst volksstam phoeniciers oorspronklijk god bron bewijs metaphysisch taalstam verschillend beschouwen oorsprong zoroaster mozaisch geschiedenis zelfstandig naam eigendommelijk
Proportion: 0.012458576
Document: *De oudste bronnen van onze Metaphysische begrippen.*
Document id: _gid001184801_01_0014 **Probability:** 0.99969333

Topic 19: stand taal onderwijs hoogl latijn stad gevoelen reden grond oud platteland hoogeschool gemeente geleerde aanvoeren geneesheer hoogleeraar studie heer_da_costa doctor schrijver stelsel geneesheeren latijnsch plattelandsheelmeester wetenschap moeijelijk geschikt student zoeken
Proportion: 0.011743373
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0019 **Probability:** 0.9984553

Topic 24: schrijver verbindtenis overeenkomst art elle trekker handel wissel prediker les koopman w preek wet nemer formulier lumiere koophandel ses gelegenheid ai k rotterdam werk voorkomen waarde verstaan jesuschrist persoon daad
Proportion: 0.009564674
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0061 **Probability:** 0.8369005

Topic 26: schrijver geschiedenis werk ruggevat buiten dier bloedsomloop verhandeling vorm feit gedeelte werking onderzoek schlosser hart begrip zin ijs denkbeeld dr gloed dans jeugd geest lezen zijde waarheid gevoelen natuur verschillend
Proportion: 0.008368685
Document: *Bibliographisch album.*
Document id: _gid001184801_01_0066 **Probability:** 0.99894196

Topic 4: stad kastilie vorst aragon koning cortes adel geschiedenis rijk spanje belasting eeuw burger cronica kastieljaansch geestelijkheid wet strijd fuero arabieren los middeleeuwen strijden betalen volstrekt regt verklaren oorlog belang mooren
Proportion: 0.008286728
Document: *Kastieljaansche Letterkunde in de Middeleeuwen.*
Document id: _gid001184801_01_0010 **Probability:** 0.9302082

Topic 27: oratorium mendelssohn dramatisch toonkunst goethe kunst elias lyrisch voornamelijk vorm innig gelukkig leipzig zelter muzijkaal element vriend voortreffelijk twijfel ontwikkelen tekst kennen componist gewaarwording felix opera voorspelling muzijk rome lied
Proportion: 0.0047220783
Document: *Kunst. Mendelssohn Bartholdy. Oratorium: Elias.*
Document id: _gid001184801_01_0022 **Probability:** 0.38963717

Topic 23: koning hart bundel kind warnasarie neerlandsch_indie gedicht jaarboekje opvoeding redactie misschien moederland holland kracht liefde poezij kunst publiek meenen geschiedenis indie och bloed los lief volk ouder diep lied merkwaardig
Proportion: 0.004640848
Document: *Twee Gedichten van Piet Bogcheljoen .*
Document id: _gid001184801_01_0023 **Probability:** 0.99748725

Topic 13: volk koning hoor sluiten vorst verjongen borst puin vrijheid troonen herinring naam vaderland eer holland vrede frankrijks van_orleans balling gloren heil volksbestaan hollands kreet vloek grijs juichen breken woning val
Proportion: 0.0023031875
Document: *De weduwe van Orleans.*
Document id: _gid001184801_01_0033 **Probability:** 0.9972948

Topic 12: kapitaal huskisson spoorweg papier reciprociteit standaard belegging pond vereenigde_staten protectionist houde rente crisis navigatieacte charles_wood granby natien ongelegenheid geldsgebrek bankacte aartshertogpalatijn bedaard credietpapier metalen speculatie ijzingwekkend amerikanen financiewezen mr_herries aansprakelijk
Proportion: 0.0009130906
Document: *Staatkundig overzigt.*
Document id: _gid001184801_01_0006 **Probability:** 0.05373676

Topic 1: elders verhongren hart zegen lijden kennen heeren brood hemel vrouw 'k verheugen troost werk gevoelen licht rijk schrijven vreugd zon drukken hoeden strijd zoeken gebrek nader kunst ontbreken regel toekomst
Proportion: 0.00078694214
Document: *Eene droevige Gedachte.*
Document id: _gid001184801_01_0011 **Probability:** 0.9952567

Topic 5: ernst_ii tafelen nachtrust achtereen hartelijkheid thuringer voorst middelst toereikende beroemdheid luchtstroomen van_beeck_calkoen dorpat wiesbaden allgemeine_geographische_ephemeriden overvoeren nicolai afmatten kluchtig onthouding koeijen toestel mobius rumker observatorium hansen gotha sterrekunde sterrekundige van_de_seeberg
Proportion: 0.00027933644

Topic 8: vereischen van_de_nederlandschen indeeling naam kaart stad gebergte moab heer werk num senden woestijn palestina bruyn grens ligging schrijver wetenschap abarim israelit berg xxi punt deut nebo beek bepaling bepalen bekend
Proportion: 6.1070355e-06

Topic 3: koning stad naam zaak wetenschap arbeid staat lid beginsel wet heer vorst vrijheid volk schrijver regering grond vergadering werk belang geschiedenis erkennen kennen frankrijk stelsel mensch lezen oud kennis gevolg
Proportion: 5.98467e-06

Topic 29: oud volk naam koning taal staat stelsel zaak werk schrijven grond invloed erkennen wetenschap wet heer vorm lezen frankrijk verklaren geest lid stad bestuur schrijver vrijheid dier onderzoek kennen kracht
Proportion: 5.942186e-06

Appendix D

Experiment 2: 50 topics from 1894-1906

The topic word clusters below are a result of a qualitative inspection of the topic word distributions that were generated by a model of 50, 60, 100 and 150 topics. The results below show the most coherent topics: i.e. topics that can easily be interpreted, that show little to no overlap with other topics, and are not redundant in the total overview. The formatting is the same as in the previous experiment (Appendix C), so that with every topic the most distinctive document is given, together with the topic's probability in this document. This might help interpreting the clusters. The topics are sorted on their proportion in the corpus.

D.1 50 topics

Topic 35: mensch boek jong kunst vrouw voelen wereld schrijven liefde kind gevoel ding kennen kracht geest ziel lezen diep begrijpen natuur zoeken schrijver hart werken volk buiten gevoelen dichter misschien waarheid
Proportion: 0.10309072
Document: *Driemaandelijksch letterkundig overzicht.*
Document id: _gid001190201_01_0112 **Probability:** 0.87059647

Topic 11: vrouw kijken ooog voelen vragen hoofd kind huis jong horen nou lopen arm gezicht roepen stem trekken jij tafel deur lachen moeder begrijpen blik mevrouw donker mensch zwart wachten slaan
Proportion: 0.08721595
Document: *De doode.*
Document id: _gid001190101_01_0036 **Probability:** 0.99391496

Topic 49: onderwijs school wet leerling dienst belang toestand leger verschillend zaak bijzonder minister regeering nederland kind inrichting art vak commissie ambtenaar verkrijgen militair onderwijzer regeling eischen staat leeraar bestuur recht openbaar
Proportion: 0.076322615
Document: *Het eindexamen der gymnasia.*
Document id: _gid001190201_01_0074 **Probability:** 0.98416233

Topic 26: kerk vrouw naam eeuw vader zoon god volk stad huis kind jong mensch kennen schrijven geest dragen bekend dood moeder koning dier priester recht hoofd hart rome broeder arm rijk
Proportion: 0.06820959

Document: *Roomsche woorden.*

Document id: _gid001190101_01_0073 **Probability:** 0.90037763

Topic 5: muziek opera componist werken kunst beethoven schrijven muzikaal meester publiek orkest uitvoering wagner compositie symphonie lied dramatisch concert drama gebied opvoering theater duitsche mozart liszt parijs vorm stuk duitschland vriend

Proportion: 0.06608646

Document: *Muzikaal overzicht.*

Document id: _gid001190301_01_0020 **Probability:** 0.9993837

Topic 21: ziel hart kind zingen lief ooog liefde god nacht dood bloem blank arm voelen horen aarde lied zon hemel diep zee lucht moeder blij stem traan vragen mensch hoofd sterven

Proportion: 0.05120194

Document: *Van de kristallen torens.*

Document id: _gid001190101_01_0115 **Probability:** 0.9991674

Topic 25: recht partij politiek regeering minister heer wet belang kiesrecht liberaal beginsel grondwet meerderheid kamer strijd vraag staat ontwerp volk zaak stem macht lid kind parlementair vrijheid stelsel sociaal verkiezing zijde

Proportion: 0.03740215

Document: *Parlementaire kroniek.*

Document id: _gid001190401_01_0101 **Probability:** 0.9982389

Topic 19: schrijven brief boek schrijver lezen vriend potgieter heer huet gids stuk jong geschiedenis studie hart fruin oordeel letterkundig tijdschrift kennen amsterdam misschien herinnering dr kennis mevrouw redactie naam verschijnen nederlandsche

Proportion: 0.036561742

Document: *Bibliographie.*

Document id: _gid001190301_01_0023 **Probability:** 0.99533725

Topic 39: wetenschap mensch begrip kennis voorstelling ding geest verschijnsel gewaarwording vraag feit onderzoek plato verschillend bewustzijn kennen wereld theorie vorm leren beteekenis waarneming ziel voorwerp ontwikkeling geestelijk socrates beweging oorzaak waarheid

Proportion: 0.036295604

Document: *Modern positivisme.*

Document id: _gid001190401_01_0095 **Probability:** 0.9982198

Topic 31: vers dichter les gedicht fransch schrijven poezie schrijver kunst naam jong boek lied vorm zoeken eeuw paris dans indruk misschien taal frankrijk drama geschiedenis beeld dood studie regel plus vroeg

Proportion: 0.035017084

Document: *Fransche symbolisten .*

Document id: _gid001190201_01_0054 **Probability:** 0.9997195

Topic 2: volk boer engelsch zuidafrika engelschen zaak bevolking oorlog trekken naam kolonie regeering eiland europa hoofd vroeg spanje president japan recht generaal invloed rijk hollandsche russisch engeland gouverneur eeuw taal vijand

Proportion: 0.03480636

Document: *De rol van den Oranje-Vrijstaat in den oorlog in Zuid-Afrika.*

Document id: _gid001190101_01_0059 **Probability:** 0.90273774

Topic 48: water wit zee lucht meisje groen berg jongen donker stad mensch uur kleur jong huis kijken beneden voet heerlijk vlak diep boomen zwart gezicht hoofd bosch lief meter zoo'n blauw

Proportion: 0.0342013

Document: *Uit Canton.*

Document id: _gid001189601_01_0080 **Probability:** 0.9308549

Topic 41: 'n maurice voelen louise nou soonbeek vrouw kijken ooog horen lief zoo'n stem diep hevig nee mensch gezicht kind ziel jij florence liefde as jou heelemaal angst innig plots begrijpen

Proportion: 0.027722854
Document: *Kunstenaarsleven.*
Document id: _gid001190601_01_0023 **Probability:** 0.99979866

Topic 14: engeland maatschappij arbeid volk eeuw mensch sociaal gladstone politiek arbeider rijk macht staat kracht wereld geld rusland recht kapitaal zaak europa werken arm engelsch japan boek beginsel toestand invloed belang
Proportion: 0.024309214
Document: *Charles Hall's kreet.*
Document id: _gid001190201_01_0102 **Probability:** 0.9409885

Topic 0: frankrijk engeland koning oorlog keizer lord prins zaak macht republiek europa vrede politiek minister regeering rusland vriend vorst fransch wereld parijs brief plan hart leger duitschland sluiten hoofd staat altoos
Proportion: 0.023929307
Document: *De laatste regeeringsjaren van Willem III. (1698-1702)*
Document id: _gid001189801_01_0110 **Probability:** 0.99942535

Topic 46: ich goethe das hamlet nicht ist so ein sie sich mit shakespeare dem dichter auf es mir mich regel schrijven an geest nur im dass drama brief wereld zu eine
Proportion: 0.02090833
Document: *Nederlandsche Shakespeare-kritiek.*
Document id: _gid001190001_01_0084 **Probability:** 0.98640364

Topic 36: straf misdadiger rechter persoon gevangenis misdaad krankzinnig misdrijf vraag congres invloed feit veroordeeling zedelijk voorwaardelijk oorzaak veroordelen maatschappij mensch karakter onderzoek omstandigheid individu bijzonder verschillend wet gevolg vroeg strijd verklaren
Proportion: 0.019971298
Document: *Nieuwe strafrechtspolitiek.*
Document id: _gid001190101_01_0050 **Probability:** 0.999754

Topic 29: verhaal taal studie grieksch letterkunde geschiedenis eeuw heer naam wetenschappelijk vorm dier sprookje vertelling volk modern fransch museum bekend verschillend jong verzameling wetenschap vrouw schrijver lezing historisch prof paris behandelen
Proportion: 0.01798134
Document: *Komen onze sprookjes en vertellingen uit Indië?*
Document id: _gid001190201_01_0092 **Probability:** 0.9997116

Topic 28: indie heer minister indische regeering millioen zaak wet hall koning nederland begrooting goud zilver financien belasting kamer middel prijs geld verklaren vraag gelden gevolg lid toestand voorstel voorstellen commissie zm
Proportion: 0.017622069
Document: *Het muntvraagstuk in Britsch-Indië.*
Document id: _gid001189401_01_0022 **Probability:** 0.9984866

Topic 33: stuk vrouw les drama liefde balzac vous kind jong kunst bedrijf dramatisch publiek mensch schrijven rol plus hartstocht wereld pour gevoel roman schrijver dame persoon talent dans tooneelspeler fransch si
Proportion: 0.017235296
Document: *Honoré de Balzac.*
Document id: _gid001189601_01_0076 **Probability:** 0.9761971

Topic 16: cats huygens ick holland amsterdam kunst zaak eeuw volk daer schrijven fransch haer stuk soo nederlandsche vs republiek aen compagnie werken dichter vgl lid vergadering vondel meester uyt invloed vroeg
Proportion: 0.015981307
Document: *Korte inleiding tot de Bataafsche geschiedenis, 1795-1798.*
Document id: _gid001190601_01_0074 **Probability:** 0.7609702

Topic 4: vrouw prins cornelie voelen vragen urania rome jong keizer ooog kind prinses kamer arm misschien lachen dame schrijven begrijpen avond vreemd lijn lief liefde stad glimlachen san vader idee huwelijk
Proportion: 0.015867326
Document: *Langs lijnen van geleidelijkheid.*
Document id: _gid001190001_01_0049 **Probability:** 0.9248241

Topic 9: gild raad stad utrecht spinoza lid zaak koning katholiek prins verbieden buiten burger persoon vreemd recht bedrijf brief betalen kerk reden utrechtsche verklaren geestelijk naam eeuw behoren maart bepaling besluit
Proportion: 0.011906041
Document: *De gilden en het regeeringstoezicht op handel en nijverheid in de middeleeuwen.*
Document id: _gid001189701_01_0060 **Probability:** 0.9997577

Topic 13: recht dl kind wet onderzoek staat vaderschap internationaal vader verbod arbitrage regeering islam mohammedaansch artikel belang heer zaak bevolking feit vv koningin lijst huwelijk conferentie arabisch moeder gezag europeesche geschil
Proportion: 0.010717626
Document: *Onderzoek naar het vaderschap.*
Document id: _gid001189601_01_0004 **Probability:** 0.99649173

Topic 45: plant ontstaan dier verschillend eigenschap bloem natuur vorm verschijnsel talrijk cel bekend eicel verschil vormen kern varieteit naam gebied mensch belangrijk bevruchting beteekenis bastaard blad zaad uiterst onderzoek geschieden de_vries
Proportion: 0.009283437
Document: *Hugo de Vries' Mutatie-theorie.*
Document id: _gid001190101_01_0028 **Probability:** 0.9564597

Topic 7: stuk drama rol dramatisch vertooning m bedrijf tooneelspeler oedipus blijspel vertaling stelling recht spelen rivier publiek tragedie koning amsterdam eeuw toestand heer gedeelte voet toeschouwer diep slot nederlandsch_tooneel waal schrijven
Proportion: 0.008956353
Document: *Dramatisch overzicht.*
Document id: _gid001189801_01_0006 **Probability:** 0.99867517

Topic 42: jan huis jongen boer heur lijk buiten vragen ding voelen kijken mensch vader zoeken ooog eten dorp avond dolf geld rond moeder wachten dood knaap wijf peerden werken jong hoofd
Proportion: 0.008548555
Document: *Langs de wegen.*
Document id: _gid001190101_01_0026 **Probability:** 0.9998107

Topic 44: schip officier kapitein boord vijand fort zee resident soldaat militair wal eiland luitenant vlag haven vaartuig matroos troep kommandant baai kust nacht geweer uur vloot strand gereed reis bevinden bevel
Proportion: 0.007590747
Document: *Bladen uit het memoriaal van den vice-admiraal J. Boelen.*
Document id: _gid001190301_01_0110 **Probability:** 0.9997596

Topic 22: expeditie kaart belangrijk deg dr diepte onderzoek m ijs schilderij zee gebied eiland tocht kust schip verschillend resultaat temperatuur verkrijgen water geologisch bekend waarneming nansen richting wetenschappelijk gedeelte zuiden breedte
Proportion: 0.007505021
Document: *De Zuidpool-campagne van 1901-1904.*
Document id: _gid001190501_01_0072 **Probability:** 0.9995526

Topic 27: berenice vorst lombok otto brief compagnie bali schrijven raffles gouverneur java engelschen koning zaak grootmoeder zenden hoofd batavia sultan vragen martinus jong engelsch you contract baliers volk palembang eiland sir

Proportion: 0.0072958455

Document: *De geschiedenis van een Engelschen raid op Hollandsch grondgebied.*

Document id: _gid001189801_01_0013 **Probability:** 0.98807454

Topic 20: woning gemeente stad bouwen PS huis terrein straat amsterdam gemeentebestuur grond perceelen verbetering persoon volkshuisvesting wet commissie onteigening eigenaar plan kamer gebouw plein particulier wijk woningwet mensch bevolking londen m

Proportion: 0.0060706753

Document: *Het woningvraagstuk in eenige Britsche steden.*

Document id: _gid001190201_01_0110 **Probability:** 0.8523193

Topic 10: ziel schrijven boek aylva kunst emilie voelen mathilde mensch jong lief lezen liefde hugo vreemd kamer geluk vragen mevrouw leed misschien dood begrijpen werken ziek waarheid vrouw mama prinses roman

Proportion: 0.00584473

Document: *Van de prinses met de blauwe haren.*

Document id: _gid001190101_01_0093 **Probability:** 0.763262

Topic 8: rembrandt wet thompson arbeid huis vrouw boek arbeider steinlen recht zola rijkdom hodgskin heer mensch jacob driebergen eigendom persoon stad vestigen kapitalist prent schrijven naam doel middel werken buiten product

Proportion: 0.0055746054

Document: *De pest en hare bestrijding in vroeger eeuwen.*

Document id: _gid001190001_01_0006 **Probability:** 0.86976314

Topic 40: vers gedicht poezie dichter hooft sonnet dr j dichteres riebeek lied g bundel h nederlandsche haer sijn vorm soo w prof vroeg jong mr amsterdam inleiding lyriek regel dichterlijk schrijven

Proportion: 0.005532644

Document: *Bibliographie.*

Document id: _gid001190301_01_0034 **Probability:** 0.89943457

Topic 43: bolland dansen beweging duncan symbool beatrice gebaar meisje dans lichaam miss electriciteit signalement dante lijn buis grieksch kunst verschillend afstand stof lengte beschrijven bekend lucht divina_commedia straal maat verklaring verkrijgen

Proportion: 0.0045710313

Document: *Isadora Duncan, haar kunst en haar ideaal.*

Document id: _gid001190601_01_0043 **Probability:** 0.99674386

Topic 6: meneer nie da vreemdeling hanna 'n tonia heeren ooog tante moar moeder janus secretaris boer glas vrede dorp kijken veur ha vragen vrouw roepen as e huis deur gezicht joa

Proportion: 0.004271662

Document: *De steunpilaren der Ope van vrede.*

Document id: _gid001190301_01_0048 **Probability:** 0.9610657

Topic 38: kind taal school leerling onderwijzer spelling onderwijs dr schrijven winkel nederlandsch hollandsch schrijftaal jong leren boek de_vries onderwijzeres moedertaal e beschaafd geslacht lezen kennis opvoeder uitspraak kennen heer cosijn vrouwelijk

Proportion: 0.004241765

Document: *Paedagogiek der ervaring.*

Document id: _gid001190601_01_0008 **Probability:** 0.9980621

Topic 34: ziekte pasteur dier zout stof bloed lucht onderzoek bacterie hoeveelheid mensch organisme verschillend oorzaak wetenschap invloed zuurstof voorkomen bacil ooog zedelijk laboratorium dr hooggebergte verschijnsel ontstaan lichaam leren bijziend verkrijgen

Proportion: 0.0037521797

Document: *Het middel van Behring-Roux tot voorkoming en genezing van diphtherie.*

Document id: _gid001189401_01_0113 **Probability:** 0.99925345

Topic 23: bedrijf grond ha boekerij arbeider openbaar weber wet landbouw rente stad werken boek engeland duitschland sachs gebruik betalen gemeente david amerika bibliotheek leeszaal spontini particulier volk duitsche cultuur boer middel
Proportion: 0.0034205783
Document: *Openbare boekerijen en leeszalen.*
Document id: _gid001190101_01_0060 **Probability:** 0.9995137

Topic 17: ruskin fabel turner chr grieksch tijger eeuw indische vorm indie achten dier herodotus koning boek benfey verhaal dr hellas lezen aristophanes leeuw erkennen jakhals voornaam oorsprong kennen mensch schilderkunst schilder
Proportion: 0.002663002
Document: *Over Grieksche en Indische fabels.*
Document id: _gid001190301_01_0101 **Probability:** 0.97349936

Topic 3: taal latijn romaansch spreektaal fransch natuurkeus darwin indogermaansch latijnsch uitspraak verschillend beteekenis weismann bekend strijd dier vraag lord vorm schrijven volkstaal uitspreken gebied salisbury engelsch woordenschat sterk huxley klank frankrijk
Proportion: 0.0024002432
Document: *Afstammingsleer en Darwinisme door Lord Salisbury geoordeeld.*
Document id: _gid001189401_01_0092 **Probability:** 0.99938804

Topic 1: vrucht pruim kruising bloem varieteit cultuur californie zitting tafel e kruisen vragen verkrijgen advokaat wild duidelik verschillend kweeker geest verschijnsel smaak doel eigenschap bastaard kleur professor zaad naam kans vraag
Proportion: 0.0021029334
Document: *Een zitting met Eusapia Paladino.*
Document id: _gid001190301_01_0072 **Probability:** 0.9743912

Topic 37: vertaling nederland auteur nederlandsche vertalen conventie vertaler werken recht taal belgie vreemd uitsluitend oorspronkelijk bescherming verschijnen auteursrecht uitgeven schrijver internationaal nadruk belang fransch uitgave uitgever nederlandsch vreemdeling wetgeving beginsel tractaat
Proportion: 0.0021013108
Document: *Nederland en de Berner Conventie.*
Document id: _gid001189601_01_0091 **Probability:** 0.9993911

Topic 12: mens taal tussen zoals fries enig ogenblik zo'n dadelik natuurlik mogelik horen zoveel nederlands vroeg betekenis eeuw hollands nodig sterk begrijpen helemaal klank frans presie bekend veranderen heten altans wijzen
Proportion: 0.0019061684
Document: *Afscheidsgroet.*
Document id: _gid001190301_01_0077 **Probability:** 0.9378856

Topic 32: da costa vriend de_clercq wereld literatuur willem_de_clercq hart da_costa gevoel querido boek vriendschap geest lezen mensch dichter volk poezie god ding huis bilderdijk sirocco levensgang kracht vragen kunstenaar horen van_eeden
Proportion: 0.001805681
Document: *Letterkundige kroniek.*
Document id: _gid001190201_01_0018 **Probability:** 0.99744296

Topic 24: school hegel absoluut onderwijs tooneelschool katholiek groen geest prinsterer tooneelspeler kerk wereld guido doel monna_vanna mensch zig natuur geschiedenis kracht protestantsche predikant wereldbeschouwing dog falck schrijven kennen buiten schoolwet karakter
Proportion: 0.0017488356
Document: *Dramatisch overzicht.*
Document id: _gid001190501_01_0093 **Probability:** 0.9984325

Topic 18: pest rat ziekte ziek dr mensch marseille toerist besmetten dier passagier huis lichaam epidemie zaak student th bacil ongedierte ontstaan reiziger quarantaine wetenschap gevaar maatregel

besmetting vroeg vuil kracht vorm
Proportion: 0.0015280454
Document: *De pest.*
Document id: _gid001190001_01_0005 **Probability:** 0.9952349

Topic 30: staat imaginair partij malade oosten candidaat kind stem democraat congres geneeskunst kiezen heele kol presidentschap amerika belang mevrouw westen ohio nellie zilver schrijven dokter mensch programma wereld stad kracht naam
Proportion: 0.0002780392
Document: *Buitenlandsch overzicht.*
Document id: _gid001189601_01_0079 **Probability:** 0.99647397

Topic 47: bulgarije pastoor bulgaarsche heijermans russisch nansen vrouw bedrijf moeder rusland grijpen spel eisch zinnebeeldig pastorie alexander jammer dramatisch mensch jong ding plicht eer strijd vorst kind deelnemen debat situatie biechten
Proportion: 0.00024092293
Document: *Buitenlandsch overzicht.*
Document id: _gid001189501_01_0086 **Probability:** 0.8032521

Topic 15: dreyfus leger militair kracht dreyfuszaak zola proces judas verrader frankrijk intellect volk jood verite zaak staf recht samenleving strijd groep generaal kerk verklaren meening bijzonder minister voorbeeld conservatief kring rennes
Proportion: 0.00020152022
Document: *Buitenlandsch overzicht.*
Document id: _gid001189901_01_0092 **Probability:** 0.9971947

Appendix E

Experiment 3: 300 topics from 1837-1936

The topic word clusters below are a result of a qualitative inspection of the topic word distributions that were generated by a model of 75, 125, 200 and 300 topics. Below results show the most coherent topics: i.e. topics that can easily be interpreted, that show little to no overlap with other topics, and are not redundant in the total overview. The formatting is the same as in the previous two experiments (Appendix C and Appendix D). The topics are sorted on their proportion in the corpus.

E.1 300 topics

Topic 101: kijken ooog jong hoofd voelen gezicht huis vragen kind lopen vrouw licht wit horen donker zwart arm lachen trekken avond mensch deur open meisje stem roepen lucht slaan tafel morgen
Proportion: 0.0351279
Document: *De wandeling.*
Document id: _gid001190401_01_0002 **Probability:** 0.99324316

Topic 62: stad huis uur voet vrouw paard water berg dier mensch straat kind wit avond jong dragen trekken hoofd arm morgen naam woning dorp groen nacht licht ooog grond zijde natuur
Proportion: 0.028917722
Document: *Herinneringen uit Algiers.*
Document id: _gid001188601_01_0004 **Probability:** 0.92905074

Topic 80: schrijver boek heer schrijven geschiedenis lezer lezen naam auteur roman verhaal vorm geest karakter licht kunst eeuw hart bekend publiek persoon werken taal volk onderwerp voorstelling oordeel kennis leveren zaak
Proportion: 0.027812185
Document: *In de kajuit. Transatlantische Schetsen.*
Document id: _gid001184701_01_0026 **Probability:** 0.7775706

Topic 198: hart licht liefde ziel dood aarde nacht god wereld diep schoon ooog zingen hemel stem mensch horen dichter kracht sterven geest zon voelen lief beeld roepen rust gedicht smart zee
Proportion: 0.027375475
Document: *Ave terra: praemortui te salutant*
Document id: _gid001191501_01_0020 **Probability:** 0.9965586

Topic 107: mensch wetenschap wereld natuur geest begrip ding waarheid menschelijk kennis zedelijk kracht recht vraag wijsbegeerte vorm verschijnsel god zin gevoel zoeken leren werkelijkheid werken ontwikkeling dier doel erkennen buiten begrijpen

Proportion: 0.02627746

Document: *Plato's natuurbeschouwing. Naar aanleiding van de Geschiedenis der Wijsbegeerte van wijlen Prof. C.B. Spruyt .*

Document id: _gid001190501_01_0018 **Probability:** 0.9382374

Topic 144: volk vrijheid geschiedenis staat beginsel natie staatkundig belang invloed eeuw zaak frankrijk gezag partij wet politiek omwenteling toestand koning strijd geest europa ontwikkeling vorst regering vroeg kracht revolutie denkbeeld maatschappij

Proportion: 0.021423321

Document: *Alexis de Tocqueville.*

Document id: _gid001186701_01_0077 **Probability:** 0.85920864

Topic 64: boek schrijver schrijven jong roman lezen mensch heer auteur stuk verhaal kunst vrouw talent lezer liefde indruk letterkundig gevoel ding vertellen werken publiek persoon voelen wereld hart geest zoeken misschien

Proportion: 0.021087887

Document: *Letterkundige kroniek.*

Document id: _gid001189801_01_0113 **Probability:** 0.90584874

Topic 180: god godsdienst christendom jezus geloof gods christelijk christus mensch godsdienstig kerk waarheid geest wetenschap evangelie geschiedenis boek theologie christen historisch dr openbaring geloven leren schrijven wereld vraag eeuw verklaren schrift

Proportion: 0.020813167

Document: *Levensteekenen der moderne theologie.*

Document id: _gid001186601_01_0083 **Probability:** 0.85096073

Topic 121: dichter geest hart vers poezij kunst schrijven wereld eer schoon heer lezen dier stuk kind gelukkig vriend lief misschien eeuw naam vragen talent hoofd lezer vroeg volk gedachte onderwerp jong

Proportion: 0.020564595

Document: *Quos ego! Hekelrijmen door den Autheur der Hippokreen-ontzwaveling. Groningen, bij P. van Zweeden. 1844.*

Document id: _gid001184401_01_0046 **Probability:** 0.91265655

Topic 158: schrijver kennis belangrijk heer bekend geschiedenis gedeelte boek verschillend wetenschap lezen beschrijving stuk voorkomen naam vermelden bevatten kaart lezer hoofdstuk vroeg behandelen leveren onderzoek dier uitgave onderwerp bijzonder wetenschappelijk schrijven

Proportion: 0.020102005

Document: *Tijdschrift voor algemeene munt- en penningkunde, uitgegeven door P.O. van der Chijs . Tweeden Deels, tweede Stuk (van bladz. 373 tot 668). Te Leiden, bij S. en J. Luchtmans, Academiedrukkers. 1841.*

Document id: _gid001184201_01_0050 **Probability:** 0.8629822

Topic 21: minister partij politiek heer regeering liberaal kamer meerderheid ministerie beginsel zaak kabinet strijd conservatief belang vraag stem parlementair kracht verklaren wet verkiezing oppositie lid zijde voorstel misschien recht toekomst parlement

Proportion: 0.020053381

Document: *Parlementaire kroniek.*

Document id: _gid001190801_01_0010 **Probability:** 0.97230524

Topic 143: leger vijand fransch prins troep stad oorlog koning frankrijk vesting zaak zijde hoofd strijd volk aanval trekken bevel generaal bondgenooten oranje soldaat voeren magt holland hollandsche sterk partij vorst dier

Proportion: 0.017602028

Document: *1705.*
Document id: _gid001186501_01_0057 **Probability:** 0.9949969

Topic 191: kunst geest wereld vorm schoonheid eeuw mensch historisch geestelijk boek kunstenaar modern zin beeld figuur diep sterk zuiver dichter werkelijkheid gevoel persoonlijkheid idee beteekenis voelen scheppen karakter begrijpen ideaal schrijver
Proportion: 0.017561637
Document: *Kant en Schiller.*
Document id: _gid001191001_01_0080 **Probability:** 0.7881346

Topic 6: roepen vader vriend vragen vrouw arm heer antwoorden hertog verlaten hart jong blik hoofd ooog eer zoon huis meester kind gelaat stem dochter hernemen wachten werpen ontvangen liefde schoon naam
Proportion: 0.01617182
Document: *Eene kroon voor Karel den Stouten.*
Document id: _gid001184101_01_0096 **Probability:** 0.9945932

Topic 169: vragen jong vrouw kind kijken mevrouw ooog vader huis moeder begrijpen lachen voelen misschien horen jij vertellen brief kamer meisje dokter avond lief mensch uur zoo'n praten hoofd gezicht schrijven
Proportion: 0.015074076
Document: *'T geluk hangt als een druiventros... Een verhaal uit het Florentijnsche.*
Document id: _gid001191801_01_0120 **Probability:** 0.9905418

Topic 255: les dans plus pour se au on son tout leur fransch ses sa mais frankrijk elle comme avec sur sont cette nous si ou ils ces lui aux parijs bien
Proportion: 0.014575336
Document: *Buitenlandsche letterkunde.*
Document id: _gid001191801_01_0107 **Probability:** 0.99646205

Topic 229: arbeid maatschappij kapitaal nijverheid middel arbeider zaak arm handel werken toestand kracht behoefte welvaart werkman rijk kennis belang volk wet vraag stelsel mensch rijkdom vroeg armoede stand verkrijgen prijs ontwikkeling
Proportion: 0.014539112
Document: *De werk-inrigtingen voor armen, uit een staatshuishoudkundig oogpunt beschouwd, door Mr. W.C. Mees . Te Rotterdam, bij J. van Baalen en Zonen. 1844.*
Document id: _gid001184401_01_0070 **Probability:** 0.8183025

Topic 45: regering zaak regt schrijver wet belang reggen heer stelsel regten vroeg beginsel magt gevolg bepaling staat geschiedenis stuk art volk schrijven naam kennis punt bekend gebraggen grond onderwerp vraag middel
Proportion: 0.014232552
Document: *Het Reglement op de drukwerken in Nederlandsch Indië .*
Document id: _gid001185701_01_0013 **Probability:** 0.76516694

Topic 216: muziek opera componist beethoven werken kunst muzikaal meester schrijven publiek uitvoering dramatisch compositie drama mozart wagner orkest duitsche opvoering symphonie concert gebied theater stuk zanger tekst duitschland voorstelling opvoeren gelegenheid
Proportion: 0.0125838285
Document: *Muzikaal overzicht.*
Document id: _gid001189801_01_0126 **Probability:** 0.97491926

Topic 30: grond bevolking landbouw toestand recht belasting betalen cijfer handel arbeid prijs waarde gevolg staat opbrengst wet verkrijgen stelsel bedragen eigenaar eigendom landbouwer belang vroeg verschillend bodem regeling millioen last kapitaal
Proportion: 0.011887876
Document: *De laatste regeling der landrente op Java.*
Document id: _gid001187301_01_0039 **Probability:** 0.9525175

Topic 33: vrouw koning zoon vader klooster kind god dood vorst naam mensch wereld broeder rijk geest sterven schrijven zuster huis jong eeuw stad moeder volk vriend hoofd heilig portroyal brief vroeg
Proportion: 0.011850945
Document: *Port-Royal.*
Document id: _gid001187301_01_0004 **Probability:** 0.9299696

Topic 189: sociaal politiek economisch recht maatschappelijk arbeider groep ontwikkeling belang maatschappij gebied wet staat arbeid organisatie beginsel richting gemeenschap partij bijzonder samenleving school volk strijd invloed bedrijf onderwijs verschillend vraag beteekenis
Proportion: 0.01152154
Document: *De R.K. Centrale Raad van Bedrijven.*
Document id: _gid001192001_01_0048 **Probability:** 0.9116089

Topic 145: moeder vader kind vrouw vragen huis oom arm jong lief horen meisje ooog hart thijs liefde roepen dochter hoofd begrijpen zoon avond mensch lex blik geld henk misschien kamer voelen
Proportion: 0.011439625
Document: *Moeder*
Document id: _gid001193001_01_0028 **Probability:** 0.76380205

Topic 165: schrijver mensch reggen wetenschap waarheid beginsel rigting begrip voorstelling gevoel grond geest beschouwen doel god regt zaak magt kracht onderzoek denkbeeld menschelijk handelen redenering onderwerp ontwikkeling stelling bewijzen toestand betrekking
Proportion: 0.010899317
Document: *Nieuwe Verhandelingen van het Genootschap tot Verdediging van de Christelijke Godsdienst, tegen derzelver hedendaagsche bestrijders . Voor het jaar 1835.*
Document id: _gid001183801_01_0028 **Probability:** 0.88330925

Topic 151: indie java indische inlandsche bevolking inlander koloniaal kolonie bestuur nederland ambtenaar belang europeesche regeering nederlandsche moederland toestand europeanen zaak batavia volk ontwikkeling hoofd heer gezag gouverneurgeneraal javaan indisch nederlandschindie islam
Proportion: 0.010454567
Document: *Insulinde's toekomst.*
Document id: _gid001190801_01_0078 **Probability:** 0.7276619

Topic 271: das ist nicht ein sich goethe dem sie mit ich so auf eine dass zu im es auch nur sein wir aus hat wenn aber von an schrijven durch oder
Proportion: 0.009811583
Document: *Het sprookje in de wetenschap*
Document id: _gid001192801_01_0114 **Probability:** 0.7375124

Topic 104: oorlog duitschland frankrijk belgie engeland nederland regeering internationaal mogendheid duitsche staat vrede verdrag fransch belang politiek militair conferentie zaak europa recht leger italie volkenbond belgisch nederlandsche volk gebied nationaal tractaat
Proportion: 0.009713366
Document: *Buitenlandsch overzicht*
Document id: _gid001193601_01_0095 **Probability:** 0.8897498

Topic 272: onderwijs leerling school vak kennis studie wetenschap leeraar taal ontwikkeling uur gymnasium methode wetenschappelijk middelbaar wiskunde docent leren opleiding student grieksch doel cursus latijn klasse verschillend les klas onderwijzen vormen
Proportion: 0.009698365
Document: *Het industriëel onderwijs.*
Document id: _gid001186101_01_0037 **Probability:** 0.91143566

Topic 122: wet straf recht rechter misdrijf vraag art zaak wetgever wetgeving gevangenis misdadiger persoon belang onderzoek maatschappij wetboek beginsel feit toepassing staat gelden bepaling

regel vrijheid veroordelen middel plegen gevolg stelsel
Proportion: 0.009500001
Document: *Een Nederlandsch strafstelsel.*
Document id: _gid001187901_01_0018 **Probability:** 0.9743994

Topic 32: recht grondwet wet lid belang minister art regeling macht staat stelsel meerderheid regeering kiesrecht volk koning kamer bepaling vertegenwoordiging artikel beginsel ontwerp stem parlement kiezer voorstel vergadering wijziging verkiezing vraag
Proportion: 0.009422573
Document: *Kiesrecht.*
Document id: _gid001190401_01_0041 **Probability:** 0.8762516

Topic 118: kerk kerkelijk katholiek gemeente paus staat predikant zaak vrijheid geestelijk bisschop kerkgenootschap beginsel koning lid roomsch godsdienst protestantsche gezag rome regt wet hervorming partij godsdienstig strijd geestelijkheid eeuw hervormen protestant
Proportion: 0.00891161
Document: *Kerk en Staat.*
Document id: _gid001186301_01_0025 **Probability:** 0.8717161

Topic 141: leger officier oorlog militair vijand verdediging soldaat dienst gedeelte fort schutterij marine vesting sterkte troep linie verdedigen middel zaak stelling nederlandsche nederland kracht verkrijgen sterk belang wapen volk toestand voeren
Proportion: 0.008513966
Document: *Een woord over onze schutterij.*
Document id: _gid001186301_01_0044 **Probability:** 0.86654127

Topic 173: kind ooog horen mensch liefde vragen vrouw stem voelen jong huis licht moeder arm lief hart zoeken vader roepen avond begrijpen geluk hoofd wachten ziel hilda hemel wereld lachen vreugde
Proportion: 0.007875183
Document: *De mensch van Nazareth.*
Document id: _gid001191601_01_0127 **Probability:** 0.8772323

Topic 66: that it with as for not be his he but all this which have from engelsch are they or you on so no her one will there who at what
Proportion: 0.007791224
Document: *Hedendaagsche Engelsche dichters.*
Document id: _gid001190801_01_0107 **Probability:** 0.98157597

Topic 190: voorstelling gewaarwording begrip kant waarheid ding stelling ervaring verschijnsel kennis oorzaak geest voorwerp waarneming bewustzijn ruimte mill wetenschap spinoza buiten werkelijkheid redeneering beschouwing denkbeeld verschillend feit wereld wet vraag werkelijk
Proportion: 0.0077765365
Document: *Philosophie?*
Document id: _gid001191001_01_0111 **Probability:** 0.9994008

Topic 211: zingen blank ziel hart bloem kind lief liefde licht ooog arm wind blij blauw zon nacht voelen zoet dood god ’k water horen droomen roos droef wit groen vogel traan
Proportion: 0.0077056927
Document: *Verzen.*
Document id: _gid001190101_01_0041 **Probability:** 0.99669755

Topic 42: dienst regeering personeel taak vloot belang commissie departement toestand statistiek bijzonder belangrijk omstandigheid verkrijgen buiten uitgaaf gevolg zeemacht kost eischen werken zaak militair maatregel achten minister ambtenaar behoefte vraag marine
Proportion: 0.0074736085
Document: *Bezuiniging bij de zeemacht tevens verbetering.*
Document id: _gid001190701_01_0062 **Probability:** 0.94600964

Topic 177: koning frankrijk engeland republiek prins vrede spanje gezant zaak oranje brief fransch schrijven onderhandeling regeering lodewijk_xiv hertog oorlog engelsch willem_iii staat politiek sluiten recht keizer hof belang macht parijs staten
Proportion: 0.0070057143
Document: *De Republiek der Vereenigde Nederlanden in hare staatkundige betrekkingen gedurende de eerste jaren na den Vrede van Utrecht. (1713-21)*
Document id: _gid001189901_01_0065 **Probability:** 0.8839056

Topic 111: zaak minister boer president regeering lid republiek engeland engelsch recht politiek partij volk transvaal wet zuidafrika heer vergadering huis verklaren parijs maand commissie engelschen oorlog bekend naam benoemen generaal constitutie
Proportion: 0.0070000645
Document: *Zuid-Afrika zoo als het thans bestaat.*
Document id: _gid001189701_01_0092 **Probability:** 0.9326452

Topic 226: vrouw vader dichter hart hooft jong cats schrijven huis vriend gedicht lief kind liefde vondel meisje dochter naam vers zoon mensch brief moeder lezen eer eeuw geest vragen vroeg dood
Proportion: 0.006959149
Document: *Tollens' Vrijage .*
Document id: _gid001190001_01_0107 **Probability:** 0.7799341

Topic 263: mevrouw vragen mijnheer moeder maria hart vrouw heer vader dame tante jong huis begrijpen brief antwoorden zaak ooog regina eer hoofd oom vriend geloven nicht gelaat toon kind horen avond
Proportion: 0.00694133
Document: *Het altaarbeeld van Saventhem.*
Document id: _gid001188001_01_0074 **Probability:** 0.9997195

Topic 258: vous pour les si mon votre plus ma ai dans bien mais brief comme tout moi y avec mes julie elle schrijven cette au suis se nous lui meme encore
Proportion: 0.006552815
Document: *Julie Simon. De Levensroman van R.C. Bakhuizen van den Brink. Uit brieven en bescheiden tezamengesteld door C. en M. Scharten - Antink.*
Document id: _gid001191301_01_0086 **Probability:** 0.9935617

Topic 236: mensch ziel wereld liefde geest menschelijk bewustzijn innerlijk zin werkelijkheid god geestelijk kracht ding diep natuur bewust buiten menschheid idee voelen sterk gevoel ervaring verhouding zuiver geluk vorm scheppen werken
Proportion: 0.0063706324
Document: *Marginalia.*
Document id: _gid001191601_01_0145 **Probability:** 0.98756355

Topic 48: voelen meta vragen mevrouw ooog vrouw jong schrijven ziel kind begrijpen liefde jenny gevoel arm mensch hoofd blik misschien stem brief huis geluk kamer lief zwart isa vroeg licht lachen
Proportion: 0.006112889
Document: *Mea Culpa.*
Document id: _gid001189501_01_0034 **Probability:** 0.9307741

Topic 26: vers gedicht dichter poezie bundel regel sonnet schrijven beeld rijm fransch jong kunst klank taal vorm gevoel lezen klinken proza zoeken versregel toon misschien ziel victor_hugo verlaine liefde strofe les
Proportion: 0.006041698
Document: *Fransche symbolisten .*
Document id: _gid001190201_01_0054 **Probability:** 0.80713385

Topic 50: rusland russisch oostenrijk volk frankrijk engeland duitschland koning keizer regeering russen europa zaak pruisen politiek rijk vorst duitsche mogendheid turkije staat berlijn plan mirabeau vergadering vorstius heer moskou hoofd lord

Proportion: 0.0059732697
Document: *Politiek overzicht.*
Document id: _gid001188601_01_0092 **Probability:** 0.7663836

Topic 95: volk keizer engeland politiek frankrijk koning lord macht bismarck geschiedenis recht oorlog zaak gladstone europa italie minister partij vrede wereld staat hart fruin kracht mensch schrijven ministerie pruissen napoleon parlement
Proportion: 0.0058910577
Document: *Politiek overzicht.*
Document id: _gid001186201_01_0061 **Probability:** 0.87922984

Topic 88: millioen spoorweg economisch belang staat nederland verkeer kost kapitaal gevolg lijn rente geld bedrag middel tarief regeering particulier industrie vraag belangrijk duitschland onderneming bedrijf buitenland aanleg verkrijgen betalen werken oorlog
Proportion: 0.0058272528
Document: *De electrificatie van Nederland.*
Document id: _gid001191701_01_0058 **Probability:** 0.8092479

Topic 203: stuk drama bedrijf dramatisch rol spelen publiek tooneelspeler vertooning spel voorstelling schouwburg blijspel comedie toeschouwer shakespeare kunst vrouw acteur mevrouw theatre vertonen heer schrijven moliere scene optreden jong tragedie treurspel
Proportion: 0.0055473233
Document: *Dramatisch overzicht.*
Document id: _gid001191301_01_0114 **Probability:** 0.9897085

Topic 152: lord mylady sir robert koning majesteit jane engeland koningin munt naam vriend misschien parlement thands londen pym ontvangen hoofd vragen broeder slaan strafford brief hare_majesteit roepen werkelijk wagen geloven mylord
Proportion: 0.005186042
Document: *Mylady Carlisle.*
Document id: _gid001186301_01_0063 **Probability:** 0.85141325

Topic 142: holland staten stad provincie statengeneraal zaak vergadering staat vorst prins lid amsterdam zeeland oranje utrecht regeering besluit commissie maart brief van_hogendorp heeren raad oldenbarnevelt stuk dienst recht stadhouder resolutie waardgelders
Proportion: 0.0051414375
Document: *Het stuk der waardgelders in de provincie Holland, hoofdzakelijk gedurende het ministerie van Johan van Oldenbarnevelt, toegelicht.*
Document id: _gid001185901_01_0034 **Probability:** 0.9912277

Topic 59: SS lezen e l ad schrijven schrijver vs non vertalen kai vertaling d ut boek aanhalen heer hr zin regel latijn voorkomen bewijzen tekst vers lezing beteekenis uitgave aanteekening quod
Proportion: 0.005098991
Document: *Hippocratis liber de victus ratione in morbis acutis. Edidit F.Z. Ermerins (Med. Doct.). Accedunt eiusdem observationes criticae in Soranum Ephesium de arte obstetricia morbisque mulierum. Lugd. Batav., apud S. et J. Luchtmans 1841.*
Document id: _gid001184201_01_0044 **Probability:** 0.99917245

Topic 188: school onderwijs kind onderwijzer opvoeding leerling openbaar ouder godsdienstig jongen bijzonder wet christelijk onderwijzen mensch heer beginsel lezen meisje jeugd leren opleiding mengen schrijven vragen klas zedelijk verschillend hoofd volksschool
Proportion: 0.004978119
Document: *De algeheele kosteloosheid van het lager onderwijs in Frankrijk.*
Document id: _gid001189901_01_0067 **Probability:** 0.79319173

Topic 12: koning spanje congres keizer vorst frankrijk oostenrijk frederik statistiek spaansch pruisen denemarken regering dood zaak oorlog rijk volk hertog duitschland ferdinand sleeswijk zoon minister italie verklaren napoleon holstein hertogdom sectie

Proportion: 0.004866509
Document: *Sleeswijk-Holstein tegenover Denemarken.*
Document id: _gid001186401_01_0010 **Probability:** 0.96321344

Topic 134: stad eeuw huis archief gebouw woning gemeente recht boek utrecht leiden vereeniging stichten bibliotheek bouwen eigenaar verzameling inrichting stichting bestuur openbaar graaf kerk lid geld geschiedenis zaak regeering belangrijk boekerij
Proportion: 0.0048127924
Document: *Armoedig noorden.*
Document id: _gid001191301_01_0133 **Probability:** 0.74733603

Topic 289: kunst kunstenaar museum tentoonstelling gebouw bouwkunst werken monument eeuw voorwerp architect stijl meester arbeid bouwmeester beeldend schilder vroeg bouwwerk jong beteekenis kracht kunstwerk vorm zoeken academie belangstelling schilderkunst zaak smaak
Proportion: 0.0047762827
Document: *Langs den ouden weg in nieuwe landen. Beschouwingen over de maatschappelijke positie der beeldende kunstenaars en over de taak hunner vereenigingen.*
Document id: _gid001191401_01_0133 **Probability:** 0.89278454

Topic 248: rome italie stad kerk romeinsch keizer paleis italiaansch beeld eeuw nero verheffen stuk tempel beroemd napels naam gebouw vroeg paus venetie florence bekend muur overblijfsel milaan schoon rijk oudheid romeinen
Proportion: 0.004621904
Document: *Rome.*
Document id: _gid001187101_01_0017 **Probability:** 0.94243205

Topic 219: examen onderwijs universiteit taal vak wet studie recht faculteit opleiding kennis regeling hoogleeraar commissie minister inrichting school student wetenschappelijk hoogeschool eischen regeering nederlandsche middelbaar gymnasium belang wetenschap gelegenheid instelling bijzonder
Proportion: 0.0045688795
Document: *Universitaire studie. (Congres-herinneringen.)*
Document id: _gid001190001_01_0112 **Probability:** 0.7955503

Topic 36: schilder schilderij kunst schilderen kunstenaar portret rembrandt meester stuk kleur werken eeuw landschap schilderkunst school doek rubens licht natuur figuur beeld tentoonstelling museum ets rijk verzameling schoon penseel teekening antwerpen
Proportion: 0.004536851
Document: *Unger's laatste etsen.*
Document id: _gid001188401_01_0015 **Probability:** 0.9467366

Topic 120: studie wetenschap letterkunde geschiedenis taal historisch wetenschappelijk eeuw volk geleerde methode duitsche pierson letterkundig hoogleeraar geleerd philologie onderzoek gebied germaansch universiteit geest schrijver schrijven arbeid grieksch klassiek fransch paris kennis
Proportion: 0.0045087687
Document: *Wetenschappelijke beoefening der moderne letterkunde.*
Document id: _gid001190101_01_0016 **Probability:** 0.7057168

Topic 239: regeering handel engeland engelsch staat kolonie nederland zaak millioen cijfer fr belgie vroeg tractaat stad unie uitvoer europa belang vestigen sluiten heer rijk toestand gebied recht bedragen frankrijk nederlandsche zilver
Proportion: 0.0043317927
Document: *Onze handel met Perzië en de Levant.*
Document id: _gid001190601_01_0025 **Probability:** 0.82868564

Topic 274: japan china engeland japansch amerika volk europa oorlog chineesche europeesche regeering mogendheid amerikaansch vloot macht generaal japanners engelsch politiek rusland troep turksch turken wereld rijk chineezen vereenigde_staat turkije keizer kracht

Proportion: 0.0042716977
Document: *Vlootpolitiek en ontwapening*
Document id: _gid001193201_01_0118 **Probability:** 0.8037647

Topic 233: plant dier ontstaan verschillend eigenschap vorm ontwikkeling darwin geslacht onderzoek vormen cel ras groep verschil organisme invloed bekend individu bloem kenmerk verschijnsel natuur beteekenis feit voorbeeld onderzoeking gebied jong omstandigheid
Proportion: 0.004143042
Document: *Bastaardeering en bevruchting.*
Document id: _gid001190301_01_0058 **Probability:** 0.9996872

Topic 262: verschijnsel psychisch oorzaak werking verschillend waarneming wetenschap voorstelling feit wensch mensch vraag toestand persoon onderzoek gebied psychologie bewustzijn grond patient gevolg handeling bepalen ontstaan individu neiging methode verklaring theorie omstandigheid
Proportion: 0.0041283085
Document: *Een wereldbeschouwing.*
Document id: _gid001191601_01_0029 **Probability:** 0.9336485

Topic 16: taal naam volk bekend stam vroeg boek lezen schrijven reis reiziger verschillend europa vreemd kennis schrijver eeuw rijk oorspronkelijk vrouw dier misschien beschrijving gebruiken zaak hollandsche trekken dr gebruik beschrijven
Proportion: 0.004079317
Document: *Indogermaansche oudheden.*
Document id: _gid001188401_01_0053 **Probability:** 0.70175767

Topic 245: athene mensch volk vrouw wereld stad god natuur atheners atheensch naam attische dier rijk amerika geest karakter verhaal oorlog staat zaak gevoel schrijver strijd kunst werken aarde dragen vroeg kracht
Proportion: 0.004025662
Document: *Het Attische volk en de kunst van Phidias.*
Document id: _gid001188401_01_0079 **Probability:** 0.9409674

Topic 29: amsterdam handel vrouw stad vreemd engeland londen amerika vreemdeling waarlijk pct daarenboven koopman volk gelukkig rijk eer haast vroeg wereld jong haven engelsch onmiddellijk uitnemend huis gedurig zoeken daarvoor evenzeer
Proportion: 0.0039413786
Document: *Verre vrienden.*
Document id: _gid001188501_01_0026 **Probability:** 0.74854165

Topic 27: naam stam stad arabisch geschiedenis eeuw waarschijnlijk volk vroeg rijk woestijn gebergte overlevering arabieren bekend streek grens vermelden oosten heer berg voorkomen egypte geslacht vgl koning islam zoon mohammed vorst
Proportion: 0.0039290637
Document: *De Edomieten en Nabateërs.*
Document id: _gid001185001_01_0056 **Probability:** 0.8560024

Topic 238: roman boek schrijven schrijver lezen lezer vrouw liefde verhaal historisch mensch werken goethe persoon jong heer kind hart kennis zaak misschien geschiedenis schrijfster wereld naam geest vragen eeuw zin auteur
Proportion: 0.0038155075
Document: *Letterkundige Kroniek.*
Document id: _gid001188701_01_0009 **Probability:** 0.7325449

Topic 296: brief potgieter schrijven huet vriend gids multatuli lezen jong bosboom letterkundig vrouw redactie boek maand schrijver mevrouw letterkunde tijdschrift herinnering gedicht amsterdam correspondentie stuk busken_huet geloven bakhuizen werken heer briefwisseling
Proportion: 0.0038091906
Document: *Drie brieven van Potgieter aan J.F Willems*
Document id: _gid001193101_01_0023 **Probability:** 0.90682983

Topic 241: schrijver taal boek woordenboek handschrift tekst voorbeeld vertaling naam beteekenis bekend dr lezen volk schrijven eeuw arabisch gebruiken voorkomen heer gebruik stuk reg vorm zin verklaring uitgave uitgeven oorspronkelijk prof
Proportion: 0.0037716515
Document: *Grammatica Arabica , breviter in usum Scholarum Academicarum conscripta a T. Roorda .*
Document id: _gid001183701_01_0095 **Probability:** 0.95514774

Topic 186: daer soo haer aen sijn sal hy maer ick ofte soude oock compagnie sy uyt schrijven brief sonder dees amsterdam wy sul seer onse wesen recht koning souden naer hadde
Proportion: 0.0037447931
Document: *Jan van Riebeek, de stichter der Kaap-kolonie.*
Document id: _gid001189401_01_0067 **Probability:** 0.90177107

Topic 13: taal schrijven spelling vlaamsch klank fransch nederlandsch vorm letter dr uitspraak regel schrijftaal nederlandsche beschaafd gebruiken gebruik horen volk de_vries vlamingen vreemd spreektaal zin vlaanderen lezen moedertaal winkel dialect e
Proportion: 0.0036450373
Document: *Het spelling-vraagstuk. De vereenvoudigde een gevaar voor volk en stam.*
Document id: _gid001191101_01_0014 **Probability:** 0.9851373

Topic 2: wet commissie geneeskundig gemeente belang staat lid gemeentebestuur maatregel geneeskundigen bepaling art regeling openbaar regeering zaak bevoegdheid voorschrift provinciaal persoon verordening onderzoek bestuur toestand plaatselijk raad geneeskunde onderwerp ziekte toezicht
Proportion: 0.0036213466
Document: *De werkzaamheden der gezondheids-commissiën.*
Document id: _gid001190401_01_0031 **Probability:** 0.93578917

Topic 54: that les be lord engeland it which on as not ged for dans this any or have an brief zaak castlereagh with pour would schrijven will britsche has government maart
Proportion: 0.0035747436
Document: *Buitenlandsch overzicht.*
Document id: _gid001192101_01_0028 **Probability:** 0.9019446

Topic 204: 'n voelen louise nou vrouw kind kijken ooog gezicht horen maurice diep moeder sprotje heemsbergen stem hevig zoo'n lief jij plots week as vragen ko lachen nee hoofd jong jou
Proportion: 0.0034931381
Document: *Kunstenaarsleven.*
Document id: _gid001190601_01_0062 **Probability:** 0.82417834

Topic 58: mens tussen volk enig zoals frans eeuw taal russiese betekenis nodig nederlands grieks mogelik zoveel sterk naam boek dostojewskij vrouw verschillend jong natuurlik geest voelen horen geschiedenis invloed staat begrijpen
Proportion: 0.003428409
Document: *Het slavendom en zijn historie.*
Document id: _gid001192801_01_0027 **Probability:** 0.98491806

Topic 282: nou mens jij kind mina vrouw christiaan voelen diep vragen horen mevrouw begrijpen zo'n misschien cavarna jou kijken huis hein nee vroeg zwijgen angelina gelukkig dadelik ogenblik maggie 'n verlangen
Proportion: 0.0033623704
Document: *Liefdeleven.*
Document id: _gid001191501_01_0045 **Probability:** 0.99978757

Topic 281: mensch geest schrijver wetenschap beschaving wereld eeuw volk boek natuur menschelijk kracht vorm naam denkbeeld maatschappij zaak verschillend voorstelling schrijven gevoel toestand ontwikkeling jong dier karakter zin vroeg waarheid leren

Proportion: 0.0032910875
Document: *Ralph Waldo Emerson.*
Document id: _gid001186101_01_0027 **Probability:** 0.7920088

Topic 223: poezie gedicht boek dichter schrijven eeuw schrijver vers mensch taal lezen volk vorm nederlandsche heer kunst dr misschien invloed geschiedenis naam wereld zin dier studie licht geest letterkunde gevoel bundel
Proportion: 0.00322259
Document: *Bibliographie.*
Document id: _gid001193001_01_0010 **Probability:** 0.766119

Topic 139: ich mir das mich nicht ein sie ihr mein so dichter ist dich mit dir im liefde dem drama sich auf kunst schiller vrouw gedicht sein jong kind wereld hart
Proportion: 0.0031287714
Document: *Maandelijksche praatjes.*
Document id: _gid001187501_01_0065 **Probability:** 0.88489914

Topic 105: ziekte lucht invloed water stof ontstaan kind gezondheid temperatuur klimaat lijder lichaam ziek oorzaak voedsel vorm verschijnsel toestand hoeveelheid mensch werking dier verschillend kracht koud voorkomen organisme sterk dr bloed
Proportion: 0.0029830579
Document: *Aanteekeningen over de Scarlatina , door D r . J.A. Wendt , getoetst aan eigene ervaring, door F.v.d. Breggen, Cz. Med. Doct. en Hoogleeraar te Amsterdam .*
Document id: _gid001183801_01_0044 **Probability:** 0.82658935

Topic 266: belang kiezer stem spaarbank beginsel partij volk mill persoon invloed staat zaak maatschappij zedelijk middel geluk candidaat getal gevolg doel algemeene verkiezing handeling mensch kracht vraag vereeniging maatschappelijk kiezen verkrijgen
Proportion: 0.0028603345
Document: *Democratie en constitutionele monarchie.*
Document id: _gid001187101_01_0021 **Probability:** 0.8324815

Topic 136: paul germaine vrouw vader kind jong vragen herman moeder laetitia avond constant mornar mireille benedictus stad hoofd huis horen misschien vriend oom vertellen meisje lopen voelen zoeken arm verlaten brief
Proportion: 0.0027555572
Document: *In den lusthof Arkadië.*
Document id: _gid001192001_01_0092 **Probability:** 0.8465406

Topic 285: prins prinses huwelijk koningin vrouw dochter brief schrijven maurits zoon cornelie moeder kind hof vader jong zuster huwen broeder huis urania elisabeth echtgenoot trouwen nassau maria duco oranje vorstin koning
Proportion: 0.0026528996
Document: *De jeugd van Louise Henriette d'Orange.*
Document id: _gid001186601_01_0041 **Probability:** 0.81814474

Topic 74: paulus brief schrijven schrijver apostel echtheid petrus zaak vs gemeente gal baur gevoelen verklaren xv christus bewijzen vroeg lezen bewijs heer_da_costa lucas geschrift jeruzalem punt evangelie dr schrift bekend grond
Proportion: 0.0026319288
Document: *Een belangrijk kritisch punt, door m r . Is. Da Costa tot eene afdoende beslissing gebragt, andermaal aan de gronden der waarschijnlijkheid getoetst.*
Document id: _gid001184901_01_0011 **Probability:** 0.99924934

Topic 178: bonaparte thands barras talleyrand generaal vragen eduard roepen fouche andwoorden sieyes begrijpen directoire hernemen heer gelaat andwoord republiek ottilie gohier horen geloven naam hoofd ontvangen vriend vernemen vroeg zijde zaak
Proportion: 0.0026318328

Document: *Achttien Brumaire.*

Document id: _gid001185201_01_0039 **Probability:** 0.999754

Topic 298: kapitein voorwerp resident stad volk eeuw zaak theramenes steenen strateeg rijk beschaving vroeg storm roberts dier vorm officier zee wal verschillend bronzen gebouw vaas zijde dop adelborst muur beeld invloed

Proportion: 0.0025609322

Document: *Schliemann's Troje.*

Document id: _gid001188201_01_0008 **Probability:** 0.8645065

Topic 103: keizer othomar licht prins ooog jong gouden wit keizerin ziel roos volk voelen kind lucht stad goud mensch hoofd vrede lief rijk liparie vreemd horen bloem vrouw kleur trekken kroonprins

Proportion: 0.0024537938

Document: *De Wajang Orang in Jogjåkartå (24-27 juni 1899).*

Document id: _gid001189901_01_0117 **Probability:** 0.8870931

Topic 295: schip zee eiland kust boord reis haven kapitein vloot varen zeeman water zeilen matroos golf vaart baai bemanning zeil anker dek augustus oceaan stuurman admiraal straat ondernemen schepen ontdekken tocht

Proportion: 0.0024380982

Document: *De ontdekkingsgeschiedenis van de Straat van Magellaan .*

Document id: _gid001187901_01_0051 **Probability:** 0.66966313

Topic 11: koning vorst portugeezen indie portugal schip rijk kust vloot handel zenden stad portugeesche keizer sultan muzelmannen albuquerque eeuw gouverneur goa eiland vesting naam vijand begeven lissabon vroeg turken bevinden ontvangen

Proportion: 0.0024290285

Document: *De Portugeezen in het Oosten.*

Document id: _gid001187701_01_0069 **Probability:** 0.82057285

Topic 9: israel volk koning profeet joden god vs joodsche boek naam geschiedenis egypte tempel verhaal priester david israels heilig juda heer jood salomo godsdienst godheid mozes jeremia israelieten stam jehova israelietisch

Proportion: 0.002317126

Document: *Een stap vooruit.*

Document id: _gid001186401_01_0036 **Probability:** 0.999173

Topic 127: schilderij amsterdam kunst stuk schrijven gids rembrandt veth heer van_de_helst portret bekend licht doek afbeelding werken dr eeuw rijksmuseum naam reproductie dyserinck schilder vroeg boek tijdschrift rembrandts schuttersmaaltijd verschijnen gravure

Proportion: 0.002297637

Document: *Een geschonden meesterstuk?*

Document id: _gid001189501_01_0058 **Probability:** 0.9719235

Topic 292: pruissen oostenrijk duitschland duitsche keizer bond pruissisch rijk regering vorst staat oostenrijksch beijeren eenheid vrede plan magt wurtemberg gevolg oorlog dier kracht rijksdag verdrag bondsvergadering doel verbinden saksen verkrijgen stand

Proportion: 0.0022973504

Document: *Vijftig jaren der Duitsche bondsgeschiedenis.*

Document id: _gid001186801_01_0066 **Probability:** 0.861566

Topic 5: amsterdam stad burgemeester amsterdamsch regent huis heer burgerij vader schrijven zaak regeering burger prins buiten ter partij stadhouder vroedschap gouw stadhuis vriend amsterdammers kerk holland utrecht straat patriot brief naam

Proportion: 0.0022945812

Document: *De zegepraal der hervorming te Amsterdam. (Vervolg en slot van blz. 114.)*

Document id: _gid001187801_01_0028 **Probability:** 0.8524861

Topic 257: muziek balzac water allard brouwer mensch jong horen spelen kapitein vrouw roepen vader kind vragen stad rivier schip hart schrijven licht begrijpen huis ooog dipo_negoro vroeg wind stem klinken wereld
Proportion: 0.0022784478
Document: *De avonturen van den muzikant aan het water.*
Document id: _gid001192601_01_0063 **Probability:** 0.99962246

Topic 99: reine kind helene vragen meisje generaal oscar vrouw huis hart mijnheer jong eleonore vader hoofd ooog arm grootmoeder trekken zuster moeder avond horen mensch mevrouw dood liefde antwoorden gelaat lief
Proportion: 0.0022721423
Document: *Victor d'Avlyn.*
Document id: _gid001187301_01_0062 **Probability:** 0.97821695

Topic 92: koning vrouw mevrouw anne jong mijnheer kind zoon filips vader pompadour huis bruno louis oven moeder heer hoofd mensch dame vriend alba zei graaf vorst naam nicht vragen dochter hart
Proportion: 0.0022315122
Document: *De schatgraver.*
Document id: _gid001184001_01_0102 **Probability:** 0.8107836

Topic 72: stof ligchaam dier organisch kracht scheikundig verschillend natuur verschijnsel ontstaan vormen vorm aarde water werking scheikunde eigenschap verandering bestanddeelen zuurstof hoeveelheid vorming verbinding chemisch warmte levenskracht ontwikkeling dierlijk waarneming wetenschap
Proportion: 0.002229023
Document: *Het leven.*
Document id: _gid001185201_01_0036 **Probability:** 0.7843234

Topic 217: vrouw kind meisje vrouwelijk gezin moeder schrijfster ouder opvoeding taal jong ziel huwelijk ontwikkeling recht tolstoj geest schrijven proust maatschappelijk werken verbeelding school dochter liefde doofstom wijsheid mensch leren wereld
Proportion: 0.0022015627
Document: *Het recht van de phantasie, en de opleiding van den onderwijzer.*
Document id: _gid001190201_01_0072 **Probability:** 0.6756735

Topic 181: heleen mensch ding voelen huis wereld kind ashley zoeken wilde gevoel geest lamb uur hart avond schrijven ooog sterk buiten vrouw vreemd vriend vragen boek plotseling brief vast werken lijken
Proportion: 0.0021435756
Document: *Heleen. Fragmenten van een roman. Tweede Gedeelte.*
Document id: _gid001191201_01_0125 **Probability:** 0.9995973

Topic 15: verhaal sprookje vertelling vertellen vorm sage legende vrouw geschiedenis dier held kind jong koning slot ontstaan moeder verschillend voorbeeld bevatten lezing feit fabel bekend overlevering thema nacht broeder vertelsel stof
Proportion: 0.002139306
Document: *Fabels en vertellingen der Kongo-negers.*
Document id: _gid001190901_01_0005 **Probability:** 0.9863033

Topic 46: 'n nou as frits da moeder nie moar oe an jan vragen veur niks mien deur vrouw mit roepen vader jong ion horen kijken sokrates doar triene lopen huis joa
Proportion: 0.0021368884
Document: *Hageveld. Roman van een klein-seminarist. (Fragmenten).*
Document id: _gid001191801_01_0058 **Probability:** 0.8758865

Topic 147: schrijver boek geschiedenis schrijven les lezen lezer zaak hoofdstuk behandelen vorm verhaal oordeel naam heer voorbeeld waarheid wetenschap [?] verklaren onderwerp werken vertaling

licht bewijzen historisch eeuw kritiek gebruik dans
Proportion: 0.0020914047
Document: *Korte berigten over boekwerken, vlugschriften enz., aankondigingen van vertalingen, letterkundig nieuws, enz.*
Document id: _gid001183801_01_0090 **Probability:** 0.49816602

Topic 256: god volk mythe mythologie naam voorstelling oorspronkelijk godsdienst hemel godheid zon indische aarde zin verklaring taal mensch wezenlijk vorm oorsprong schoon verschijnsel stam bekend ongetwijfeld geest natuur indra indiers mythologisch
Proportion: 0.0020907244
Document: *Vedenstudiën.*
Document id: _gid001187101_01_0038 **Probability:** 0.86763257

Topic 208: goud waarde zilver geld prijs munt wissel standaard metaal hoeveelheid daling vraag gouden gevolg koers oorzaak dalen ruilmiddel ricardo stijgen wisselkoers verhouding bedrag belangrijk crediet verandering millioen aanbod zilveren bank
Proportion: 0.0020779856
Document: *Wisselkoersen.*
Document id: _gid001188301_01_0006 **Probability:** 0.9751269

Topic 201: militair leger officier toestand staf chef scharnhorst volk opperbevel generaal generalen wapen taak opleiding dienst kracht eischen geest minister vraag gelegenheid behoren gebied oefening hoofd leiden buiten oorlog rang academie
Proportion: 0.0020646306
Document: *Over waardeering van strategische bekwaamheid*
Document id: _gid001193201_01_0082 **Probability:** 0.83455014

Topic 261: nou nee jij herman kijken vragen voelen 'n och meneer he kamer lachen horen ooog kind jaap hoofd jou vrouw ding gezicht zoo'n mensch praten johan lopen juffrouw jullie niks
Proportion: 0.002002603
Document: *In het voorbijgaan.*
Document id: _gid001190601_01_0114 **Probability:** 0.9129333

Topic 155: belasting gemeente heffing heffen accijns inkomst uitgaaf opcenten wet inkomen grondbelasting accijnsen direct huurwaarde falck afschaffing rijk opbrengst pct financien minister personeel belastingstelsel betalen inkomstenbelasting vermogen schatkist last belasten bedrag
Proportion: 0.001917357
Document: *Erratum.*
Document id: _gid001191901_01_0114 **Probability:** 0.7918121

Topic 43: othello jago vlaamsch volk vlaanderen mensch desdemona vrouw plicht shakspeare recht liefde gevoel vlamingen kracht hart maatschappij wereld cassio strijd natuur gevoelen schrijven begrijpen sterk daad kind speeksel vriend macht
Proportion: 0.0019078903
Document: *Eene Othello-studie.*
Document id: _gid001188301_01_0056 **Probability:** 0.99201

Topic 297: sultan vorst sumatra hoofd gezag atjeh bevolking bestuur expeditie palembang westkust rijk resident eiland gouvernement binnenland engelschen vestiging kampong onderwerping vestigen gevolg compagnie zenden rivier oorlog contract gebied bali padang
Proportion: 0.0018846266
Document: *De uitbreiding van het Nederlandsch gezag op Sumatra.*
Document id: _gid001188701_01_0080 **Probability:** 0.7677512

Topic 227: hy zy zyn gy wy brief myn schrijven valckenaer betje vriend hart zyne wolff letteroefening vrouw vaderland jaarg zig deeze myne heer lief beyma vader vv altoos mensch vryheid uitgeven
Proportion: 0.0018831844

Document: *In de Beemster-pastorie. (Naar onuitgegeven brieven.)*
Document id: _gid001190301_01_0016 **Probability:** 0.9592519

Topic 213: beweging hersenen prikkel orgaan bewustzijn licht indruk spier toestand sterk invloed voorstelling cel lichaam dier waarneming reactie verschijnsel evenwicht gewaarwording verschillend zenuw zenuwstelsel werking kracht functie psychisch zintuig waarnemen gevoel
Proportion: 0.0018651618
Document: *Mensch of automaat? Inleiding tot eene studie over het Hypnotisme.*
Document id: _gid001188801_01_0087 **Probability:** 0.9785575

Topic 161: prof dr cicero schrijven vs boek rinkes schrijver stuk redevoering lezen lulofs boot maerlant bekend brief eeuw uitgaaf naam inleiding uitgeven lezer kiehl voorbeeld cobet vertaling tekst summa zaak misschien
Proportion: 0.0018246793
Document: *Catilina.*
Document id: _gid001185701_01_0042 **Probability:** 0.8441512

Topic 23: water bodem vormen gesteente bron rand meter bedekken zout helling krater berg gedeelte diepte groen spleet grond oppervlakte vlak diep wit hoogte groeien anna_maria bosch naam voet vrouw vulkanisch koken
Proportion: 0.0018169014
Document: *Het Yellowstone-park.*
Document id: _gid001190501_01_0014 **Probability:** 0.99953514

Topic 41: krankzinnig gesticht darwin lijder recht strijd dier stichten theorie vraag schrijven wet verschillend gebied prof mensch rechtsbewustzijn feit patient natuurkeus voorbeeld krankzinnigengesticht vroeg bijzonder beschouwen geneesheer erkennen vorm verplegen verpleging
Proportion: 0.0017959253
Document: *Het openbare Krankzinnigenwezen, vooral met betrekking tot ons land.*
Document id: _gid001184801_01_0020 **Probability:** 0.709713

Topic 34: koning verhaal shakespeare willem dood frankrijk fransch gedicht eeuw ridder naam stuk karel held geschiedenis dichter persoon keizer vader heer historisch spanje graaf zoon verhalen vorst romeo heldendicht bekend horen
Proportion: 0.0017851443
Document: *De middeleeuwsche gedichten over Willem van Oranje.*
Document id: _gid001185401_01_0029 **Probability:** 0.9583361

Topic 207: bilderdijk da costa da_costa vriend bilderdijks wereld de_clercq geest capadose brief groen schrijven mensch eeuw hart amsterdam willem_de_clercq vader lezen gevoel misschien reveil huis tocqueville boek studie leiden ziel dichter
Proportion: 0.0017751405
Document: *De jeugd van Isaac da Costa. (1798-1823.)*
Document id: _gid001189301_01_0058 **Probability:** 0.9839735

Topic 278: grieksch grieken homerus griekenland ilias athene god dichter zeus plato apollo troje odysseus gedicht homerisch held griek godin eeuw achilles odyssee sage epos homeros agamemnon aarde hellas griekschen zoon zanger
Proportion: 0.001750271
Document: *Over bronnen en samenstelling der Ilias.*
Document id: _gid001191501_01_0094 **Probability:** 0.76232195

Topic 247: heer geschiedenis eeuw stad naam stuk broeder bekend gasthuis graaf drenthe schrijver kind schrijven vroeg heeren zoon mr boek broederschap bisschop gelderland vrouw overlijden lezen vader utrecht secundus holland behoren
Proportion: 0.0017242696
Document: *Het geslacht der Nicolai, en de portretten van Joannes Secundus.*
Document id: _gid001183901_01_0125 **Probability:** 0.93014956

Topic 273: tolstoi mensch ruskin schrijven kracht liefde werken wetenschap vrouw boek brief sterk gevoel dood arbeid kunst geestelijk maurice diep geloof waarheid geloven maatschappij geest rijk tacitus ding hart menigte kind
Proportion: 0.0016886289
Document: *Over Tolstoi.*
Document id: _gid001192901_01_0123 **Probability:** 0.9994511

Topic 146: expeditie tocht ijs water noorden diepte zee noordelijk breedte rivier richting m uur zuiden kust bevinden gebied waarneming temperatuur zuidelijk afstand groenland belangrijk maand kaart onderzoek diep vast oostelijk drijven
Proportion: 0.0016579747
Document: *Nansen's Noordpool-expeditie.*
Document id: _gid001189701_01_0002 **Probability:** 0.94139415

Topic 119: aarde planeet zon ster maan afstand sterrekunde beweging galilei wet waarneming schrijver hemel bekend massa wetenschap bepalen zonnestelsel saturnus komeet punt sterrekundige verklaring berekening stelsel sterrekundig mars verkrijgen schrijven as
Proportion: 0.0016546185
Document: *De wet van Kirkwood, toegelicht en beoordeeld.*
Document id: _gid001185301_01_0027 **Probability:** 0.99970007

Topic 89: martinus schedel gall verschillend vroeg geest bekend leren eeuw naam mensch stelsel schrijven prof vorm hoofdstuk voorhoofd trekken hoofd schrijver volk invloed studie beschaving horatius bevolking streek plato eenzaam denkbeeld
Proportion: 0.0016382774
Document: *De bevolking van ons vaderland.*
Document id: _gid001191201_01_0002 **Probability:** 0.7926092

Topic 150: beeld beeldhouwer kop grieksch lichaam kopie kunst houding eeuw hoofd antiek grond kunstwerk vorm bekend recht gedeelte groep arm beeldhouwwerk rome licht beroemd waarschijnlijk zeus origineel praxiteles vervaardigen relief dragen
Proportion: 0.0016336651
Document: *Antieke beeldhouwwerken in het Vatikaan.*
Document id: _gid001186901_01_0064 **Probability:** 0.9997298

Topic 220: vriend lezen david vrouw jong amsterdam vragen floris jacob avond kitty horen begrijpen wereld vader zoeken antwoorden doris schrijven boek stad volk mensch wisch starkadd wijck roepen trekken huis papier
Proportion: 0.0016166946
Document: *Eene Hollandsche revue.*
Document id: _gid001186701_01_0004 **Probability:** 0.9153872

Topic 210: dl dr j h mr i_blz w g m d l e jaargang boek prof amsterdam bladz nederlandsche geschiedenis gids overzicht nederland redactie jan jhr 2e pn bibliographie 1e inleiding
Proportion: 0.0015911142
Document: *Register der in dezen jaargang behandelde werken.*
Document id: _gid001186901_01_0067 **Probability:** 0.99714714

Topic 10: mohammed water machteld strand bart marretje marie grond zee moeder koran diep meter visschers vrouw mohammes heer vader kind sura schrijven trekken mennekens voet visch jong oever diepte vroeg huis
Proportion: 0.0015673728
Document: *Worstelend Zeeland.*
Document id: _gid001189801_01_0119 **Probability:** 0.7286398

Topic 133: volk eeuw kracht jong mensch heer lezen invloed schrijven taal vorm wereld schrijver zaak dichter waarheid nationaal werken ontwikkeling vreemd auteur kind rijk gevoel beweging doel nederlandsche fransch strijd gevoelen

Proportion: 0.0015427745
Document: *Pro Patria.*
Document id: _gid001189801_01_0083 **Probability:** 0.9277399

Topic 287: lessing boek droom nathan mensch lezen wereld peter waken vriend geest sokrates droomen misschien ding gevoel tete schrijven or dood kunst dromen gewoonte voelen hart jong liefde waarlijk stad begrijpen
Proportion: 0.0015163614
Document: *Simon Of over den droom aan Dr. H. Onnen Sr. Dialoog*
Document id: _gid001191801_01_0029 **Probability:** 0.89450467

Topic 243: ann juan kind lize cyril ooog francis broeke max gelaat antwoorden blik mira misschien plotseling voelen begrijpen moeder huis sieper waterhoek schelde trachten stem vragen bill schrijven jet lander hoofd
Proportion: 0.0015066469
Document: *Twee meisjes en ik*
Document id: _gid001193001_01_0080 **Probability:** 0.82380044

Topic 232: el water dijk veen dam zand veenen bodem polder bunder gedeelte rivier newcastle friesland groningen duin eeuw provincie ingenieur vroeg bentinck sandwich vloed chesterfield stanley voet mest kaart zuiderzee wad
Proportion: 0.001446265
Document: *Ameland.*
Document id: _gid001186901_01_0002 **Probability:** 0.98656714

Topic 71: moeder rijkert kind jong vrouw ooog hart letje huis felicia hoofd mensch diep zwijgen guido stem buiten voelen licht arm dood vader vragen wereld plotseling heur vroeg oma ziel dragen
Proportion: 0.0014262614
Document: *Mors et Vita.*
Document id: _gid001188701_01_0081 **Probability:** 0.60225827

Topic 86: ick iblz iiblz iiiblz ivblz ghy stuk dr hooft daer soo maer geschiedenis sy jonckbloet hy so schrijven haer amsterdam uyt aen breeroo myn waer gheen gedicht schrijver oock dichter
Proportion: 0.0014203918
Document: *De bronnen van Breeroos romantische spelen.*
Document id: _gid001188501_01_0020 **Probability:** 0.73397326

Topic 93: mensch liefde ziel mathilde boek wereld roman vrouw schrijven kunst dichter heer genieten kind jong kracht vorm brod kingsley zin aarde lilith gevoel natuur god gedicht voorwerp hart geest menschelijk
Proportion: 0.0014103687
Document: *De kunst van het genieten.*
Document id: _gid001191001_01_0168 **Probability:** 0.9950771

Topic 174: de_groot schrijven vriend brief cervantes parijs br recht vaderland huis vrouw spaansch misschien descartes vgl boek verblijf vroeg spanje frans ontvangen holland bekend bibl stuk maria zaak broeder licht amsterdam
Proportion: 0.0013766853
Document: *Hugo de Groot's tweede ballingschap.*
Document id: _gid001190301_01_0085 **Probability:** 0.90705377

Topic 91: bank kapitaal crediet pct papier millioen rente geld bankier nederlandsche_bank leening circulatie instelling circulatiebank bedrag crisis octrooi uitgifte publiek billet particulier voorschot middel gevaar beurs schuld zaak betaling waarborg operatie
Proportion: 0.0013651039
Document: *Giftvrij lichtgas.*
Document id: _gid001191301_01_0121 **Probability:** 0.92434525

Topic 98: recht grond agrarisch heer schrijven taal rijswijck eigendom volk horen belang vreemd beschikkingsrecht flanor sicilie lid vroeg zaak verklaren zoeken familie naam kind jong boer stuk lezen buiten bevolking vragen
Proportion: 0.0013515959
Document: *Uit Zuid-Afrika.*
Document id: _gid001187901_01_0053 **Probability:** 0.749831

Topic 19: gordon egypte egyptisch arabieren woestijn stam abdelkader algerie emir soedan troep regeering nijl beeld bevolking provincie khartoem maand zenden engelsch mahdi officier kameelen roo schrijven slaaf japansch khedive verkrijgen gouverneur
Proportion: 0.0013422495
Document: *Iets over photographie.*
Document id: _gid001185601_01_0033 **Probability:** 0.79230225

Topic 240: nederland nederlandsche minister lichting milicien belgie deg antwerpen gedeelte heer maand militie dienst dienstplicht oorlogsschip petermann toestand verkrijgen schip zee belgisch bataljon oefening rotterdam kanaal zaak mobilisatie kader militair dienstplichtig
Proportion: 0.001338973
Document: *De Eemsgrens*
Document id: _gid001193001_01_0102 **Probability:** 0.80029905

Topic 160: vrouw moeder liefde charlotte roman bourget kind boek schrijver hart graniet heer meisje jong wereld vragen maupassant verhaal hartstocht schrijven lezer persoon dame france geest zoeken contessa_margherita mensch huwelijk stijl
Proportion: 0.0013105347
Document: *Letterkundige kroniek.*
Document id: _gid001189501_01_0010 **Probability:** 0.70388865

Topic 183: jongen meisje lief horen mijnheer paul vertellen jong hervor orvarodd corrie vroeg zei vragen huis mensch avond oom mevrouw vriend zoo'n ding licht geloven hialmar roepen heer gelukkig dame antwoorden
Proportion: 0.0013103284
Document: *Uit de herinneringen van Adriaan Gildemeester.*
Document id: _gid001192501_01_0076 **Probability:** 0.95155364

Topic 199: compagnie batavia gouverneur handel china formosa valentijn canton zendeling hollanders malakka chineezen chineesche luzac hollandsche schrijven zenden regeering zending vv britsche donker oostindische_compagnie fort macao raad factorij nederlandsche thee gouverneurgeneraal
Proportion: 0.0013012142
Document: *Onze vaderen in China.*
Document id: _gid001191701_01_0036 **Probability:** 0.8179531

Topic 196: jan huis heur jongen lijk vina boer buiten ding vragen zoeken wijf kwaad mensch geld voelen dorp peerden eten werken lijf vast rond dolf pastor keren luid hof ginder kijken
Proportion: 0.0012723166
Document: *Langs de wegen.*
Document id: _gid001190201_01_0001 **Probability:** 0.7650742

Topic 102: taal fabel vorm eeuw invloed chr vroeg geschiedenis volk ornament dialect uitdrukking dier voorbeeld grieksch beteekenis bekend ontwikkeling naam zuiden schrijven noorden verschillend fries germaansch verdam wijzen boek verschil indische
Proportion: 0.001242228
Document: *Over Grieksche en Indische fabels.*
Document id: _gid001190301_01_0101 **Probability:** 0.93673843

Topic 108: necker eeuw geestelijk geschiedenis wereldlijk kerk paus schrijver vereeniging waldens stael augustinus vraag geest waarheid gebied vroeg bewijzen dier voorstelling schrijven middeleeuwen

rome waldenzen licht bisschop verhandeling stuk bewijs onderzoek
Proportion: 0.0012185575
Document: *Over de vereeniging van de geestelijke en wereldlijke oppermagt in den kerkelijken staat.*
Document id: _gid001185301_01_0001 **Probability:** 0.8167274

Topic 63: maurice 'n soonbeek voelen hm flora louise geld vrouw mensch mevrouw macht fleury zoo'n jagen meneer dood goethe liefde angst natuur stem genot begrijpen ooog durven prachtig heelemaal horen veiling
Proportion: 0.0012029684
Document: *Kunstenaarsleven.*
Document id: _gid001190601_01_0072 **Probability:** 0.6418728

Topic 170: curtius mulder eeuw naam vroeg bekend studie boek oudheid student geschiedenis geschrift schrijven dioscuren grieksch bron schrijver vraag zaak gymnasium geneeskundigen engeland trekken engelsch jong school bericht aristoteles denkbeeld vragen
Proportion: 0.0012024537
Document: *Angelen en Saksen op de Friesche terpen.*
Document id: _gid001191901_01_0059 **Probability:** 0.85190094

Topic 20: napoleon chesterfield wellington waterloo pruissisch pruissen slingelandt uur blucher amsterdam grouchy keizer quatrebras townshend ney divisie korps slagveld aanval brigade ligny stuk bredero walpole leger thiers schrijven strijd erlon punt
Proportion: 0.0011852041
Document: *Thiers' beschrijving van den veldtogt van 1815.*
Document id: _gid001186201_01_0058 **Probability:** 0.76821464

Topic 167: student wetenschap hoogeschool onderwijs sarcey les examen faculteit professor studie vak geest hoogleeraar naam klikspaan modderman akademische misschien betrekking college leeraar onderwerp jong zaak maatschappij besluit belang kennis leiden artikel
Proportion: 0.001171565
Document: *Hooger onderwijs.*
Document id: _gid001186801_01_0015 **Probability:** 0.90350837

Topic 193: jury e gezworene stuk schrijven terentius proces menander uitspraak zaak regel dekker plautus heer engeland engelsch bewijs beschuldigen gothisch voorbeeld pro schrijver tweeklank arm siegenbeek procureur deo aard spelling komedie
Proportion: 0.0011680498
Document: *De Nederlandsche spelling.*
Document id: _gid001186201_01_0032 **Probability:** 0.85075396

Topic 222: bets huug hans olga vrouw sofie maarte jong voelen vragen martini marie kind agnes schrijven freddy rob josua brief misschien lijken begrijpen huub huis lief van_de_willigen gelukkig horen lachen maarten
Proportion: 0.0011665311
Document: *De gave gulden*
Document id: _gid001193401_01_0058 **Probability:** 0.8879816

Topic 77: m kanaal rivier water friesche rijn sluis semietisch naam scheepvaart vaart friezen friesland waterweg verbinding diepte breedte waal maas verbetering friesch verkeer vormen waterstand stroomen lengte heer beteekenis ijmuiden zee
Proportion: 0.0011352934
Document: *Waterwegen in Nederland.*
Document id: _gid001191901_01_0096 **Probability:** 0.6568037

Topic 250: liszt schrijven bulow gids brief parijs stuk redactie artikel lodewijk concert van_wagner lohengrin tijdschrift blad berlioz plan wagner quack groningen publiek werken chopin juni prins saintsaens kunst tannhauser emden redacteur

Proportion: 0.0011349929

Document: *Wagner in Frankrijk.*

Document id: _gid001188701_01_0089 **Probability:** 0.6849975

Topic 153: mevrouw mijnheer markies francois prinses diana orsini vrouw aubigny abt hernemen sieur jong hertog dubois frankrijk sainbertot begrijpen armentieres dame graaf baronne spanje gravin monseigneur edelman waarheid parijs chanteloup camereramajor

Proportion: 0.0011330218

Document: *Diana.*

Document id: _gid001184601_01_0064 **Probability:** 0.5733358

Topic 264: maria god job richard vader hart vriend michiel irene mensch seneca gods lijden ep kind staeg lib liefde diep boer geluk aagje ooog holland zonde slaan ding gedicht dood franschman

Proportion: 0.001115697

Document: *Het leven dat wij droomden*

Document id: _gid001193101_01_0120 **Probability:** 0.79754454

Topic 228: dolf leeuw saul brief plinius paula dier rob sam circus trajanus boek schrijven geld vrouw vragen zestig napels tenslotte naam mensch vriend temmer brussel stuk misschien briefwisseling keizer krant publiek

Proportion: 0.0011083422

Document: *In Napels hongeren zestig leeuwen*

Document id: _gid001193401_01_0027 **Probability:** 0.70933664

Topic 230: hamlet shakespeare ophelia koning geest hamlets vader polonius prins claudius daad perspectief horatio moeder laertes natuur koningin plaatsen brandes brutus omgeving gramberg denemarken vriend glas schijn zoeken vorm stuk kunstenaar

Proportion: 0.0011060858

Document: *Hamlet.*

Document id: _gid001188201_01_0035 **Probability:** 0.81747395

Topic 293: wet lijfsdwang partij heer zetel harold art recht methode middel feit artikel stem schuldenaar mill vraag eodem zaak verdeeling minister bijzonder stammler grond blad vragen dier misschien publiek naam kind

Proportion: 0.0010896119

Document: *De verdeeling der zetels over de verschillende partijgroepen bij evenredige vertegenwoordiging.*

Document id: _gid001191101_01_0121 **Probability:** 0.8129819

Topic 214: noorwegen thiss titia noorsche licht wind zee christiania keller zweden stem rita hart horen geluk wereld koning fjord mensch ooog buiten vrouw diep zingen noren donker schoon lopen machtig roepen

Proportion: 0.0010842539

Document: *Een juli-dag.*

Document id: _gid001189201_01_0053 **Probability:** 0.7202287

Topic 179: gild raad bali taal vreemd japansch verbieden stad lid eeuw baliers balisch utrecht japan ha betalen eiland bepaling bedrijf verschillend gildebroeder burger yedo brouwer zaak buiten verkoopen laken invloed recht

Proportion: 0.0010686363

Document: *De gilden en het regeeringstoezicht op handel en nijverheid in de middeleeuwen.*

Document id: _gid001189701_01_0060 **Probability:** 0.94082373

Topic 288: auteur conventie vertaling auteursrecht werken bescherming deensch schrijver recht denemarken no artikel uitgeven uitgever denen verschijnen nederland vertalingsrecht dr berner berlijn unie boek uitsluitend schrijven nederlandsche oorspronkelijk zaak bepaling taal

Proportion: 0.0010552481

Document: *De Berner-Conventie, te Berlijn herzien.*

Document id: _gid001190801_01_0147 **Probability:** 0.76451635

Topic 60: pasteur ziekte bacterie water onderzoek visch verschillend dier kg proefstation visschen vangst invloed smetstof schadelijk oorzaak voorkomen visscherij stof onderzoeking wetenschappelijk gisting middel ziek gebied belang verkrijgen bekend koch jong
Proportion: 0.001047388
Document: *Het onderzoek van drinkwater.*
Document id: _gid001189301_01_0036 **Probability:** 0.95150447

Topic 18: brederode lodewijk groen prins heer landvoogdes brief tang schrijven edele boeddha willem koning verbond hoorne boek t_ii boeddhistisch archives antwerpen egmont hoogstraten geestelijk keizer convent smeekschrift zaak broeder arch vianen
Proportion: 0.001043363
Document: *I. Hendrick graaf van Brederode, mede-grondlegger der Nederlandsche vrijheid, verdedigd door Mr. M.C. van Hall , Staatsraad, enz. Met Platen. Amsterdam, Johannes Müller, 1844. xvi, 241 blz. 8 o . H. Antwoord aan M r . M.C. van Hall, staatsraad, enz. (over a. Hendrick, Graaf van Brederode; b. Uitgave van Brieven; c. Historische Kritiek), door Mr. G. Groen van Prinsterer . Leiden, S. en J. Luchtmans. 1844. 104 blz. 8 o .*
Document id: _gid001184501_01_0025 **Probability:** 0.80914646

Topic 110: huygens ick vondel vs cats ged poezie stuk vers hooft gedicht dichter vgl schrijven soo haer daer aen volk vlgg constantyn diamant dier eeuw geest ghij tonen werken christiaan maer
Proportion: 0.0010368885
Document: *Constantyn Huygens.*
Document id: _gid001190001_01_0060 **Probability:** 0.98316807

Topic 290: kunst teekenen vorm frans platoon kunstenaar leerling geest nijverheid voorbeeld gedeelte les van_effen verschillend vormen denkbeeld schrijven school leren ornament schoon werken model middel industrie vreemd smaak stijl gevoel taal
Proportion: 0.0010346817
Document: *Eene vraag des tijds.*
Document id: _gid001186401_01_0035 **Probability:** 0.73235404

Topic 83: thorbecke koning hall minister schimmelpenninck ministerie schrijven moleschott brief heer commissie zm donders vriend groen mr van_de_brugghen lid deen maart kempenaer zaak tweede_kamer ontvangen heidelberg utrecht rappard ontslag simons betrekking
Proportion: 0.0010224764
Document: *Thorbecke's Geldersche reis VI. - W.A. Schimmelpenninck van der Oye aan L.N. van Randwijck, de Poll 26 Nov. 1852.*
Document id: _gid001193201_01_0130 **Probability:** 0.68591213

Topic 106: kleist brief caroline schrijven vriend baco macaulay vrouw liefde thomson berlijn geest schelling hart essex broeder graaf licht schlegel huis vragen vriendin toestand jena mensch naam mainz indruk dood vroeg
Proportion: 0.0010083914
Document: *Caroline.*
Document id: _gid001187201_01_0009 **Probability:** 0.8724972

Topic 267: wetenschap water leeuwenhoek mikroskoop kennis dr boek bekend onderzoek zaak licht lewes schrijven bladz duin amsterdam comenius bredero helmholtz theorie mensch euangelie vormen natuur vriend belangrijk landbouwschool gebruiken scheikunde verkrijgen
Proportion: 0.0009580832
Document: *De drinkwatervoorziening in ons land voorheen en thans.*
Document id: _gid001192201_01_0050 **Probability:** 0.7099251

Topic 254: slaaf neger suriname slavernij kolonie kind fabriek blank toestand emancipatie werkman werken plantage vrouw arbeid heer paramaribo wet slavin uur meester westindische jong gouverneur verschillend surinaamsch moeder vroeg nacht de_neger
Proportion: 0.00095638423

Document: *De arbeidsenquête.*

Document id: _gid001188701_01_0052 **Probability:** 0.86697316

Topic 172: taal schrijven kind vertaling leerling boek lezen vertalen school leesboek schrijver vorm hollandsch heer zin onderwijs vertaler gebruiken lay moedertaal proza beteekenis leren uitdrukking mensch onderwijzer onderzoek van_deyssel begrijpen zuiver

Proportion: 0.0009561685

Document: *Kunsthistorische methoden.*

Document id: _gid001192601_01_0123 **Probability:** 0.96990836

Topic 112: lijn x SS vlak punt d figuur hoek y vergelijking wiskunde cirkel reeks toepassing krom oppervlak evenwijdig e oplossing formule term behandelen q oneindig vraagstuk plat m ruimte bepalen buigen

Proportion: 0.0009495811

Document: *Vervolg op de Beginselen der Hoogere Meetkunst , bevattende de Theorie der gebodene oppervlakken en kromme lijnen van dubbele kromming, benevens Formules voor de Hoogere Meetkunst, door J. Jonkhert , Math. et Phil. Nat. Cand.*

Document id: _gid001183801_01_0042 **Probability:** 0.9985502

Topic 90: linnaeus lucht mohr meter punt stelsel uur recht tov werken top tunnel water r3 gids vlak herakles heer observatorium teste jong buiten kracht licht stuk hoogte beneden les upsala trekken

Proportion: 0.00094937655

Document: *Een dag aan den Mont-Cenis-tunnel.*

Document id: _gid001186801_01_0079 **Probability:** 0.9628266

Topic 82: victor eva marga elsje hermine liefde nout richard kaptein voelen richardson heer edith vrouw werken roman gevoel clarissa berghem verlangen jong begrijpen lovelace werkelijk misschien gelukkig bertus fabriek revolutie gedachte

Proportion: 0.0009355872

Document: *Demonen.*

Document id: _gid001192001_01_0042 **Probability:** 0.630464

Topic 7: schrijven schrijver boek de_foe naam zaak heer bekend kielland geschrift geest rijk jong stuk publiek wereld lezen licht augier hart vroeg naber uitgever hoofd lezer verhaal werken persoon verschijnen pen

Proportion: 0.0009355362

Document: *Daniel de Foe.*

Document id: _gid001187001_01_0023 **Probability:** 0.77812606

Topic 51: boek leopold naam voorbeeld tekst schrijver eeuw vertaling lezen regel zin plein shakespeare ampere heer brochure schrijven vraag gedicht oorspronkelijk uitgave straat belangrijk verschillend waarde vorm werken bekend lezing voorkomen

Proportion: 0.00093334966

Document: *Shakespeare?*

Document id: _gid001190901_01_0149 **Probability:** 0.93783593

Topic 182: japan roode_kruis japansch volk mensch vereeniging oorlog lid rus hulp russisch massa dienst japanners zaak gewond vroeg schrijven vrouw vreemd westerling compagnie menigte anarchisme staat jokohama zm kennis werken departement

Proportion: 0.000930444

Document: *Het Roode Kruis van Japan.*

Document id: _gid001190601_01_0014 **Probability:** 0.83610606

Topic 61: amsterdam stelling haarlem manzoni polder heer drukken water haarlemmermeer kunst inundatie uitvinding boekdrukkunst ap holland m zuiderzee millioen vijand uitvinder rijnland waarschijnlijk fort eer lucia kanaal noordzee mainz werken stad

Proportion: 0.00092999253

Document: *Zwakke zijden.*

Document id: _gid001187501_01_0057 **Probability:** 0.9065654

Topic 206: deken martha betje hildebrand meter vriendin boek hectare schrijven roman geest sentimenteele wolff droogmaking jong bezit willem diep werken hart molengraaff licht leevend vriendschap liefde brief mensch kruger druk plan
Proportion: 0.0009282653
Document: *De samenwerking van Wolff en Deken.*
Document id: _gid001192201_01_0091 **Probability:** 0.5304141

Topic 75: straal theorie atoom lijn meetkunde punt licht natuurkunde energie snelheid trilling negatief golflengte getal electrisch positief wet afstand verhouding figuur bijv physicus eigenschap electroon verschillend seconde cirkel aard spectrum bepalen
Proportion: 0.00090125954
Document: *De quantentheorie van Max Planck.*
Document id: _gid001192701_01_0060 **Probability:** 0.80157954

Topic 135: jezus opium florence verhaal tsjechow lukas liefde vissering schrijver matthaeus ziel schrijven lief boek zin sluikhandel schl mensch lezen luc voelen matth kist christus odo diep smokkelhandel gouden johannes goddelijk
Proportion: 0.00090053806
Document: *Stijn Streuvels' Minnehandel.*
Document id: _gid001190401_01_0032 **Probability:** 0.58469343

Topic 175: inkomen belasting opbrengst vermogen masaryk inkomstenbelasting kapitaal arbeid waarde inkomst bedrag staat beschouwen tsjechen geld ontstaan vraag grondslag bron grond oedipus prijsstijging vertering behoefte middel genot drukken doel aanmerking verkrijgen
Proportion: 0.0008892377
Document: *Problemen der inkomstenbelasting.*
Document id: _gid001192701_01_0067 **Probability:** 0.923641

Topic 205: vennootschap keizer w waldeck hertog wallenstein dl nationaliteit d schrijver naamlooze dr zaak zweden aandeelhouder willem forster wirth wallensteins vreemd leger letteroefeningen keurvorst pl heer officier vreemdeling bekend persoon nos
Proportion: 0.00087842526
Document: *De zoogenaamde nationaliteit der naamlooze vennootschap.*
Document id: _gid001191801_01_0053 **Probability:** 0.6592357

Topic 47: rembrandt saskia roman kind boek schrijven persoon schrijver vrouw wereld licht mensch liefde donders ding jong jesaja heyse krul goncourts jes hoofdstuk hanslick heer_r model cyrus literatuur karakter de_barok bewijzen
Proportion: 0.0008776765
Document: *Een didactische roman.*
Document id: _gid001187401_01_0036 **Probability:** 0.9131386

Topic 294: erasmus smyrna huygens ick vrouw schrijven heer hooft constantin tesselschade lydia romola jong wisby eeuw dichter tito vondel stuk leendertz ghy naam vader liefde ephese bekend bjornson arabiese licht lezen
Proportion: 0.0008742125
Document: *De liefdesgeschiedenissen van twee Nederlandsche dichters.*
Document id: _gid001187101_01_0009 **Probability:** 0.8028468

Topic 115: preek leerrede tekst lid commissie synode gemeente vergadering klassikaal kerkbestuur provinciaal bundel hoorder vertegenwoordigen reglement bestuur kiezen synodaal rapport auteur onderwijzeres stem benoemen onderwerp lezen curator bepalen gelegenheid art ten
Proportion: 0.00086644845
Document: *De Vertegenwoordiging der Nederlandsche Hervormde Kerk door de Synode*
Document id: _gid001184801_01_0047 **Probability:** 0.81372094

Topic 265: francesco grootvader jacobi uberto agata luyken god schrijven savina tito giovanella lezen gherardo letizia mino volk trebiano pietro alba soldaat brief schoolonderwijs hemel spanje zaak el

juan jong y spinoza
Proportion: 0.0008435278
Document: *De Nederlandsche opstand in de Spaansche letteren*
Document id: _gid001193001_01_0108 **Probability:** 0.71811736

Topic 4: mensch aap wetenschap ontwikkeling economisch staathuishoudkunde menschelijk vraag gebied dier naam eeuw verschijnsel invloed doornik natuur heden historisch aarde feit verschillend streven beschouwen heim wijzen vin jong volksgemeenschap volk boek
Proportion: 0.0008400978
Document: *Retardatie en foetalisatie, een nieuwe beschouwing van het vraagstuk der anthropogenese.*
Document id: _gid001192701_01_0058 **Probability:** 0.57344794

Topic 276: jozette madame aristide jeanne celestin monsieur carpentier mademoiselle leguenne lourty etienne herz vrouw van_loo nee frank dutoit alphonse tuin tafel hortense raam loge tom barbara ooog huis gezicht avond gabrielle
Proportion: 0.0008355315
Document: *Vedeldeuntjen, april 1860.*
Document id: _gid001186001_01_0024 **Probability:** 0.87035304

Topic 192: reelen anna_paulowna 'n polder zuster anna drank martens sterk maatschappij vrouw moeder brandewijn vader voelen hotel berg zon huis kind opnieuw trouwen calamiteuse stad gesprek bestuurder vroeg beneden stien nou
Proportion: 0.0008235939
Document: *Een trouwdag.*
Document id: _gid001191701_01_0095 **Probability:** 0.77183676

Topic 277: kanaal steven pericles el water zee gedeelte suez grond zand mijl stad dezelve ondermeester voet knip gevoel vroeg diep roode_zee verheffen bevinden vormen middel eberhard hoogte bijzonder werken geest middellandsche_zee
Proportion: 0.00080253626
Document: *Zandkorrels uit de landengte van Suez.*
Document id: _gid001186901_01_0021 **Probability:** 0.9794144

Topic 44: frank bertie nel eve balfour sir vragen bake bessie phil begrijpen ooog patty voelen god misschien hoofd horen tim will bang genade kind vreemd arm stem brief deur vroeg archibald
Proportion: 0.00080183306
Document: *Noodlot.*
Document id: _gid001189001_01_0086 **Probability:** 0.8560992

Topic 184: doodstraf duitsche congres regeering rodenburgh dood britsche misdadiger lombroso vraag lamprecht duitschland grond armande criminaliteit celia politiek bismarck moord bekend stuk lulli vragen rapport gebied zaak lope academie koloniaal recht
Proportion: 0.0007960644
Document: *Karl Lamprecht.*
Document id: _gid001191501_01_0096 **Probability:** 0.7401085

Topic 131: heer brederode groen schrijver hall boek stuk schrijven zaak vroeg lezen reggen regt grond bladz amsterdam verdriet geschiedenis naam houtsoort brief hoofdstuk jeronimus bewijs heeren brants waarheid hoofd mensch beschouwing
Proportion: 0.0007948755
Document: *I. Hendrick graaf van Brederode, mede-grondlegger der Nederlandsche vrijheid, verdedigd door Mr. M.C. van Hall , Staatsraad, enz. Met Platen. Amsterdam, Johannes Müller, 1844. XVI, 241 blzz. 8 o . II. Antwoord aan m r . M.C. van Hall, staatsraad, enz. (over a. Hendrick, Graaf van Brederode; b. Uitgave van Brieven; c. Historische Kritiek), door Mr. G. Groen van Prinsterer . Leiden, S. en J. Luchtmans. 1844. 104 blz. 8 o .*
Document id: _gid001184501_01_0022 **Probability:** 0.6860827

Topic 246: art kerkmeester zaak schutterij militie heer admiraal kerk dienst rijk vast kader regeering japan de_ruyter tijdelijk keizerlijk jedo oefening provincie actief wedstrijd ick lieden abh rang verschillend brief h breero

Proportion: 0.0007933674

Document: *De Ruyter's journaal tijdens de expeditie naar Denemarken.*

Document id: _gid001190801_01_0002 **Probability:** 0.7690394

Topic 35: indianen tao confucius vogue indiaan indiaansch volk ding hoofdstuk mensch boek china lao_tsz rijk naam schrijven chineesche zendeling li tollens blank chateaubriand paula amerikaansch sterven dood jong lezen hart wereld

Proportion: 0.00078174716

Document: *Wu Wei. Eene studie, naar aanleiding van Lao Tsz's filosofie.*

Document id: _gid001189701_01_0109 **Probability:** 0.99737173

Topic 252: taine dansen dans duncan beweging inkomen gebaar putte meisje kunst belasting miss kracht lichaam danskunst vraag inkomstenbelasting dienst schrijven stand beteekenis aulard mensch persoon volstrekt punt muziek heer incometax begrijpen

Proportion: 0.00078136777

Document: *De inkomstenbelasting.*

Document id: _gid001187101_01_0001 **Probability:** 0.69209474

Topic 17: buys strauss luther vriend hut duitschland schrijven strijd overl brief volk ziekte moeras geb erasmus eeuw hoogleeraar vroeg naam turkije hutten wetenschap ulrich ballot ridder wereld geschiedenis rome studeren psychiatrisch

Proportion: 0.0007781532

Document: *Ulrich van Hutten.*

Document id: _gid001185801_01_0024 **Probability:** 0.91754615

Topic 209: ambtenaar taal onderwijs dienst indie indische kennis lijst javaansch stelsel opleiding vak zetel delft studie instituut administratief inrigting examen besluit instelling maleisch verdeeling betrekking de_akademie bepalen bepaling bedoelen volkenkunde gevolg

Proportion: 0.00077319844

Document: *De Delftsche akademie als instituut tot opleiding van ambtenaren voor Neerlandsch Indië.*

Document id: _gid001185301_01_0033 **Probability:** 0.97242314

Topic 242: marokko columbus baldaeus naam lezen schrijven eeuw grant congo taal bekend stad nestorianen kerk boek marokkaansch ceylon stuk vader zoeken hooft vroeg nestoriaan anslo zoon verklaren schrijver spanje heer leopold

Proportion: 0.00077145366

Document: *Columbus voor 1492.*

Document id: _gid001189201_01_0047 **Probability:** 0.6395272

Topic 259: marx treub gemeente hegel boek kapitaal schrijven wet arbeid engels systeem vraag progressie schrijver methode waarde wetenschap theorie starter werkelijkheid beteekenis materie prof recht kuiper algemeene zin sociaal heer von_scheffel

Proportion: 0.00076909194

Document: *Professor Treub's Marx.*

Document id: _gid001190401_01_0021 **Probability:** 0.57100546

Topic 24: neus dr doedes spanje mensch boek janus stevenson spaansch schrijven tonia lezen joseph vader wereld else prof geest leren misschien god tin ding naam groninger kind buiten groningers stuk spanjaard

Proportion: 0.0007645783

Document: *Over neuzen.*

Document id: _gid001183901_01_0122 **Probability:** 0.9791418

Topic 70: kind moeder vrouw ziel plant kracht naam boom tak rakshasa leggen vaderschap verbod geest dier dood koning grond wortel lichaam molieres stuk mensch von_schack dragen sterven onderzoek bloem volk roepen
Proportion: 0.00076396734
Document: *Het volksgeloof aan het bovennatuurlijke in het rijk der planten.*
Document id: _gid001188101_01_0043 **Probability:** 0.83819985

Topic 215: floris pascal historisch eeuw koning ronsard blancefloer bossuet hooft les leycester vrouw geschiedenis schrijven roman vijftien kunst lezen malherbe vorst jezuiet eon persoon hart graaf naam stuk tartufe taal licht
Proportion: 0.0007637835
Document: *De graaf van leycester in nederland, door A.L.G. Toussaint .*
Document id: _gid001184701_01_0032 **Probability:** 0.92775285

Topic 129: george ierland iersche gladstone lord sir miss koningin victoria kolonel mr oom engeland londen ieren mary zaak chamberlain engelsch brief mrs philip dublin louise ier wereld christina filips partij beaconsfield
Proportion: 0.0007614102
Document: *Buitenlandsch overzicht.*
Document id: _gid001192101_01_0117 **Probability:** 0.46412745

Topic 67: pensioen ambtenaar staat m philips afdeelingen fonds scheffel hettner lid timon antonio commissie weduwe maatschappij wet portugal algemeene pensioenfonds kind dier memorie graanwet zaak dood schrijven muziekaal buiten dienst omstandigheid
Proportion: 0.00075672916
Document: *Ambtenaarsbelangen.*
Document id: _gid001188301_01_0034 **Probability:** 0.62761265

Topic 137: heer lucht valentin schrijver economisch palm verhouding vorm houten koolstofzuur hoeveelheid omstandigheid beschouwing verschijnsel zuurstof verschil stikstof uitademen bepalen grond berekenen vroeg invloed verschillend bekend oorzaak water gevolg proef gedeelte
Proportion: 0.0007563115
Document: *Natuurkunde van den mensch, door G. Valentin , Hoogleeraar te Bern. Uit het Hoogduitsch, door J.G. Rooseboom , M.D. te Gouda. Met 234 tusschen den tekst geplaatste houtsneê-figuren. Gouda, G.B. van Goor, 1845, 2 deelen 8 o ., 764 en 930 blz.*
Document id: _gid001184601_01_0033 **Probability:** 0.86547476

Topic 87: geel peer kohlbrugge frank da_costa dio doopsgezinde peer_gynt schrijven villetard weerloosheid god louis doopsgezinden godsdienst lezen vriend christen sietske mensch vosmaer de_clercq geest mensinga dragen reisje brief grieksch kunst beets
Proportion: 0.0007474414
Document: *Jongere tijdgenooten.*
Document id: _gid001188901_01_0022 **Probability:** 0.8566549

Topic 22: naam water zwaan eeuw schrijver olympia trekken sage plan beteekenis onderzoek gebied invloed volk grond germaansch vroeg zaak licht streek bekend zee stam god ridder verschillend toestand cornelius regge stuw
Proportion: 0.0007445357
Document: *Over den oorsprong van den Ridder met den zwaan.*
Document id: _gid001189401_01_0093 **Probability:** 0.8585763

Topic 235: joden militair officier onderofficier uitgaaf dienst leger infanterie oefening kazerne wapen korporaal soldaat jood kader begrooting cavalerie verschillend compagnie dienstplicht soldij samen korps regiment som bedragen formatie geld luitenant koning
Proportion: 0.00073947984
Document: *De Nederlandsche oorlogsbegrooting, voorheen, thans en in de toekomst.*
Document id: _gid001189701_01_0028 **Probability:** 0.94472736

Topic 249: romeinsch sheridan alice eeuw vesting vroeg schrijver naam granville karakter voorstelling bataven waarschijnlijk bekend zoeken onderzoek vechten hoofd castellum werkelijk romeinen laocoon doel bouwen stuk beteekenis nederzetting dier lezer zaak
Proportion: 0.0007346188
Document: *Het eerste hoofdstuk onzer vaderlandsche geschiedenis.*
Document id: _gid001191001_01_0019 **Probability:** 0.8287444

Topic 130: socialisme socialistisch arbeider socialist hart villier patroon hoorne marx vb anjou debussy sorel schrijven marxisme boek kapitalisme revolutie overeenkomst bekend antwoord partij kamer strijd patroonsvereeniging onmiddellijk stad zenden buiten heer
Proportion: 0.00072677626
Document: *De groote lockout in Denemarken.*
Document id: _gid001189901_01_0097 **Probability:** 0.69314224

Topic 237: schr ref zuidafrika afrikaansch taal volk boer dr geschiedenis afrikaners hollandsch engelsch ijsland ijslandsche afrikaner schrijven kaapstad nederlanders nederlandsch reykjavik hollandsche wet patricier ds pretoria heer tribuun margaretha neumann prof
Proportion: 0.0007202671
Document: *Het Hollandsch in Zuid-Afrika.*
Document id: _gid001190801_01_0116 **Probability:** 0.9654393

Topic 96: lied zingen volkslied carlyle volk muziek melodie couplet verzameling eeuw volks irving 16de maat toon mond horen oosten toonkunst edinburg 15de westersche bekend vermeulen oostersche wijs maal dichter volksdichter schrijven
Proportion: 0.00071197894
Document: *Een Straatliedje, van Piet Bogcheljoen.*
Document id: _gid001184601_01_0041 **Probability:** 0.96630967

Topic 55: spengler instrument haydn orkest livius niebuhr ennius muziek nationaalsocialisme viool trompet schrijven vorm annale gebruik werken cicero hobo eeuw instrumentaal toon vormen gebruiken annalen romeinsch maximi verschillend album bazuin symphonie
Proportion: 0.0007043705
Document: *Muzikaal overzicht.*
Document id: _gid001190701_01_0125 **Probability:** 0.71250075

Topic 224: kautsky vh ekonomisch goethe maatschappelijk ii_blz maatschappij volk moraal geschiedenis iii_blz iv_blz drift arnold zedelijk gorter wereld ontwikkeling boek sociaal kracht verhouding schrijver frankrijk macht belang werken gevoel zending politiek
Proportion: 0.0007026764
Document: *Histoire de France en History of England.*
Document id: _gid001192701_01_0018 **Probability:** 0.75911814

Topic 52: woning lombok huygens marinier bouwen koning willem_iii PS londen gemeente persoon vermelden arbeiderswoning korps huis journaal mensch volkshuisvesting onteigening stad bekend dagboek goncourt perceelen gelegenheid plan engelsch huisvesting betrekking buurt
Proportion: 0.0006992133
Document: *Het woningvraagstuk in eenige Britsche steden.*
Document id: _gid001190201_01_0110 **Probability:** 0.6001733

Topic 79: javaansch java taal heer roorda spreekwoord bataviaasch_genootschap gericke genootschaps javanen vertaling eysinga raffles homoeopathie woordenboek genootschap boek studie uitgave spraakkunst kawi javaan dr javaansche maleisch bijbelgenootschap crawfurd sumatra soerakarta schrijver
Proportion: 0.0006969276
Document: *Tot de onafhankelijkheid, onpartijdigheid en beleefdheid van een onzer vaderlandsche tijdschriften.*
Document id: _gid001183901_01_0145 **Probability:** 0.91247547

Topic 69: cromwell parlement koning macbeth engeland strijd kracht beowulf macht volk engelsch deken recht diep naam lady_macbeth omstandigheid neptunus jong eeuw grendel heks werkelijkheid zaak vriend vroeg held hoofd werken congo
Proportion: 0.0006938043
Document: *Nieuw-Amerika.*
Document id: _gid001191801_01_0126 **Probability:** 0.77516407

Topic 116: willem baltus theo meneer klara lucette vragen boer moeder vader hart opeen dorp heer dorry misschien hoeve jong horen pastoor lezen andre oorlog mensch boek recht heuvel mina leonce vroeg
Proportion: 0.00069369766
Document: *Stemmen uit de redactie*
Document id: _gid001193401_01_0067 **Probability:** 0.53809667

Topic 81: joden chinesche ballon taal volk naam boer jood schrijven stad werken china rijk chinezen boek lezen zendeling amerika bekend vertaling verschillend luchtschip christen ds zaak engelsch hoogte christendom gedeelte holland
Proportion: 0.00069222954
Document: *Robert Morrison.*
Document id: _gid001184901_01_0029 **Probability:** 0.8260023

Topic 14: visscherij water haren vaartuig pantseren huygens leiden monitor stuk ram veluwe acte graaf voet hertog kustverdediging tromp hoogheid zaak eeuw toren ithaka schip vroeg visschen stad lezen gebruiken prof beteekenis
Proportion: 0.00068716146
Document: *Eenige bedenkingen over de vroonvisscherij van Leiden.*
Document id: _gid001185801_01_0030 **Probability:** 0.91791815

Topic 128: ziel beeld zee licht mensch vrouw dienstbode chineesche kwan_yin onzijdig ding boek boeddha chinees gewaad onzijdigen ooog heinrich china wonder heilig berg lichaam britsche blank schoonheid recht kunst wit rein
Proportion: 0.0006806177
Document: *Kwan Yin, De godin der genade.*
Document id: _gid001189601_01_0013 **Probability:** 0.93307596

Topic 138: maartje guus boek van_hooren ziekenhuis misschien been goa dood mutsaers dr cato mensch nietzsche vriend edmond kracht sterk maalwijck charles hart stad schrijven baars van_maalwijck arm spier doctor doll vrouw
Proportion: 0.00067143503
Document: *Oerleven.*
Document id: _gid001192101_01_0016 **Probability:** 0.6425746

Topic 154: heer lodewijk brief koning schrijven zaak maria_antoinette dr_nuyens kambodja beeldenstorm reden verhulst nassau strada plan bewijzen graaf naam uitgaaf invloed gen mechelen hoofd persoon loge geschiedenis groen kennis journaal schrijver
Proportion: 0.0006654836
Document: *De Elzeviers.*
Document id: _gid001188001_01_0073 **Probability:** 0.7254648

Topic 164: albert granvelle regering landvoogdes philips emma raad koning egmont edele nederland state oranje inquisitie vriend plakaat rene plan geld viglius consistorie ligue hart overheid ketterij kardinaal vragen protestant bergen compromis
Proportion: 0.00066403236
Document: *Bibliographisch album.*
Document id: _gid001188001_01_0072 **Probability:** 0.9991777

Topic 140: oreste warner vader onderteekening antoinette moeder candidaat joan betty denise aeschylus lijst fen stem stemming vragen aegisthus zaak stovenzetster partij treden misschien zoon vrouw

van_der_laan evenredig naam verharding bloed ida
Proportion: 0.00065061223
Document: *De drie Electra's.*
Document id: _gid001189901_01_0064 **Probability:** 0.72860223

Topic 270: tristan gedicht leopold iseut vers dichter leopolds liefde koning hart bedier poezie cheops naam thomas eenzelvigheid innerlijk gedachte vorm diep universiteit koningin student vrouw karakter schrijven individualisme gewaarwording bron vragen
Proportion: 0.00064908434
Document: *Het universiteitsfeest van Edinburg.*
Document id: _gid001188401_01_0042 **Probability:** 0.7541299

Topic 114: brand gaspard ajax schrijfster herodotos marathon neef broeder bram atheners sophocles slag gevoelen hermine dood naam athene moeder vader huis grond perzen stuk elise agne roepen henri mensch frances kind
Proportion: 0.0006476491
Document: *Blikken in de werkelijkheid.*
Document id: _gid001185301_01_0032 **Probability:** 0.7253768

Topic 57: zwerver jeanne koning romance vrouw vader gedicht mirzaschaffy rubens serafijn rijk prinses lize jong dichter zoon dochter rodrigo heer graaf maleine ich andrea eeuw stuk vragen vroeg bodenstedt parijs mieke
Proportion: 0.0006446145
Document: *Kastieljaansche Letterkunde in de Middeleeuwen.*
Document id: _gid001184801_01_0028 **Probability:** 0.9856996

Topic 275: e italiaansch florence kaart ch naam del no non eeuw celebes eiland che italie vita si dl punt dante con atlas schrijven l cit lo dood mi io blaeu mio
Proportion: 0.00064241944
Document: *De beteekenis van den naam Celebes.*
Document id: _gid001192101_01_0123 **Probability:** 0.9894924

Topic 94: macaulay rank schrijven lady mary brief koning lezen vrouw schweitzer mensch geschiedenis raphael jakob pope kennis verhaal vermelden engeland persoon dichter ontvangen vorst bach gedicht willem karel taal belangrijk chios
Proportion: 0.00062873284
Document: *Firdausi.*
Document id: _gid001189601_01_0116 **Probability:** 0.68161184

Topic 197: eeuw stroomen geschiedenis archief aspasia vloed wetenschap attalus noordzee heer waterbeweging tijdperk kust vriesland zeespiegel hoek_van_holland boek bekend gebied schrijven perikles de_helder daling bibliotheek oever strand lyon rome hoogwater laagwater
Proportion: 0.00062713
Document: *Onze westelijke nabuur, de Noordzee.*
Document id: _gid001188601_01_0060 **Probability:** 0.56966376

Topic 185: sirius schip odd enna vader ida thijm moeder engelschen analogie snuffelaar vragen strachey hollandsche engelsch kapteyn schouten oover bergen evenredigheid oorlogschip vertellen wal haven boord hollanders leeven zee zoon weeten
Proportion: 0.0006243149
Document: *Lotgevallen van een Hollandsch retourschip in 1665.*
Document id: _gid001189601_01_0052 **Probability:** 0.775187

Topic 84: jezuiet bourgondisch banks bourgondie reserve orde naam wilhelmus hertog federal philips jezuiten ellendorf jezuit karel eeuw frankrijk michelet quinet damloup procent fransch gelden geschiedenis deposito isabeau member bourgondiers van_de_spiegel zaak
Proportion: 0.00061890093
Document: *Het Federal Reserve System in de Vereenigde Staten.*
Document id: _gid001192401_01_0004 **Probability:** 0.6176487

Topic 279: eric spel kaart schaakspel spelen naam kaartspel rousseau schrijven vries ernest engeland volk speler cyriel boek bredero eeuw fransch koning oscar mr_pickwick blijspel gladstone kleur dries jeanjacques hollanders frankrijk vroeg
Proportion: 0.0006153438
Document: *Speelkaarten.*
Document id: _gid001189501_01_0123 **Probability:** 0.9995952

Topic 1: vdh dn h naam publiek schrijven statistiek biographie g d heer_h misschien gijsbert verband karel eeuw newton koning vriend leibniz obermann lezen karel_ii willem_ii boek brun cromwell les waarde brunetiere
Proportion: 0.0006103256
Document: *Biographie.*
Document id: _gid001185901_01_0032 **Probability:** 0.8087523

Topic 100: wells schrijven brief prof kerkvergadering heer_e debussy fransch schrijver gorter boek spanje eeuw rome bron gregorius spaansch geschiedenis joden bewijzen strijkquartet sturm lulofs misschien licht vroeg coornhert naam vorm bisschop
Proportion: 0.0006071047
Document: *Prof. Lulofs en de Gids.*
Document id: _gid001183801_01_0158 **Probability:** 0.8850142

Topic 132: daniel schrijven getal grotius burman heer witt de_groot ich geer roomsch neet wiskunde zin tafel heeroom babylonisch saxe la_court d gebruiken brief professor cijferschrift nr breuk court eeuw interest uitdrukking
Proportion: 0.0006037476
Document: *Prae-Helleensche wiskunde*
Document id: _gid001193501_01_0073 **Probability:** 0.89565873

Topic 113: beweging snelheid newton lichaam mechanica theorie einstein ruimte proef bewegen licht relativiteitstheorie richting absoluut lorentz aarde aether gebeurtenis waarnemer verschijnsel natuurwet grootheid mathematisch eenparig vigny klassiek kneppelhout systeem natuurwetenschap uitkomst
Proportion: 0.0005946953
Document: *Leekenvragen ten opzichte van de relativiteitstheorie.*
Document id: _gid001192101_01_0035 **Probability:** 0.71549225

Topic 109: renaissance l middeleeuwen bezit SS d zaak ad eeuw bezitter actie middeleeuwsche begrip art italie humanist actio recht burckhardt petrarca heer eischer boccaccio gelden eigenaar belang si bayle proces res
Proportion: 0.0005915602
Document: *Disputatio juridica inauguralis, continens Annotationes ad Codicis civilis Belgici Libri secundi Sectiones duas priores, defend. L.H. Buse , a.d. 8 April. 1842. Lugd. Bat., apud Gebhard et socios. 52 pag. 8 o .*
Document id: _gid001184201_01_0061 **Probability:** 0.74111944

Topic 157: ardengo papa grootvader tante grootpapa edina mademoiselle renza diletta olaf jong mama zei clementina huis palermo plotseling rosalia vragen cavallaro abbatella vader voelen naam moeder bibliotheek villa_cavallaro morgen vroeg haten
Proportion: 0.00059150666
Document: *Darby Doyle's Reize naar Quebec. (Proeve uit een Iersch Tijdschrift .)*
Document id: _gid001183801_01_0102 **Probability:** 0.6811296

Topic 97: hearn nora gudrun israels helmer vrouw carausius schrijven rodenbach japan wate liefde kudrun vader bedrijf laura werken drama karakter handeling kind onderwerp chamberlain strijd ibsen boek lafcadio_hearn slot brief vriend
Proportion: 0.00059056166
Document: *Dramatisch overzicht.*
Document id: _gid001188901_01_0039 **Probability:** 0.7280307

Topic 3: kind vader leopardi taal moeder germanen newman vaderschap celte naam afstamming erkenning galliers volk romaansch caesar bewijzen artikel verschillend schrijven gallie bewijs onderhoud onderzoek dier romeinen celten duitsche erkennen schrijver
Proportion: 0.0005865992
Document: *Celten en Germanen.*
Document id: _gid001185501_01_0041 **Probability:** 0.9505976

Topic 269: socrates xenophon dionyzos hooglied schrijven wereld liefde faun herinnering boek bruid circe vrouw schrijver gesprek romaansch zoeken indus geest jong mensch god toren aran victor_hugo couperus benjamin vlakte schoon gevoel
Proportion: 0.00058469013
Document: *Floris Verster . (Leiden 9 Juni 1861-21 Januari 1927).*
Document id: _gid001192701_01_0033 **Probability:** 0.59224874

Topic 68: friesche fictie vrijheid lotze art beweging mensch denkbeeldig schoonheid gelden buiten jaspers friesland bloot waarde veeartsenijkunde ist jongfriesche aesthetisch das friesch makelaar vgg oder mogelijkheid naam doel luxemburg bepalen dem
Proportion: 0.00057606987
Document: *Het vrijheidsbeginsel .*
Document id: _gid001192201_01_0040 **Probability:** 0.49840873

Topic 39: arbeid bastiat lothair tenir wet bucher staat schrijver mensch stelsel verschillend kracht vorm theorie beginsel rhythmisch dienst vraag burgemeester tient denkbeeld boek zaak raad welstand beschouwing zoeken gezang hogendorp behoefte
Proportion: 0.00055725157
Document: *Poëzie en arbeid.*
Document id: _gid001190201_01_0081 **Probability:** 0.7677981

Topic 291: faust goethe mephistopheles gretchen tijger judas van_den_faust tooneelen gedeelte pilatus drama christus optreden maria_magdalena goethe_faust licht geest dichter koning tafereel treden stirling escuriaal voeren fragment dier nacht werpen voet dood
Proportion: 0.0005533211
Document: *Mozaïk.*
Document id: _gid001185401_01_0005 **Probability:** 0.69335794

Topic 225: felix kind chrisje mijnheer brakel lief vreemd bosch roepen majoor parijs moeder hart publiek reggen keller mevrouw koning berkel pop start lezer lijk stuk fransch jong oldenbarnevelts horen jager dichter
Proportion: 0.00055146666
Document: *Het vreemde kind. (naar Hoffmann).*
Document id: _gid001184101_01_0079 **Probability:** 0.7067598

Topic 148: genestet joost peter jeanne jong dominee mis priester vriend kind lief roomsch kerk reinette montalembert etioles bidden vrouw roomschen gebed hart tencin beteekenis lezen naam madame wereld koning mensch van_dale
Proportion: 0.0005406705
Document: *Roomsche woorden.*
Document id: _gid001190101_01_0073 **Probability:** 0.8477583

Topic 187: pirenne schrijven belle forster heer geneeskundigen brief vrouw marina penn zaak boris mainz fransch zoon heinrich arts constant belgen vorst polen acht verklaren zin vroeg dood kind staring boek betrekking
Proportion: 0.00053390954
Document: *Georg Forster.*
Document id: _gid001186301_01_0061 **Probability:** 0.815991

Topic 218: weber kind hart lief vragen drinken slapen huis graf vrouw zon moeder maan paard zingen meisje veld stralen vertellen lied aarde bloed zuster dragen spontini ster water blij bloem rood

Proportion: 0.0005296364
Document: *Roemeensche balladen. Naar de fransche prozavertaling van Hélène Vacaresco .*
Document id: _gid001190501_01_0062 **Probability:** 0.8647137

Topic 166: dante beatrice poezie kate dichter ten van_dante periode divina_commedia productie paradijs commedia verbeelding gevoel voorspoed vraag dat_dante vita_nuova depressie verklaring hel luthersche gevoelen kunst zaak verbruiksgoederen liefde verklaren tell geassocieerden
Proportion: 0.0005252734
Document: *Het conjunctuurprobleem*
Document id: _gid001192801_01_0012 **Probability:** 0.6744735

Topic 283: berlioz plantijn boek bijbel drukken tekst juliette schrijven gulden exemplaar arias jaures werken reyer ontvangen bibliotheek koning antwerpen parijs brief vertaling papier maart volksbibliotheek zenden frankrijk juni rome elektra goedkeuring
Proportion: 0.000520547
Document: *Plantijns Koninklijke Bijbel. Geschiedenis van een boek in de XVI e eeuw.*
Document id: _gid001188001_01_0049 **Probability:** 0.86801434

Topic 163: si hi maria ic vrouwe moeder klooster sy vrouw mi heur haer van_looy daer ghi lief gods horen kinen soe arm legende ridder mit hadde zebedeus kind koning gode mer
Proportion: 0.00052010984
Document: *De gravinne van Salisbury . 1342.*
Document id: _gid001188401_01_0046 **Probability:** 0.81158584

Topic 159: euripides treurspel phedra theseus stuk racine koor handeling sachs voedster artemis persoon dichter dood drama liefde dramatisch toon zingen aphrodite sophocles grieksch verlaten noodlot beckmesser kunst zoeken hartstgen lied oenone
Proportion: 0.00050898263
Document: *Dante en Petrarca tot Maria .*
Document id: _gid001192101_01_0098 **Probability:** 0.73705494

Topic 8: gambetta rama licht han fox lord san engeland wilson sita bismarck strijd invloed koning si geest beweging natie parlement pitt electriciteit kohl molekul mandarijn zaak verklaren magnetisch tonen kiezen besluit
Proportion: 0.000508724
Document: *Het zeeman-verschijnsel.*
Document id: _gid001190301_01_0026 **Probability:** 0.5557184

Topic 200: shelley kalff eeuw mystiek 'n 14de geest eeuws 13de licht god boek ziel godwin mie maerlant wereld kennis poezie mary verwording 15de nait schoonheid iii_blz liefde prof hart schrijven invloed
Proportion: 0.000506759
Document: *Semietische mystiek.*
Document id: _gid001191901_01_0017 **Probability:** 0.83985275

Topic 73: beaumarchais vo marianne lomenie schrijven gevoel boek vader vriend bewijs schrijver zaak hy dominee voltaire brief uitgave vrouw selma rec gids stuk vlaamsch zyne heer barbier zoon bewijzen francs van_der_hoeven
Proportion: 0.0005060964
Document: *I. Jaek, of een arm Huisgezin, door P.F. van Kerckhoven . Te Antwerpen, bij Karel Oberts. 1842. 221 bl. 8 o . II. De koopmansklerk. Eene Antwerpsche Zedenschets, door P.F. van Kerckhoven . Antwerpen, Drukkerij van J.E. Buschmann. 1843. Teekeningen door Eug. de Block, 76 bl. 8 o . III. Hoe men schilder wordt. Eene ware geschiedenis van eenen schilder, die nog leeft, door Hendrik Conscience . Antwerpen, Drukkerj van J.E. Buschman. 1843. Met Houtsneden. 71 bl. groot 8 o . IV. Wat eene moeder lyden kan. Ware Geschiedenis, door Hendrik Conscience . Vijftig Houtsneden, geteekend door J. Mathysen, op hout gesneden door H. Brown. Antwerpen, Drukkery van J.E. Buschman, Uitgever. 1844. 66 bl. 12 o .*
Document id: _gid001184401_01_0016 **Probability:** 0.88490254

Topic 253: eed jeugd doopsgezinde reeks e functien heer gids hel jeunesse eden doopsgezinden gedurig breuk van_vollenhoven conventie functie koning jacobijnen mei formulier afleggen geschiedenis eedt zweren wet waarde convergentie jacobijn term
Proportion: 0.0004992906
Document: *Eenige zeetermen in de Nederlandsche vertaling van Shakespeare's Tempest*
Document id: _gid001192801_01_0085 **Probability:** 0.6338658

Topic 286: savonarola roman shakspere warhold stad sidney van_oordt koning historisch eeuw geschiedenis pl greene lorenzo irmenlo oordt florence italie profeet rijk ta vorst tolpoort lyly scudery engelsch handeling middeleeuwen historie vertaling
Proportion: 0.0004979484
Document: *Kastieljaansche Letterkunde in de Middeleeuwen.*
Document id: _gid001184801_01_0010 **Probability:** 0.7613516

Topic 202: europesche kolonisatie tol denemarken kolonist klimaat europeanen sterfte commissie kolonie journalistiek verdrag holland dr vestigen pandora rotterdam tropisch luchtstreek haiti journalist vermeulen werken heete bezwaar regering bombay zaak maatschappij zond
Proportion: 0.0004947339
Document: *Europesche Volkplantingen in tropische gewesten.*
Document id: _gid001187001_01_0013 **Probability:** 0.7985097

Topic 56: fechner menno wereld eeuw wijsbegeerte gebouw vorm systeem geest indische vos stijl denkbeeld studie gebied constructie god wetenschap kinakultuur graniet bouwen boek schepping mensch grondslag vroeg geschiedenis kennis kant evert
Proportion: 0.0004868809
Document: *Indische theosofie.*
Document id: _gid001191101_01_0049 **Probability:** 0.8720555

Topic 149: vakonderwijs oberlin gild patroon duitsche middenstand gezel ambachtsnijverheid nijverheid handwerk leiden grootindustrie bosscha organisatie nederlandsche duitschland tellen bedrijf ambacht prof kleinnijverheid gildewezen handwerker bijv leerlingwezen elize zoodoende belang vak baas
Proportion: 0.00047787317
Document: *Middenstandskernen.*
Document id: _gid001190601_01_0030 **Probability:** 0.63405114

Topic 162: jacobi volkenrecht recht pp hodgskin schelling thomasius l eiwit seqq vet natuurrecht wet mensch wetenschap gewoonte bron regel arbeid voedsel voorraad hamann oorlogsrecht bekend naam waarde stellig voorbeeld schrijver inkomen
Proportion: 0.00047630284
Document: *Een en ander over Christian Thomasius (geb. te Leipzig 1655. Professor aldaar 1684, - te Halle 1690, 1728)*
Document id: _gid001192101_01_0036 **Probability:** 0.7919203

Topic 40: maurits nieuwpoort oostende signalement saidjah voetvolk ruiterij spaansch buffel schr havelaar valentin regiment duin kornet hoofd zijde schans bandiet ernst spanjaarden leger faustus adinda lengte stella verschillend no bertillon albertus
Proportion: 0.00047122748
Document: *De slag van Nieuwpoort.*
Document id: _gid001186901_01_0058 **Probability:** 0.459978

Topic 65: werkloosheid mensch duitschland arbeider mark geldeenheid dood stahl inflatie lichaam arbeid fontane pierre verhouding sterven jong orgaan middel werkgelegenheid ouderdom bethmann element gevolg ontwikkeling marken leeftijd arnim werken geestelijk zin
Proportion: 0.00046603088
Document: *Over oud-worden.*
Document id: _gid001191701_01_0019 **Probability:** 0.6059097

Topic 212: jezus luk vw paaschfeest johannes maria kunst schrijver eeuw galilea matth moeder jeruzalem byzantijnsch typhus kiemplasma feest claes mark heer rafael markus lukas giotto traditie no vi school vraag trekken
Proportion: 0.00046277622
Document: *Beoordeelingen en aankondigingen.*
Document id: _gid001184701_01_0033 **Probability:** 0.99905473

Topic 251: militair coehoorn pest rat mensch boek schrijven van_der_hoeven vlo obdam delict krijgsraad huis schrijver fransch woningverbetering pestbestrijding marlborough eekeren saargebied naam soldaat vauban kommissie ziel dr jong eeuw roman weiss
Proportion: 0.0004545634
Document: *Overzicht der Nederlandsche letteren.*
Document id: _gid001190701_01_0041 **Probability:** 0.85218537

Topic 168: velsen sage lied floris rekenkamer nederlandsche geraert overlevering koning deensch uitgaaf graaf vorst bron motief vrouw str floris_v dood bombay heer historisch hi naam bericht gedicht jong dienstjaar invloed ontstaan
Proportion: 0.0004366623
Document: *Het lied van Geraert van Velsen.*
Document id: _gid001189901_01_0051 **Probability:** 0.9805112

Topic 124: witt grey engeland churchill ic aw brief w downing ick asquith wilson educatie haldane schrijven kant engelsch willem_ii morley politiek japikse beleving prinses_royaal kabinet augustus holland oranjepartij daer regel engelschen
Proportion: 0.00042942804
Document: *Aanstaande verandering van Amsterdam's wapen.*
Document id: _gid001184101_01_0104 **Probability:** 0.7675798

Topic 221: timor koepang portugezen portugesche nederland tijger belgie radja nederlandsche sonebait aap koper portugal compagnie nederlandsch timorezen eiland olifant gebied bloed heer_brouwer volk belgisch portugeesch reus solor aanzien leggen rotti vlag
Proportion: 0.0004238975
Document: *De onjuistheid der grondstellingen van de memorie van antwoord nopens het Nederlandsch-Belgische tractaat.*
Document id: _gid001192601_01_0104 **Probability:** 0.6857521

Topic 244: rosny brug americaansch van_leeuwen russell ameland nit taal do tongval heer_adams bouwer verhaal pastoor schrijven priester volk meredith uitrusting mensch rivier beschaving brief aegidi d eiland beschaafd prinses volkenregt januari
Proportion: 0.00042276524
Document: *Nieuwe Duitsche dichters.*
Document id: _gid001187601_01_0037 **Probability:** 0.5985069

Topic 234: mencius steunpunt koning hiram hubrecht bloed sidon middel zaak mentor dier kracht bacil uitkomst mont loe dollar gevolg rijk verschillend inspuiting regeering jong heer_stieltjes staat werken gevaar vorst heer_s onvatbaar
Proportion: 0.00040982364
Document: *China's volkstribuun*
Document id: _gid001193101_01_0049 **Probability:** 0.73111576

Topic 268: haring bosse stendhal bergman gilbert heer schrijver latijnsch lezen beyle rector dostojevski levensschets school model schrijven huis grenoble ton eugene leerling zaak honderd grieksch gothisch noot conrector bladz notaris frank
Proportion: 0.00040664096
Document: *Antwoord op de Antikritiek van de in De Gids , 1841, N o . 9, voorkomende beoordeeling van: Levensschets van Frans Antoni Bosse, enz., geplaatst in De Recensent ook der Recensenten , voor 1842, N o . 2, blz. 49-66.*
Document id: _gid001184201_01_0103 **Probability:** 0.7894705

Topic 76: emmy schaepman kind giulio eline ooog diep voelen volker rudolf ding orfeus regel rene_de_clercq stem jan_walch vreemd rust hoofd donker horen strophe licht gedacht end pijn verleden ru fel werkelijkheid
Proportion: 0.00038685076
Document: *Weifeling.*
Document id: _gid001191801_01_0123 **Probability:** 0.9958751

Topic 38: hamann pierre_leroux moslim sovjet leroux rijk restitutie ha faustisch schrijver eeuw dier enver belangrijk eigendom boek toerkestan geschiedenis millioen tijdperk geslacht vorm heer recht buddhisme jong vader mensch artikel vroeg
Proportion: 0.00038571513
Document: *In memoriam Vlaamsche arbeid*
Document id: _gid001193101_01_0093 **Probability:** 0.8948642

Topic 37: metz genoveva effen bazaine koning mac_mahon zweden sedan opera zweedsche venizelos siegfried college fransch augustus chalons college_de_france spelen parijs strijker amsterdam schouwburg plan les legerkorps frankrijk recht duitschers harald seminarium
Proportion: 0.00038073384
Document: *Genoveva op den hoogen semmer.*
Document id: _gid001192001_01_0044 **Probability:** 0.5399145

Topic 194: werken vgl focquenbroch perzie amsterdam boek licht stuk eeuw vers scarron bekend verschijnen vertaling uitgave fransch vroeg schrijven vdh roodkapje druk kind blijspel gans gedicht drukken les god naam uitgeven
Proportion: 0.0003721976
Document: *Focquenbroch.*
Document id: _gid001188101_01_0067 **Probability:** 0.948785

Topic 299: dollar billiton zaak vrouw geld kind papier huis mensch honderd tinerts brief boek wallstreet tin diep vragen humperdinck banka onderneming dikwijls naam verzekeren eerlijk toestand woning trekken rente vreemd pikol
Proportion: 0.0003685038
Document: *Een handelsroman .*
Document id: _gid001186301_01_0009 **Probability:** 0.999629

Topic 0: suiker oedipus geschiedenis historisch kern anton sophocles raffinaderij stuk nederlandsche baron jacobus rede schrijven schrijver ruw wetenschap duitsche engelsch zaak heer broeder zondeloosheid kennis jezus_christus kilogram vertaling godgeleerdheid godsdienstwetenschap publiek
Proportion: 0.0003484159
Document: *Bibliographie.*
Document id: _gid001191101_01_0034 **Probability:** 0.99406743

Topic 85: renan lear edmund ziekte gloster cordelia edgar vee dier goneril regan shakspeare runderpest idee liefde dochter eros albany levensdrang dood besmetting mensch besmetten guido monna_vanna koning gevolg afmaken nar waanzin
Proportion: 0.00034542382
Document: *De runderpest.*
Document id: _gid001186601_01_0020 **Probability:** 0.654325

Topic 31: blank coen handel compagnie ras kleurling schip g banda blanke batavia indie e oorlogschip indianen sul additioneel kleuren particulier heer reael middel slaaf amboina can varen sijn ambon engelschen luyden
Proportion: 0.0003433745
Document: *Gekleurd en blank*
Document id: _gid001193401_01_0017 **Probability:** 0.5824615

Topic 123: schneevogen kanon schneevoogt fransch cm m gast uur david krupp kg buys heer fabriek hotel essen naam generaal piccini gebied kennis jong verschillend augustus geest kunst buitengasthuis firma gelegenheid trekken
Proportion: 0.00032773113
Document: *Het internationale feest bij den Kanonnenkoning.*
Document id: _gid001187901_01_0065 **Probability:** 0.7862085

Topic 49: bellamy 'n gedicht van_alphen schrijver mineraal delfstof schrijven brief kunst pascal dichter alphen stad derkinderen hoofdstuk vriend mineralogie wandschildering utrecht heer e smaak natuur fillis oorspronkelijk naam palermo beteekenis branden
Proportion: 0.00032149677
Document: *Stemmen uit de redactie*
Document id: _gid001193201_01_0089 **Probability:** 0.96708333

Topic 25: gemeente h montherlant goldberg rijk barres auteur cod ouida boer gemeentelijk religieus din van_de_berg ideaal schrijven brichanteau sterven arm puck opcenten groux jong cat ah religie vergoeding hoofdstuk bierens_de_haan twente
Proportion: 0.0003198944
Document: *Rijk en gemeenten.*
Document id: _gid001192901_01_0066 **Probability:** 0.6226172

Topic 78: thomas tekst imitatio kempis handschrift cap boek heer bavinck prof geert_groote thomas_van_kempen kristendom auteur stuk latijn variant latijnsch nederlandsche gerson lezen brugman mystiek schrijver devotie drukken vertalen grondtekst slot navolging
Proportion: 0.0003163827
Document: *Het beste dat wij aan de wereld gaven.*
Document id: _gid001192901_01_0022 **Probability:** 0.9993035

Topic 260: mary charles disraeli eliza vrouw graziella schrijven vrouwenbeweging mensch wereld schlegel liefde conservatief hart engelsch boek vlieger lucht toestel tante brief uiting toekomst gevoel naam karakter amerika jong vader licht
Proportion: 0.00030741197
Document: *Duitsche physica*
Document id: _gid001193601_01_0120 **Probability:** 0.7533494

Topic 231: minister opperbevelhebber beleid materie oorlog verantwoordelijkheid foix ministerraad heer ding beweging graaf staatsrechtelijk opperbevel krijgsbeleid verantwoordelijk gaston aristoteles vorm regeering functie verhouding zaak marine causa dragen buitenlandsche sfeer positie heemstede
Proportion: 0.00030531792
Document: *Aristoteles' Metaphysica*
Document id: _gid001193601_01_0113 **Probability:** 0.748406

Topic 176: oom vincent kareltje jonkvrouw mauve wetenschappelijk wetenschap kritiek kunst mensch theo waals mama vincent_van_gogh trekken kamer huis negen admiraal zwaard ooog misschien kleinduimpje boek heeren ding slaan roepen volkenkunde werken
Proportion: 0.00029319257
Document: *Buitenlandsch overzicht.*
Document id: _gid001189501_01_0066 **Probability:** 0.9530076

Topic 28: duitschland engeland boekerij aant duitsche d blair amerika vg openbaar m hanne engelsch heer_lulofs berlijn montesquieu asquith alphabet henschel aanteekening londen lepsius jong zaak frankrijk prijs bulow hauptmann mark sophie
Proportion: 0.0002792484
Document: *Specimen inaugurale de montesquivio - quod - publico ac solemni Examini submittit J. Heemskerk Ab. f. Amstelodami. 1839. apud J.H. et G. van Heteren. 8 o . Pars I. pag. XVI. et 120. - Pars II. pag. 194.*
Document id: _gid001184001_01_0038 **Probability:** 0.5733137

Topic 156: franciscus clara orde assisi michelangelo elias heerendienst dienst blok paus canova sabatier model vlinder marmer geld carrara regel antieken werken voorbeeld van_velzen begrijpen armoede minderbroeder hoofd kapel beschouwen w gevecht
Proportion: 0.0002763362
Document: *Nog een woord over den slag bij Doggersbank.*
Document id: _gid001184501_01_0082 **Probability:** 0.86274743

Topic 280: cannes marcel_schwob andersen schrijven kind mentone mensch sarah_bernhardt mug heer coppee sterven les barres schwob ziek hotels sarah zenden parijs kroniek vader lijden uur jong vagebond herinneren hyeres karakter dikwijls
Proportion: 0.0002412234
Document: *Aanteekeningen en opmerkingen.*
Document id: _gid001190101_01_0121 **Probability:** 0.84993374

Topic 125: gide vertaling koffie zoon tekst lezing schrijven brazilie jozef chrysale maria boek rave roergebied philaminte lezen zin jezus tooneelkijker savantes eeuw andre_gide blad rijksweer tijdschrift lezer schrift licht stuk roman
Proportion: 0.00023397352
Document: *Fransche letteren*
Document id: _gid001192701_01_0087 **Probability:** 0.9987745

Topic 171: marnix richepin van_de_vluggen rabelais hageroos barres adelaert aminta cahier tableau les silvia vondel prof finland typus nederlandsche belangrijk leeuwendalers gueux ta geschrift schrijven vii leeuwendaler vi verschijnen recht van_de_vlugt geschiedenis
Proportion: 0.0002251235
Document: *Bibliographie.*
Document id: _gid001192401_01_0076 **Probability:** 0.8264302

Topic 53: luxemburg water goud suriname ariel louvois eaux rivier nous eau boschneger diogenes woerden boot kreek y goudveld neger doorsteken goudzoeker digue coupure les hangmat leck placer ennemis october paramaribo cayenne
Proportion: 0.00022360224
Document: *Het Diogenes-ideaal.*
Document id: _gid001191001_01_0167 **Probability:** 0.84288

Topic 126: overtreding belastingschuldige geuns inkomen criminaliteit klasse cijfer plegen verordening opbrengst aanslag groep anna jj bevolking badeloch aanslaan feit belastbaar pct rangschikken aanwijzing hester heffing veroordeelingen sphinx sterk kohier aangeslagenen aangifte
Proportion: 0.0002072901
Document: *Kleine criminaliteit.*
Document id: _gid001191601_01_0056 **Probability:** 0.99908745

Topic 284: swift van_hamel forster hamel whigs tories biografie biograaf esther addison whig ag gulliver schrijven fransch goethe dagboek persoonlijkheid tory staring harley misschien aanteekening engeland brief aard mensch lezen benjamin_constant tollens
Proportion: 0.00018368909
Document: *Buitenlandsch overzicht.*
Document id: _gid001189901_01_0103 **Probability:** 0.865711

Topic 195: eindexamen leeraar 5e klacht inspekteur school 6e cijfer heeren commissie werken examen direkteur 4e wantrouwen 8e paraat iblz kennis iiblz examinator kwaad gebrek meening bruggencate oordeel leerling leerstof m 9e
Proportion: 0.00015306374
Document: *Vóór het eindexamen. Onuitgesproken rede .*
Document id: _gid001191001_01_0005 **Probability:** 0.99911034

Topic 117: huerta mexico boek shakspere dichter personeel gevoel schoonheid kunst aandeelen literatuur dragen ziel peinzer querido waereld natuur meditatie exploitatie genie boomen sonnet deugd kracht mensch van_deyssel italie licht trouw strijd

Proportion: 0.0001234425

Document: *Literatuur en leven. Is. Querido, Meditaties over Literatuur en Leven. 1898.*

Document id: _gid001189901_01_0019 **Probability:** 0.9986448

E.2 Discipline proportion plots

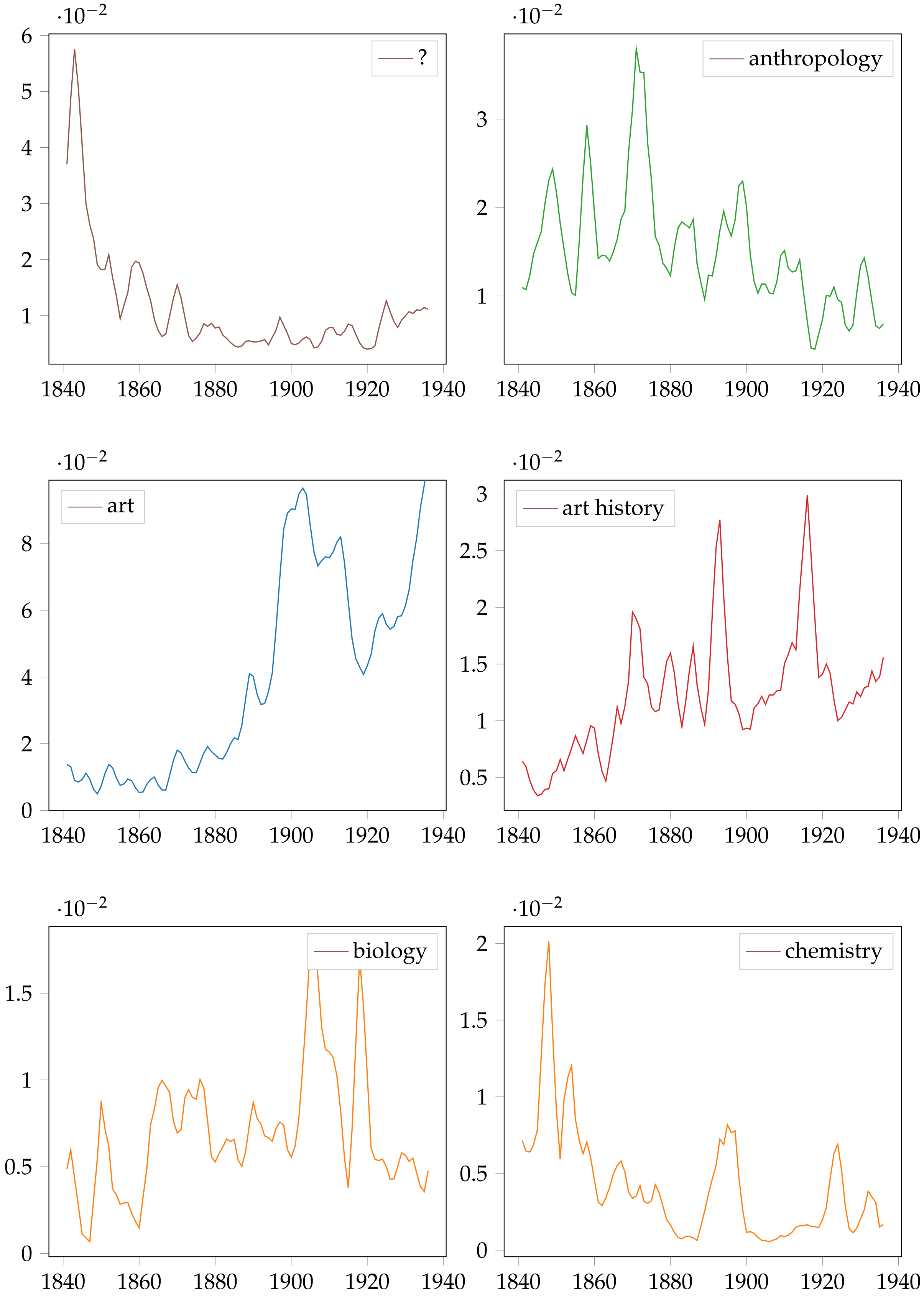


FIGURE E.1: Relative proportion of a discipline per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

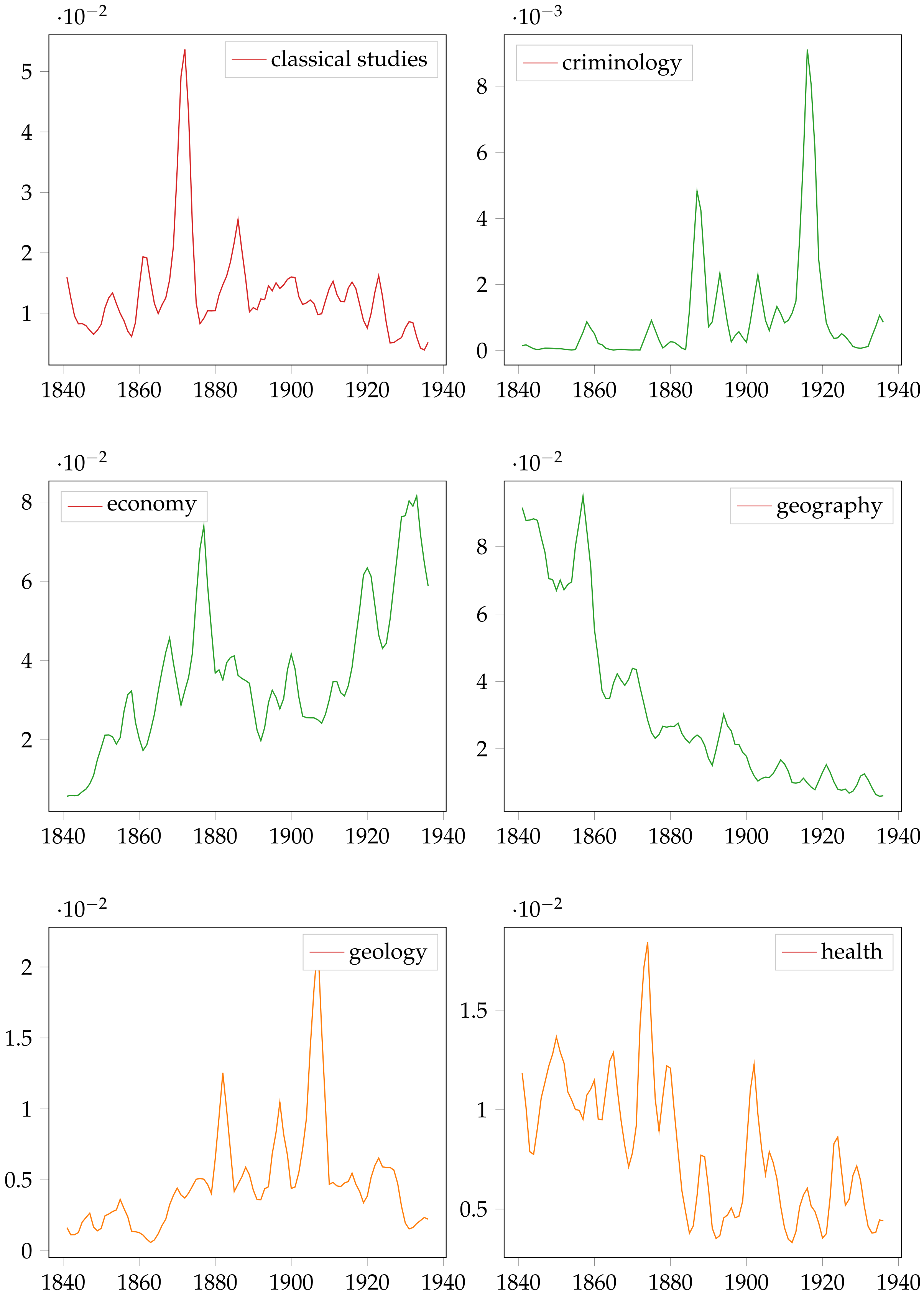


FIGURE E.2: Relative proportion of a discipline per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

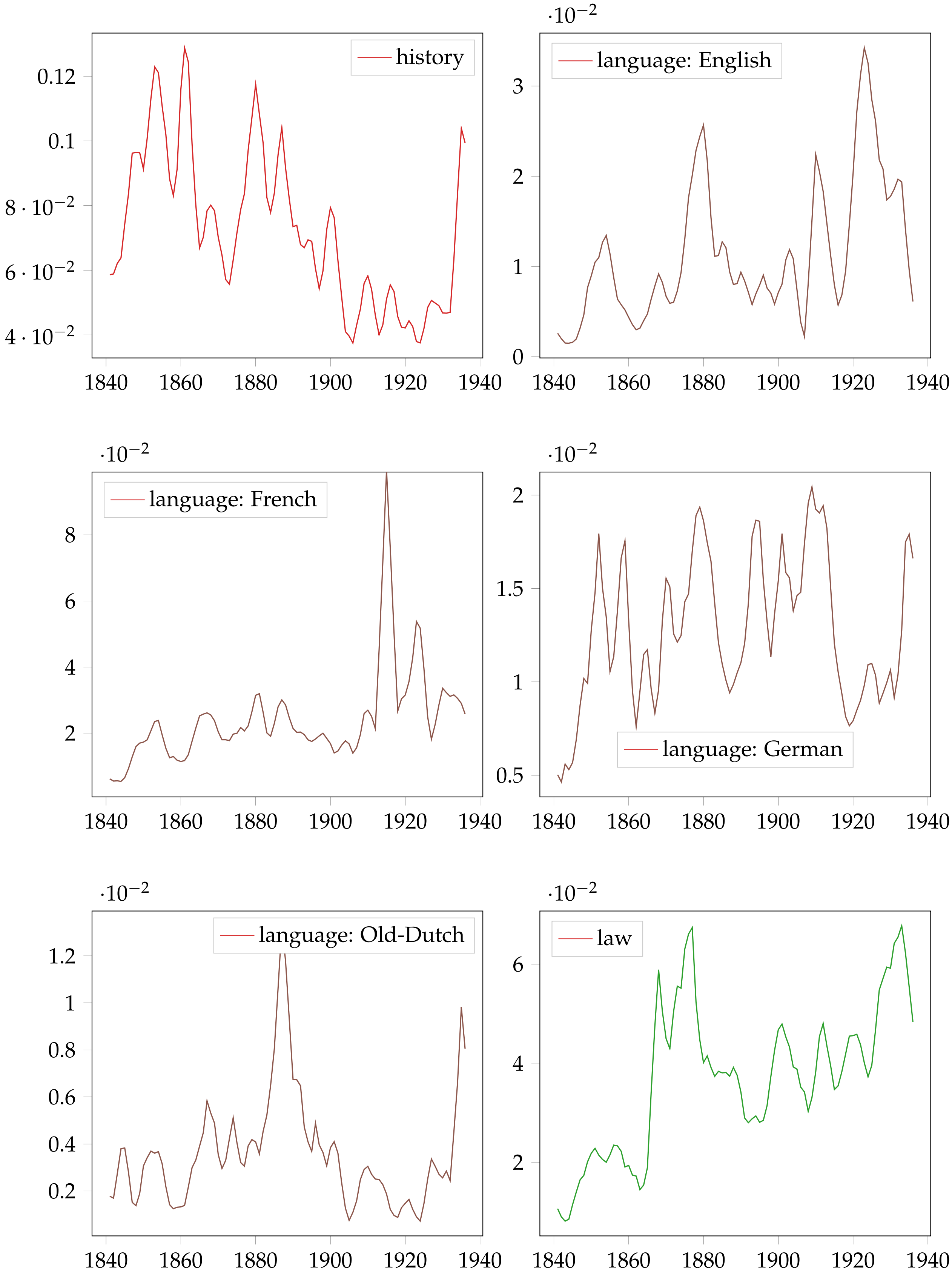


FIGURE E.3: Relative proportion of a discipline per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

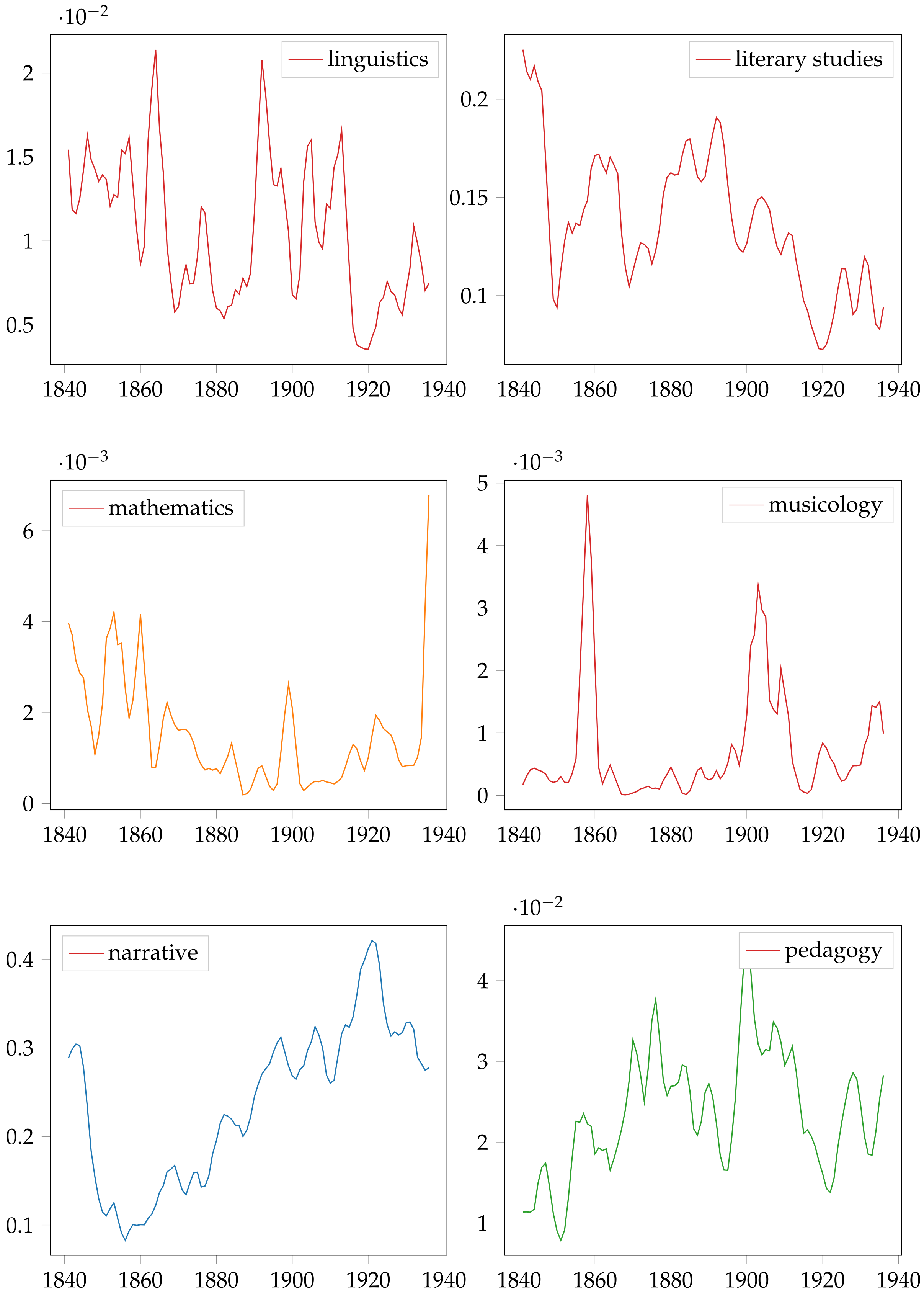


FIGURE E.4: Relative proportion of a discipline per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

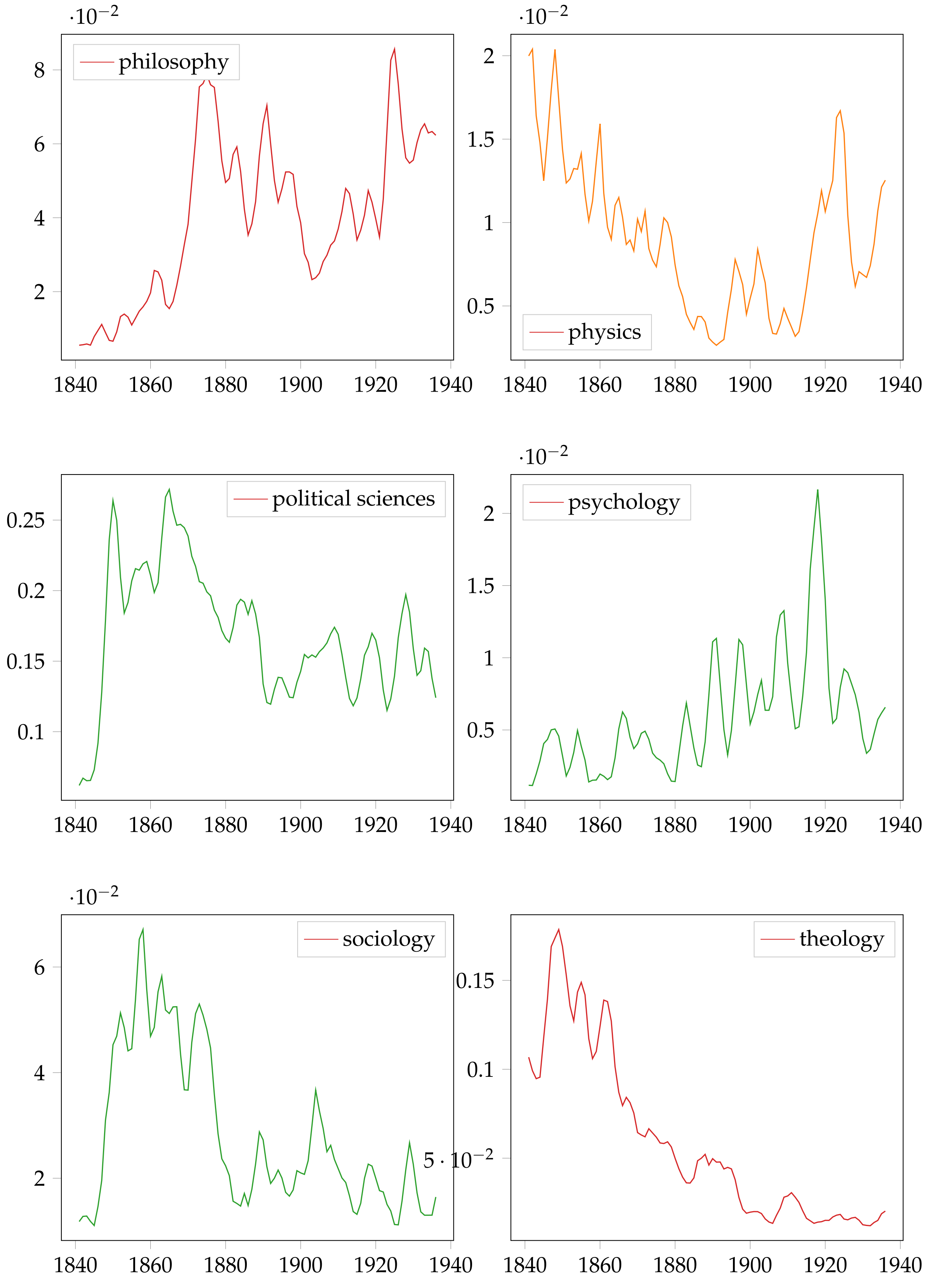
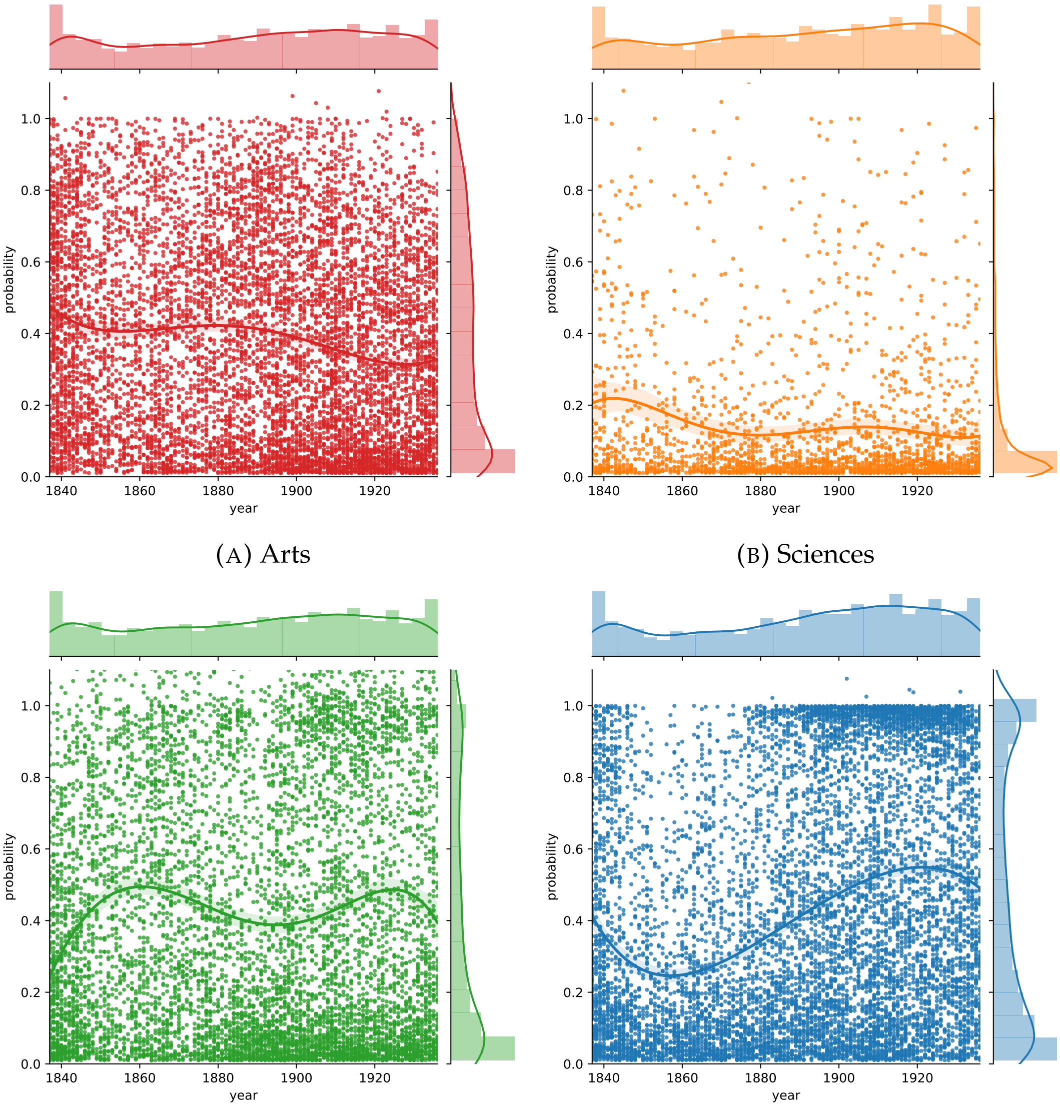


FIGURE E.5: Relative proportion of a discipline per year (volume). The graph is smoothened by taking the triangular window of 5 consecutive years.

E.3 Regression plots



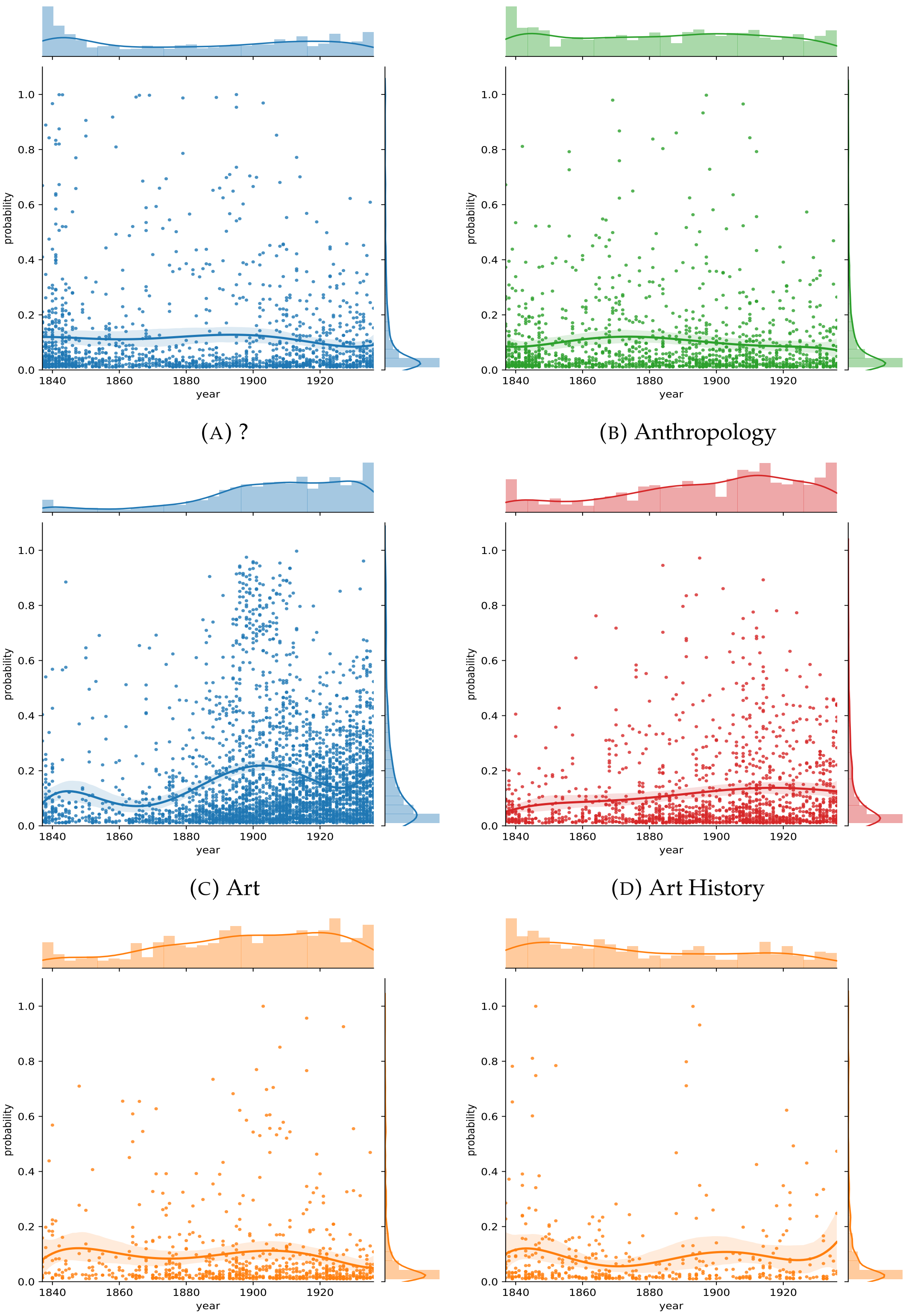
(A) Arts

(B) Sciences

(C) Social Sciences

(D) Other

FIGURE E.6: Scatter plot for each field. Every dot is a document that has a probability (y-axis) for the given field. Only documents that include the discipline (p > .01) are plotted. The margin plots are histograms and the line is a trend line that fits the points in the space.



(A) ?

(B) Anthropology

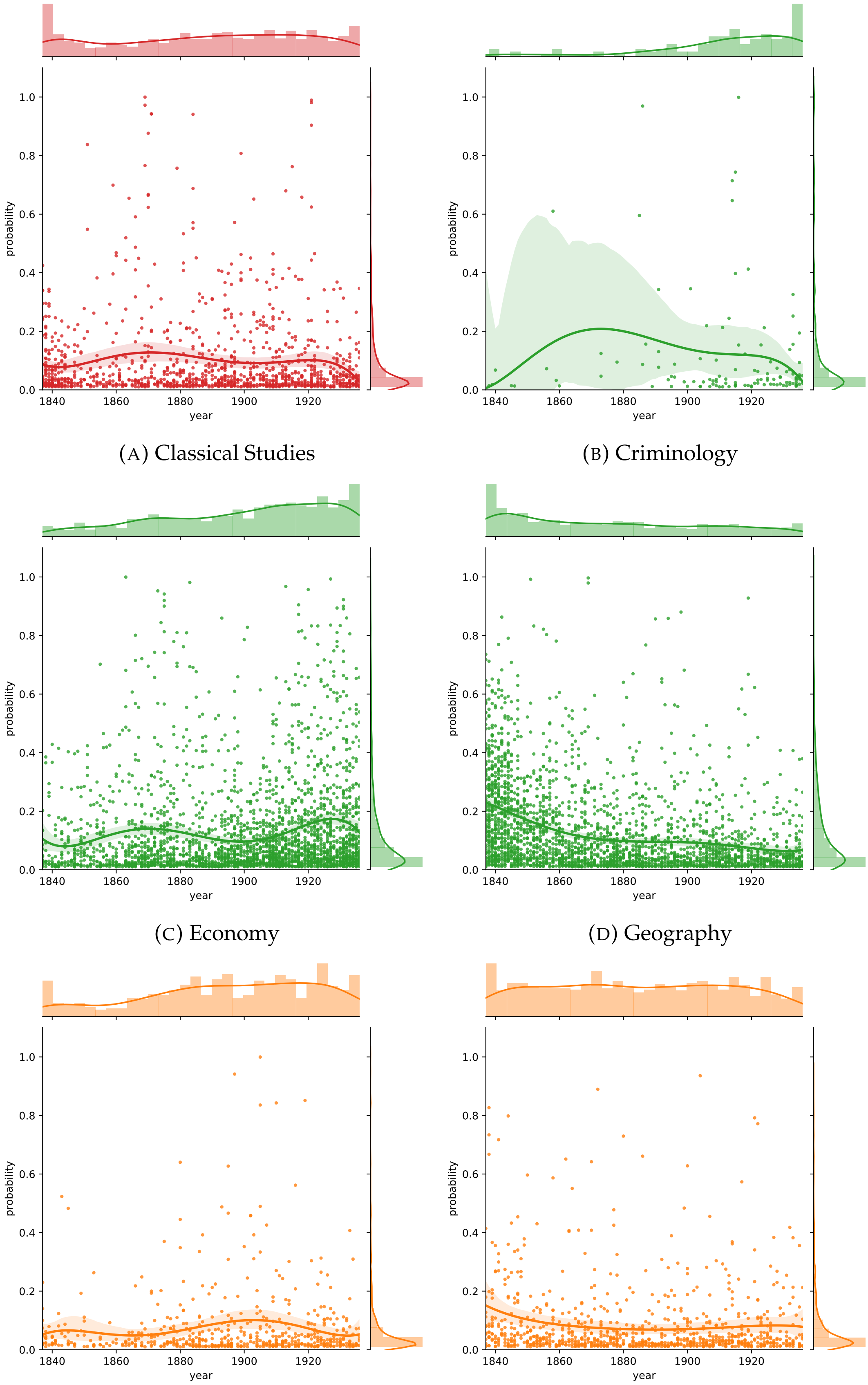
(C) Art

(D) Art History

(E) Biology

(F) Chemistry

FIGURE E.7: Scatter plot for each discipline. Every dot is a document that has a probability (y-axis) for the given discipline. Only documents that include the discipline (p > .01) are plotted. The margin plots are histograms and the line is a trend line that fits the points in the space.



(A) Classical Studies

(B) Criminology

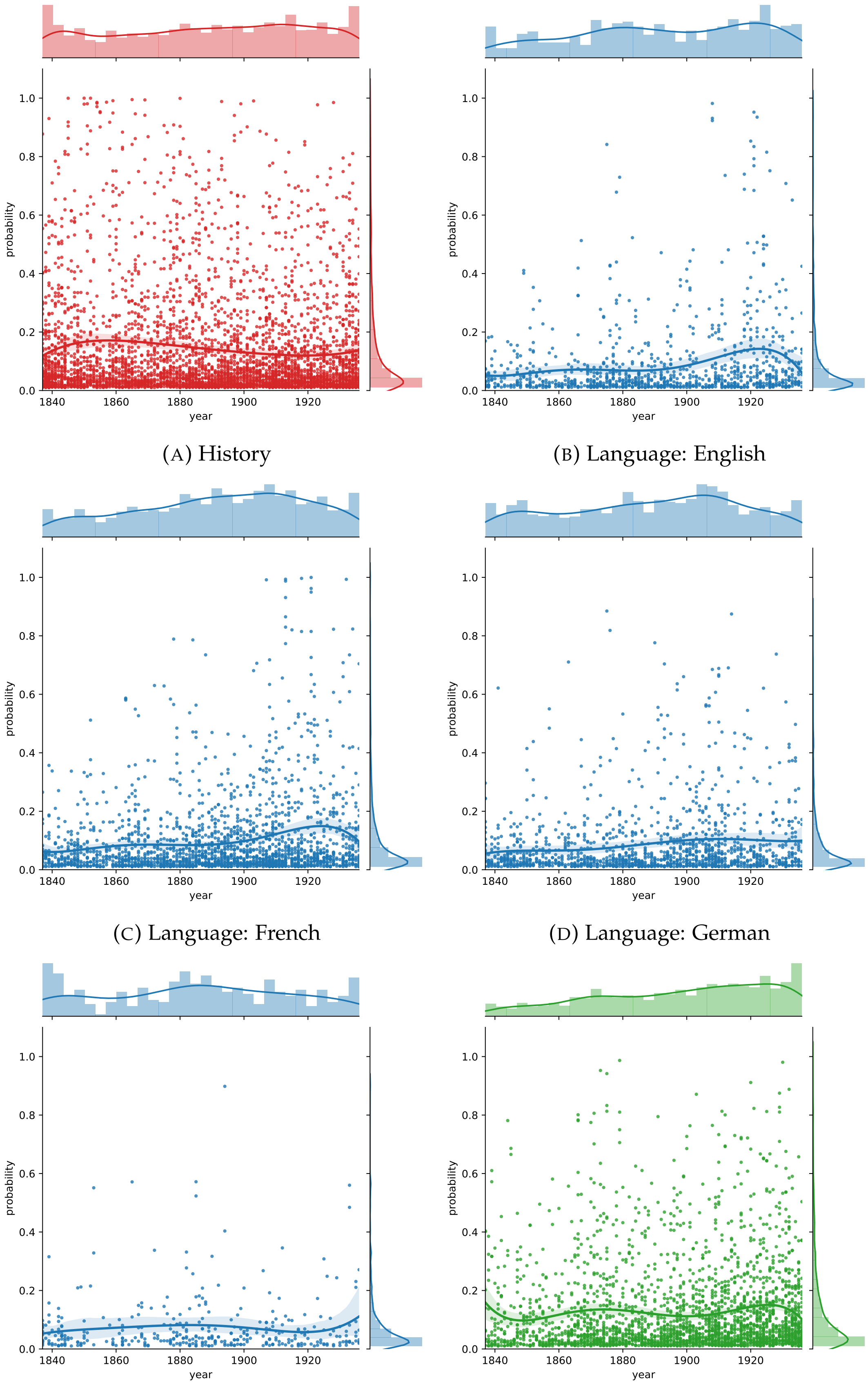
(C) Economy

(D) Geography

(E) Geology

(F) Health

FIGURE E.8: Scatter plot for each discipline. Every dot is a document that has a probability (y-axis) for the given discipline. Only documents that include the discipline (p > .01) are plotted. The margin plots are histograms and the line is a trend line that fits the points in the space.



(A) History

(B) Language: English

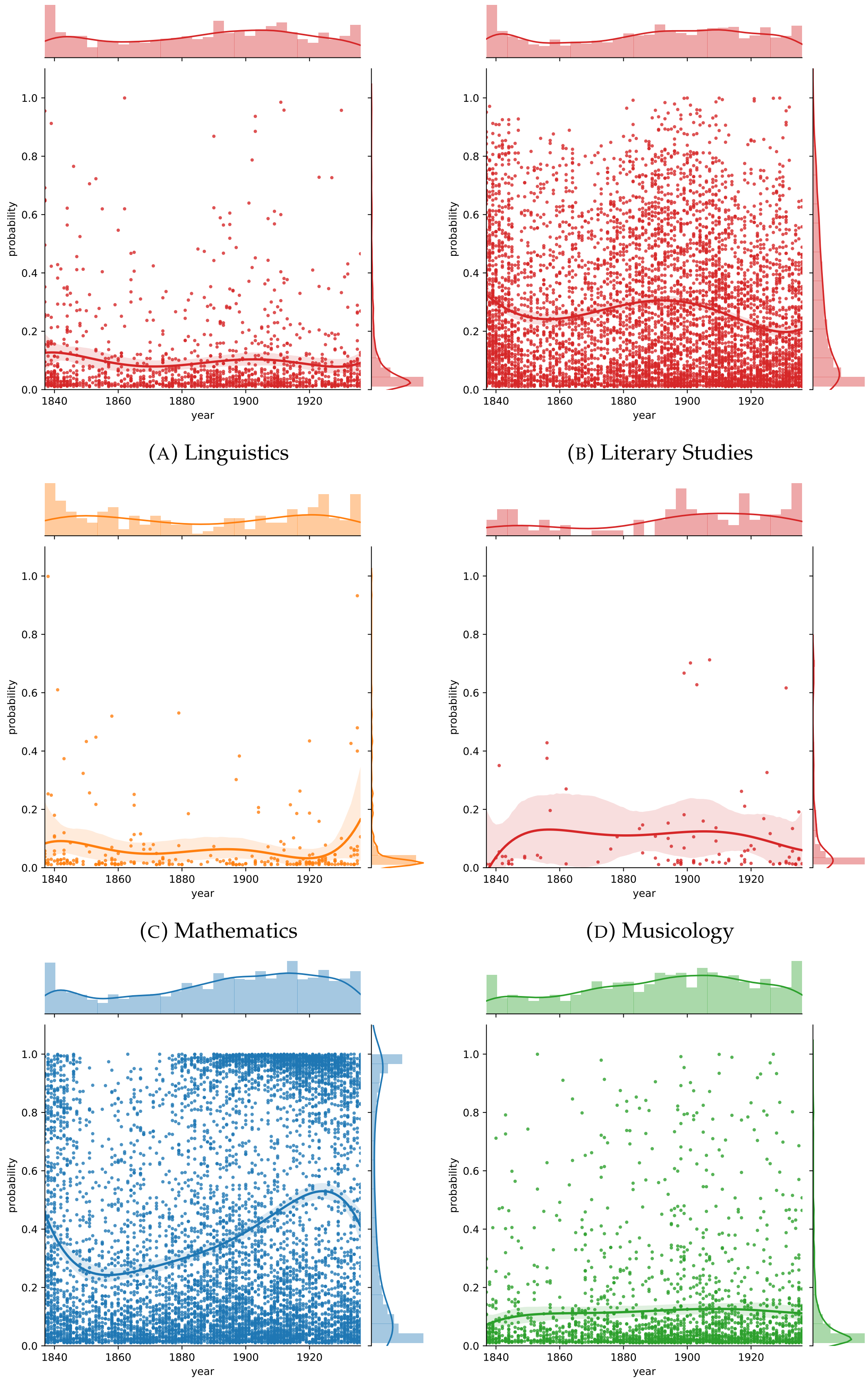
(C) Language: French

(D) Language: German

(E) Language: Old-Dutch

(F) Law

FIGURE E.9: Scatter plot for each discipline. Every dot is a document that has a probability (y-axis) for the given discipline. Only documents that include the discipline ($p > .01$) are plotted. The margin plots are histograms and the line is a trend line that fits the points in the space.



(A) Linguistics

(B) Literary Studies

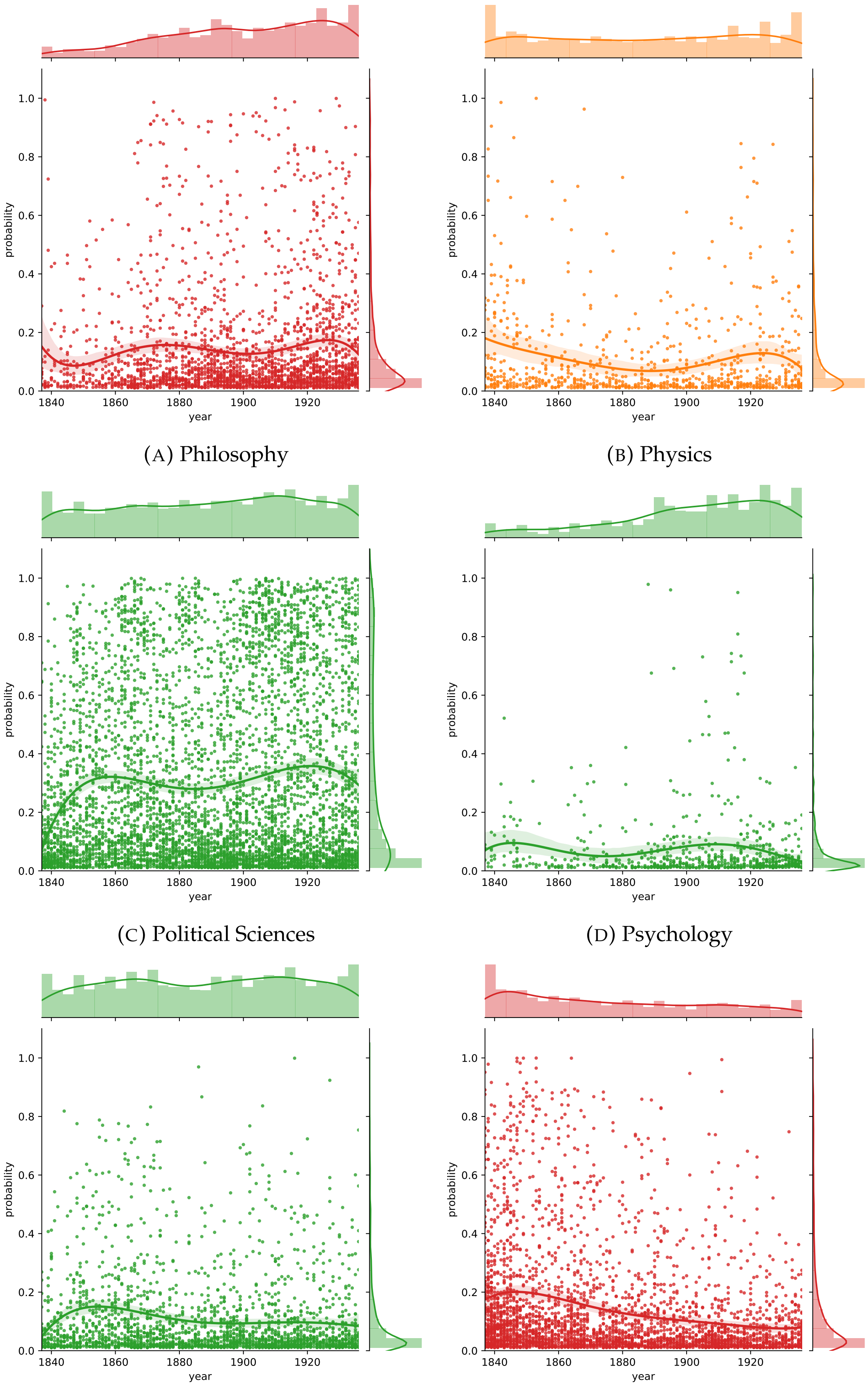
(C) Mathematics

(D) Musicology

(E) Narrative

(F) Pedagogy

FIGURE E.10: Scatter plot for each discipline. Every dot is a document that has a probability (y-axis) for the given discipline. Only documents that include the discipline ($p > .01$) are plotted. The margin plots are histograms and the line is a trend line that fits the points in the space.



(A) Philosophy

(B) Physics

(C) Political Sciences

(D) Psychology

(E) Sociology

(F) Theology

FIGURE E.11: Scatter plot for each discipline. Every dot is a document that has a probability (y-axis) for the given discipline. Only documents that include the discipline (p > .01) are plotted. The margin plots are histograms and the line is a trend line that fits the points in the space.