UPPSALA
UNIVERSITET

Tracking the Time Trends of Swedish Literature and Finding Characteristics of Authors by Using Topic Modelling

Hikaru Hashimoto

Uppsala University
Department of Linguistics and Philology
Master Programme in Language Technology
Master's Thesis in Language Technology, 30 ECTS credits

January 29, 2020

Supervisor:
Mats Dahllöf

Abstract

In this thesis, we discover the time trends in Swedish literature and characteristics of authors. We apply latent Dirichlet allocation (LDA), a method for topic modelling, to a corpus composed of 118 Swedish books and prose collected in Litteraturbanken. By using the LDA model, we observe two findings: topics that focus on daily life, such as nature or family are frequently observed in the corpus, and peaks of topics in time trends result from books on the same topic written by several authors or books written by an author in a short time. Additionally, LDA is applicable to assessments of the characteristics of authors. We list the particular topics for nine authors with more than three books in the corpus by comparing the topic distribution of those authors to the topic distribution of the entire corpus.

Contents

Preface

I would like to express my gratitude to my supervisor Mats Dahllöf for supporting me during the whole thesis. Thank you for giving me helpful advice and helping translate Swedish words into English. Without you, it would have been impossible to examine Swedish literature in my thesis. I also would like to thank Orphei Drängar, a worldly-famous men's chorus in Uppsala for accepting me as a member. When I listened to a concert held by OD in Tokyo a few years ago, I was so fascinated that I decided to go to Sweden. It is my honour to sing in OD. Finally, I would like to thank my friends and family. Having a conversation with friends in our computer room, Chomsky relaxes me even when we rush our assignments to meet deadlines. My family has always supported me mentally and financially. If I were alone, I would not have finished the thesis.

1. Introduction

1.1. Background

Reading many books (e.g. 100 books) and quickly finding the similarity or dissimilarity among them is difficult. Assuming that a person takes 2 hours to read a book, the duration necessary for her or him to read 100 books is approximately 200 hours. After reading books, she or he takes time more to analyse the texts. This type of literary analysis, reading all pages of books to discuss them, is called *close reading* (Jockers, 2013). Critics provide insights on literature through close reading and have said that authors have unique styles, that some authors affect other authors, and that a group of authors has employed movements such as romanticism or realism. For example, authors in the romanticism genre tend to express individual, subjective feelings or their religiosity. Although these types of critiques increase the understanding of works of literature or relations among them, these types of literary criticism are by nature, subjective and time-consuming.

An increasing amount of literature is available on the web in digitised, downloadable forms for free or for purchase. Thus, individuals can now read books without procuring them from a book store or a library. Researchers in computer science and other related fields such as language technology have proposed computational methods for effectively handling these types of large data.

Computational methods for handling data have advantages over literary analysis by humans. First, computers read texts fast and in a manner different from humans because computers do not understand the meaning of words. Computers count words in novels and sort them for quick retrieval and never make mistakes in reading once programmed correctly. Second, computers present objective data by sorting and do not make subjective judgements. Although subjectivity sometimes enters digital methods in the building process, the data they present can be less subjective than human reading, and facilitate the discovery of new perspectives that cannot be obtained by close reading. Computational methods for literary analysis are a subsection of digital humanities.

Digital humanities is a field of research in which researchers use digital tools such as computers in humanities. As aforementioned, computational approaches facilitate the objective reading of a lot of text quickly by providing reorganised data. For instance, search engines such as Google return documents which may relate to keywords input into the search box. In general, search engines read documents on the web and mark them in advance. Search engines reorganise documents according to users' keywords and find documents related to their interests.

Digital methods provide new perspectives on literary analysis. Researchers can now use various means to read documents. A new approach is to skim many documents faster than before by using information retrieval methods. Researchers no longer have to read each document from beginning to end to search for keywords

if documents are correctly digitised. Another perspective researchers acquire through digital methods is closer than close reading. Computers are good at counting and memorising and can count the word frequency of an entire novel and memorise it without mistakes, a task which is otherwise impossible for humans. People do not focus on word-level or character-level information such as word frequency but on sentences or content. New reading methods in digital humanities are called distant reading (Moretti, 2000).

1.2. Aim and Research Question

With the help of distant reading, new perspectives of literature are available. In this thesis, we use a digital method, namely, a topic model, to read Swedish literature distantly. The main research questions are as follows:

- Which topics do authors frequently use in a Swedish literature corpus which comprises books from 1879 to 1941?
- Which topics are time-specific in the corpus?
- Which topics characterise each author in the corpus?

In other words, we perform two tasks. The first task is to answer research questions 1 and 2. We track the fluctuation of topic trends over time and discuss what changes and what remains unchanged. Our second task is to answer the third question: thus, we attempt to characterise each author by topic.

The remainder of the thesis is organised as follows. In Chapter 2, we present the literature review. In Chapter 3, we introduce our corpus and preprocess it, build our model for experiments, and describe our experiments in detail. In Chapter 4, we present the results of our experiments and discuss them. In Chapter 5, we conclude the thesis.

2. Literature Review

Topic models are methods for discovering latent topics in documents by using unsupervised means. These models enable researchers to observe similarities among a batch of plain documents or the uniqueness of a document by using topics generated by inference algorithms. Each topic usually comprises semantically similar words such as {baseball, play, field, soccer}. In summary, topic models create groups of statistically similar words from plain texts. Researchers use the groups of words (called topics) in their downstream tasks. Topic models describe each document as a combination of topics. For instance, a model may describe a news article as a combination of 0.3 of topic 1 (e.g. politics) and 0.7 of topic 2 (e.g. economy). If two documents have similar topic distribution such that document A has 0.3 of topic 1 and 0.7 of topic 2, whereas document B has 0.25 of topic 1 and 0.75 of topic 2, the documents are similar according to the model.

2.1. Latent Dirichlet Allocation

Before discussing the literature, we introduce the terminologies we use in this paper. We follow the notation used in Blei et al. (2003) :

- A *word* is the basic unit of discrete data and defined to be an item from a vocabulary indexed by {1,...V}.
- A *document* is a sequence of N words denoted by $\mathbf{w} = (w_1, w_2, ..., w_N)$, where w_n is the nth word in the sequence.
- A *corpus* is a collection of M documents denoted by $\boldsymbol{D} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$, where $\mathbf{w}_m$ is the mth document in the corpus.

Latent Dirichlet allocation (LDA) is a generative probabilistic model for the collection of discrete data proposed by Blei et al. (2003). LDA generates documents in a corpus as follows:

1. For each k in topic $K = 1, ..., k$:
   a) Choose $\beta \sim$ Dirichlet(η)
2. For each $\mathbf{w}$ in M:
   a) Choose $\theta \sim$ Dirichlet(α)
3. For each w_n in N:
   a) choose a topic $z_n \sim$ Multinomial(θ)
   b) choose a word w_n from $p(w_n|z_n, \beta)$.

In other words, β is charactersised by Dirichlet distribution with a hyperparameter η in step 1. β is a $k \times V$ matrix where $\beta = p(w^j = 1|z^i = 1)$. In step 2, LDA characterises a parameter θ which is a topic distribution of a document w_n from Dirichlet distribution with a hyperparameter α such as (0.3, 0.2,...) where 0.3 is a probability of the generative model that chooses topic 0, and 0.2 is a probability for topic 1 in the document. In the next step (step 2-a), the model chooses a topic z_n for a word w_n from the multinomial distribution over the topics defined by parameter θ. The model chooses a word from the topic, based on a probability distribution over words β in the topic in step 2-b. Figure 2.1 presents a model of the process.

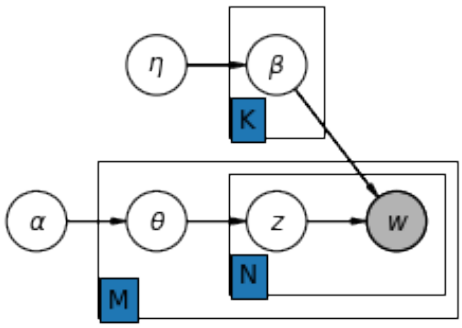


Figure 2.1.: Graphical model representation of LDA. Boxes represent iterations. Specifically, the outer box on the lower side is for document-level iteration, and the inner box is for word-level iteration, namely, the parameter θ is chosen for each document in M, and the parameter z is chosen for each word in a document N. The box on the upper side is for iterations for topics. The nodes are variables in LDA, and the edges show dependencies among them. η, α and β are corpus-level parameters sampled only once in generating documents from a corpus, θ_d is a document-level variable defined for each document, and z_{dn} and w_{dn} are word-level variables. The variable w_{dn} is observable and the others are latent variables. α and η are approximated by inference algorithms.

In building models, hyperparameters α and η are approximated based on observable variables w by using algorithms because it is impossible to compute α and η precisely (Blei et al., 2003). Algorithms for inferring them are either the variational methods Blei et al. (2003) and Hoffman et al. (2010) proposed or the sampling approaches T. L. Griffiths and Steyvers (2004) discussed. Gensim, which we use in this thesis, implements an online variational inference algorithm that follows Hoffman et al. (2010).

The LDA model requires simple modifications to approximate α and η. Figure 2.1 can be modified into Figure 2.2. In the figure, the Dirichlet parameter γ and the multinomial parameters $(\phi_1, ..., \phi_N)$ for a corpus $|D| = N$ are free variational parameters. The per-word topic assignment z is parameterised by ϕ and per-document topic weights θ si parameterised by γ. The distribution of Figure 2.2 is

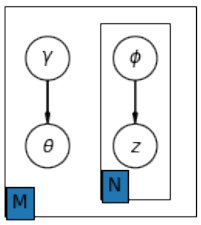


Figure 2.2.: Graphical Model of Modified LDA for Inferring Parameters.

characterised by using the parameters γ and ϕ:

$$q(\theta, z|\gamma, \phi) = q(\theta|\gamma) \prod_{n=1}^{N} q(z_n|\phi_n) \tag{2.1}$$

The variables for Dirichlet distribution α and η are approximated based on the simplified LDA model by the variational Bayesian inference, which is described by analogy to the EM algorithm. In E-step, the algorithm optimises γ and ϕ for each document holding λ, a variable which characterises β fixed. In M-step, the algorithm updates λ given ϕ and the log likelihood of the document-level variables α and η on the basis of λ, γ and ϕ. The algorithm proposed in Hoffman et al. (2010) is in Algorithm 1. ρ in the algorithm is a weight for updating $\boldsymbol{\lambda}$. In the E-step of the algorithm, the expectations are:

$$\mathbb{E}_q[\log\theta_{tk}] = \Psi(\gamma_{tk}) - \Psi(\sum_{i=1}^{K} \gamma_{ti}) \tag{2.2}$$

$$\mathbb{E}_q[\log\beta_{kw}] = \Psi(\boldsymbol{\lambda}_{kw}) - \Psi(\sum_{i=1}^{W} \boldsymbol{\lambda}_{ki}) \tag{2.3}$$

where Ψ denotes the digamma function, $\tilde{\alpha}(\gamma_t)$ is the inverse of the Hessian times the gradient $\nabla_\alpha \ell(n_t, \gamma_t, \phi_t, \boldsymbol{\lambda})$, and $\tilde{\eta}(\boldsymbol{\lambda})$ is the inverse of the Hessian times the gradient $\nabla_\eta \mathcal{L}$. $\mathcal{L}$ is defined as:

$$\mathcal{L} \triangleq \sum_t \ell(n_t, \phi_t, \gamma_t, \boldsymbol{\lambda}) \tag{2.4}$$

Algorithm 1: Online Variational Bayes for LDA

Define $\rho_t \triangleq (\tau_0 + t)^{-k}$;
Initialise $\boldsymbol{\lambda}$ randomly.;
for $t = 0$ *to* ∞ **do**
  E step;
  Initialise $\gamma_{tk} = 1$. (The constant 1 is arbitrary.);
  repeat
    Set $\phi_{twk} \propto \exp\{\mathbb{E}_q[\log\theta_{tk}] + \mathbb{E}_q[\log\beta_{kw}]\}$;
    Set $\gamma_{tk} = \alpha + \sum_w \phi_{twk} n_{tw}$;
  until $\frac{1}{K}\sum_K |\text{change in } \gamma_{tk}| < 0.00001$;
  M step;
  Compute $\tilde{\boldsymbol{\lambda}}_{kw} = \eta + D n_{tw} \phi_{twk}$;
  Set $\boldsymbol{\lambda} = (1 - \rho_t)\boldsymbol{\lambda} + \rho_t \tilde{\boldsymbol{\lambda}}$;
  Set $\alpha = \alpha - \rho_t \tilde{\alpha}(\gamma_t)$;
  Set $\eta = \eta - \rho_t \tilde{\eta}(\boldsymbol{\lambda})$;
end

2.2. Evaluation of Models

To obtain a suitable LDA model for experiments, we must find the optimal number of topics for the model. We use perplexity and coherence as a measure to indicate the best number of topics.

2.2.1. Perplexity

Perplexity is a measurement to compute how well a model predicts samples (samples are documents in the test set in this case) and shows the number of options based on the modelling. Specifically, a perplexity score of 1000 indicates that the model has 1000 options when predicting the correct word in the test set. Perplexity is formally defined in Hoffman et al. (2010) as:

$$\text{perplexity}\left(n^{\text{test}}, \boldsymbol{\lambda}, \alpha\right) \triangleq \exp\left\{-\left(\sum_i \log p\left(n_i^{\text{test}}|\alpha, \beta\right)\right) / \left(\sum_{i,w} n_{iw}^{\text{test}}\right)\right\} \tag{2.5}$$

where n_i^{test} denotes the bag-of-words representation for the ith document. We cannot directly compute $\log p\left(n_i^{\text{test}}|\alpha, \beta\right)$ in LDA because of the dependency among variables. Hence, we compute a lower bound for perplexity as a proxy which is implemented in Gensim based on Hoffman et al. (2010) as:

$$\text{perplexity}\left(n^{\text{test}}, \boldsymbol{\lambda}, \alpha\right) \leq exp\left\{-\left(\sum_i \mathbb{E}_q\left[\log p\left(n_i^{\text{test}}, \theta_i, z_i|\alpha, \beta\right)\right] - \mathbb{E}_q\left[\log q(\theta_i, z_i)\right]\right)\left(\sum_{i,w} n_{iw}^{\text{test}}\right)\right\} \tag{2.6}$$

Perplexity can be used either to find the optimal number of topics (Maskeri et al., 2008)[1] (Ostrowski, 2015) or the best learning parameter settings such as minibatch size (Hoffman et al., 2010): however, the best perplexity score does not directly indicate the best model (Chang et al., 2009).

2.2.2. Coherence

Coherence is another means to evaluate models. It evaluates the quality of topics in models and computes the semantic similarity of words in each topic, while perplexity evaluates models by how well they infer samples. Intuitively, good quality topics have similar words or words which occur in similar contexts. Several approaches to compute coherence have been proposed. The state of the art approach was presented in Röder et al. (2015). Although their approach is approximately 30 times slower than that presented in Mimno et al. (2011), who proposed the fastest method, it achieves the strongest correlation with human ratings.

Methods to compute coherence is composed of four stages; segmentation, probability estimation, confirmation measure and aggregation. In segmentation, words in a topic are divided into a pair of subsets, W^a and W^* to examine whether the existence of W^* supports the occurrence of W^a in succeeding steps. In probability estimation, probability of the words is computed by using a reference corpus. Note that, a reference corpus can be either the corpus which is used to build the model or an extrinsic corpus. For instance, Röder et al. (2015) used both the corpus which they used to build models and a Wikipedia corpus. In confirmation measure, the subsets W^a and W^* are used to compute how strong the conditioning word set W^* supports W^a. In the last step, confirmation scores of all subsets are aggregated to a coherence score.

Segmentation

All of the methods for coherence have slightly different schemes in every step. In the segmentation step, the method Röder et al. (2015), which we use in the thesis divides top n words W in each topic into a word W^a and all of the words including the word chosen as a subset W^a, which is formally described as follows:

$$S = \{(W^a, W^*)|W^a = \{w_i\}; w_i \in W; W^* = W\} \tag{2.7}$$

Probability Estimation

Computing the probability is required in the next step, confirmation. The probability of words is obtained from a reference corpus. Röder et al. (2015) used a sliding window. The window moves over documents in the corpus, and each step defines a new document. The probability of a word is computed as the number of documents in which the word occurs divided by the total number of documents.

[1] They computed maximum likelihood. Perplexity can be computed by using maximum likelihood. Therefore, maximum likelihood and perplexity are related.

Confirmation

The confirmation measure takes a pair S of words and computes how strong the pair related. Subsets W^a and W^* are represented as vectors $\vec{v}$ and $\vec{w}$ respectively to obtain confirmation scores. Elements in a vector are weighted by normalised pointwise mutual information (NPMI) in building a vector and vectors are described as follows:

$$\vec{v}(W^a) = \{ \sum_{w_i \in W^a} \text{NPMI}(w_i, w_j)^\gamma \}_{j=1,\ldots|W|} \tag{2.8}$$

γ in Equation 2.8 is a constant number. Röder et al. (2015) used 1. NPMI can be computed as:

$$\text{NPMI}(w_i, w_j) = \frac{\text{PMI}(w_i, w_j)}{-\log(p(w_i, w_j) + \epsilon)} \tag{2.9}$$

where pointwize mutual information (PMI) is computed as:

$$\text{PMI}(w_i, w_j) = \log_2 \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)} \tag{2.10}$$

The confirmation score is computed by cosine similarity.

$$Sim(\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{\|\vec{v}\|\|\vec{w}\|} \tag{2.11}$$

Aggregation

Confirmation scores from all subset pairs are aggregated to a coherence score by the arithmetic mean in the aggregation step. The pipeline is already implemented in Gensim, so we do not have to implement them from scratch.

2.2.3. Topic Labelling

Researchers have to evaluate the topics of their model manually or automatically to label them because LDA topics built by algorithms are probability distribution over words. Researchers can label topics manually by evaluating them qualitatively. By doing so, researchers choose the most suitable labels for topics in most cases. However, manual evaluation by nature subjective and lack of reproducibility.

There are methods to label topics automatically. In LDA topics, terms w_i are presented with the marginal probability $p(w_i|t_j)$ for topics t_j. Therefore, the simplest way for automatic labelling of LDA topics is to choose the words with the highest $p(w_i|t_j)$ as labels of the topics. Lau et al. (2010) proposed a method to choose the most representative word from the top 10 topic words in that topic by an unsupervised way and by a supervised way with several features such as the mutual information.

Chen et al. (2006), Mei et al. (2007) and Lau et al. (2011) proposed methods to generate labels. By and large, there are two steps for topic labelling. The first step is to generate label candidates from external data, and the second step is to choose the best label from the candidates. Chen et al. (2006) computed mutual information for bigram words to generate candidate labels. They proposed a formula, LSA-weighted frequency to choose the best bigram word from the candidates for labelling a topic.

Their method was originally for topics generated by latent semantic analysis (LSA), but it is applicable to LDA. Mei et al. (2007) generated candidate labels by both chunking (shallow parsing) and an n-gram testing approach, and they argued that the n-gram testing approach is the better option in most cases. They chose the best label from candidates by either of two methods, the zero-order relevance and the first-order relevance. The zero-order relevance used the marginal probability $p(w_i|t_j)$ to rank candidates, and the first-order relevance computed KL divergence between the probability in which the topic generates the candidate label and the probability in which a context collection (an external corpus) generates the candidate label. The first-order relevance was more robust than the zero-order relevance in their experiments. Lau et al. (2011) used the top-10 topic terms from the original topic model and English Wikipedia articles to generate candidate labels. They used a supervised method, support vector regression to rank the candidates.

Cano Basave et al. (2014) proposed a method to generate labels for topics by summarising documents. They collected documents relevant to a topic and summarised the documents to generate a label. The advantage of their method is that it does not require external data to generate labels.

2.3. Application of LDA

Dahllöf and Berglund (2019) and Barakat (2018) have applied topic models to Swedish literature. Dahllöf and Berglund (2019) applied LDA to Swedish classics and modern bestsellers and observed character-gender-biased topics and author-gender-biased topics. Barakat (2018) used LDA in a part of his thesis to determine the reasons for an audiobook's popularity. He mainly conducted a quantitative analysis by using a held-out set. They observed shared features among popular audiobooks by using several models including topic models: however, the topic models are noisy because of proper nouns and inflexion.

Topic models have been applied to Swedish documents other than literature. Magnusson et al. (2018) applied LDA to speeches in the Swedish Parliament. They tracked the trend of the immigration topics and discussed them along with two social events: the rise of a radical right party in 2010 and the refugee crisis in 2015 to describe the impacts of the two events. Norén and Snickars (2017) applied LDA to Swedish Governmental Official Reports to find the history of Swedish film politics. Their model revealed that film-related LDA topic intensively appeared from 1945 to 1955 when the Swedish government planned for the Film Reform.

Outside the Swedish context, topic models have been applied to various texts. In literature, researchers have applied LDA to different languages. Jockers (2013) built LDA models on a corpus composed of 3,346 books in the nineteenth century. He observed that nations and genders have clear thematic preferences or tendencies. For instance, he observed that a topic for affection and happiness appeared in books by female authors twice as much as in books by male authors.

Tangherlini and Leonard (2013) used LDA to Danish literature. Their approach was unique in that they built their models on sub-corpora composed of well-known Danish books and used the models to observe impacts of the books in Danish

literature corpus in which most of the books are not read today. They used the models to extract similar sentences from the books in the Danish literature corpus. They argued that the similar sentences they extracted from the corpus by using the models are the impacts of the well-known books on which they built the models.

Navarro-Colorado (2018) applied LDA to Spanish golden age sonnets corpus. They compared topics generated by LDA to topics researchers generated from the sonnets, and they observed that LDA extracts not only thematic topics but also poetic motifs. It implies that topics people assume do not always match with LDA topics.

Roe et al. (2014) built LDA models on French Encyclopédie published in the 18th century. They observed latent topics which were not observed in the way documents in Encyclopédie were classified by using LDA-topics. For instance, The authors observed the discourse about morality among many classes such as grammar and geography class. They argued that philosophers at the time of editing hid controversial opinions in the articles.

In fields other than literature, Maskeri et al. (2008) applied LDA to source codes to understand what cords to aim for. They attempted to effectively classify cord snippets inside types of software for engineers to indicate which codes the engineers wanted to fix. Bastani et al. (2019) applied LDA to the Consumer Financial Protection Bureau corpus, which comprises consumer financial complaints, to observe trends over time. To observe trends, they proposed topic popularity based on the topics they extracted.

They built their model on their corpus and computed topic popularity based on the topics generated by the model. The topic popularity expresses time trends of topics. They let $D_t = \{d_t^1, d_t^2, ..., d_t^{n_t}\}$ be the collection of documents (in their corpus, a document is a complaint from a customer.) received at a time index t, d_t^j be the jth document in the collection, and n_t be the total number of documents in D_t. The authors define topic popularity (TP) for topic id i at a time index t as:

$$\mathrm{TP}_{i,t} = \frac{\sum_{t=1}^{n_t} \theta_{i,d_t^j}}{n_t} \qquad (\forall i = 1, ..., K; \forall t = 1, ..., T) \tag{2.12}$$

θ_{i,d_t^j} denotes the probability distribution of topic id i distributed to a document d_t^j, K is the total number of topics, and T is the total number of time indices.

Other than Bastani et al. (2019), Jockers (2013), Magnusson et al. (2018) and Norén and Snickars (2017) computed topic popularities over time. In Jockers (2013), he used a five-year moving average to reduce noise in plotting time trends. In Magnusson et al. (2018), they computed time trends as proportion of all tokens. In Norén and Snickars (2017), they seem to have computed topic distribution of documents per year, namely, they used $t = 1$ *year* for computing time trends. Unfortunately, none of them described the method to compute time trends formally.

Our thesis is similar to Dahllöf and Berglund (2019) because we apply LDA to Swedish literature, and toJockers (2013), Bastani et al. (2019), Magnusson et al. (2018) and Norén and Snickars (2017) because we track the trends of topics. However, according to our review of researches, this paper is the first to describe patterns in Swedish literature over time by using topic models.

3. LDA Based Topic Modelling for Swedish Literature

In this chapter, we present an overview of our corpus in Section 3.1 and preprocess it for our experiments in Section 3.2. Next, we discuss the hyperparameters of our model and build our model, and the method for representing our model in Section 3.3.

We conduct two experiments using the model. The first experiment is to apply the model to the whole corpus to find the time trends of all the topics in Section 3.4. The second experiment is to apply the model to subcorpora composed of each author's work to visualise the differences among the authors regarding topics each author preferred in Section 3.5.

3.1. Corpus

We use the same corpus as in Dahllöf and Berglund (2019). The corpus comprises 118 Swedish novels and prose collected in Litteraturbanken[1]. The books are published from 1879 to 1941 by 35 authors. Each author and their number of works are listed in Table 3.1. Many of the documents were written between 1908 and 1935. Regarding the size of the corpus, the corpus comprises 6,518,354 words and 126,536 types, excluding punctuation. We preprocess the corpus in Section 3.2 because the original corpus is impractical for computational methods for several reasons we discuss in preprocessing.

Table 3.1.: Each author and their number of works

Name	Born-Died	# Work
Agrell, Alfhild	1849-1923	1
Andersson, Dan	1888-1920	5
Angered Strandberg, Hilma	1855-1927	2
Berger, Henning	1872-1924	2
Bergman, Hjalmar	1883-1931	14
Beskow, Elisabeth	1870-1929	1
Boye, Karin	1900-1941	8
Duse, Samuel August	1873-1933	1
Ek, Karin	1885-1926	1
Fitinghoff, Laura	1848-1908	1
Hansson, Ola	1860-1925	2

Continue to the next page

[1] https://litteraturbanken.se/

Table 3.1.: Each author and their number of works

Name	Born-Died	# Work
Heidenstam, Verner von	1859-1940	2
Hemmer, Jarl	1893-1944	2
Koch, Martin	1882-1940	4
Krusenstjerna, Agnes von	1894-1940	1
Kämpe, Alfred	1877-1936	4
Lagerlöf, Selma	1858-1940	16
Landquist, Ellen	1883-1916	1
Lybeck, Mikael	1864-1925	1
Malling, Mathilda	1864-1942	1
Månsson, Fabian	1872-1938	2
Mörne, Arvid	1876-1946	2
Nyblom, Helena	1843-1926	2
Regis, Julius	1889-1925	1
Sandel, Maria	1870-1927	5
Schildt, Runar	1888-1925	1
Sjöberg, Birger	1885-1929	2
Stéenhoff, Frida	1865-1945	1
Strindberg, August	1849-1912	23
Söderberg, Hjalmar	1869-1941	1
Söderberg, Mikael	1903-1931	2
Söderholm, Kerstin	1897-1943	1
Värnlund, Rudolf	1900-1945	3
Wahlenberg, Anna	1858-1933	1
Witt, Otto	1875-1923	1

End of the table

3.2. Preprocess

Preprocessing is the critical first step of natural language processing tasks (Uysal and Günal, 2014). Tokenisation, stopword removal, lowercase conversion and stemming have been applied to data in many studies. In this thesis, we divide all books in the corpus into small chunks first in Section 3.2.1. After that, we apply tokenisation and lemmatisation by using a tagger. Next, we remove stopwords and uncommon words. Moreover, we remove words other than nouns from the corpus because we only use nouns to build our models in Section 3.2.2.

3.2.1. Chunking the Documents

To build the LDA models, long documents are segmented into smaller word chunks to improve models' quality. LDA uses the bag-of-words representation of documents in building models. Therefore, information other than word frequency, such as word order or the structure of paragraphs, is ignored. This limitation is not

critical when each document is relatively small, such as news articles. However, each document we use to build models is a book, each of which is a quite long document. If a book is written on a few topics, modelling on books without chunking is effective. However, a book usually comprises many scenes or themes: some of them are described in a part of the book, and others are described in the whole book. Imagine that a superhero lives as a normal person at the beginning of a novel. She fights against a villain in most of the book other than in the beginning. Chunking the book helps for topic models to capture the daily life at the beginning as a topic because the models focus on local information by chunking the book into a smaller size. The chunk size differs based on the aim of the research. Dahllöf and Berglund (2019) used 12 and 24 as their chunk size and Barakat (2018) used 500. We follow Jockers (2013), who used 1,000 size chunks. Our corpus, which is originally composed of 118 documents (118 books), comprises 2718 1000-chunk documents after chunking the books.

3.2.2. Removing Stopwords and Infrequent Words

The first preprocessing method we use is tokenisation. We lowercase, lemmatise words and indicate part-of-speeches (PoS) of the words in the corpus by using efficient sequence labelling (Östling, 2018). He proposed a perceptron-based part-of-speech tagger, which outperforms a neural network tagger. In their experiments, the average error rate of their tagger was 6.7%. We observe some errors in tokenising and PoS tagging of our corpus by the tagger. For instance, some apostrophes are interpreted as nouns, which are removed in this step.

Regarding the distribution of distinct words, also called types, in natural languages, they follow the Zipf's distribution (Piantadosi, 2014). That is, a few types, such as articles account for a high proportion of texts whereas most types are rarely observed. Because the types with high frequency occur in many contexts in most documents, they are not effective for building a topic model that computes the co-occurrence of words. Specifically, the number of words in our corpus is 6,518,354, and it comprises 126,536 types. Word frequency of the corpus is plotted in Figure 3.1, where the first-ranking word type is observed more than 100,000 (10^5) times (257,888 times in total). Frequent words such as the first-ranking word are stopwords.

Stopwords refer to types which are very often observed independently of contexts. Researchers must define their own stopwords for their tasks because there is no universal list of stopwords. We define the top 100 frequent words as our stopwords and remove them from our corpus. As a result of stopword removal, the tokens in the corpus decreased to 3,030,593, and the types decreased to 126,436. Although we remove only 100 types, the words reduce by half. Note that we remove part-of-speeches other than nouns later. Therefore most of the stopwords are discarded in the preprocessing process in any case. However, four words out of the100 stopwords we have removed here are nouns (Table 3.2). Those words will not be removed in later steps if we do not remove them here.

We also remove uncommon words from the corpus. The rank of the 63,000th-word type has a 10^0 frequency in Figure 3.1, which indicates that more than half of the types in the corpus occur only once. These uncommon words are not useful

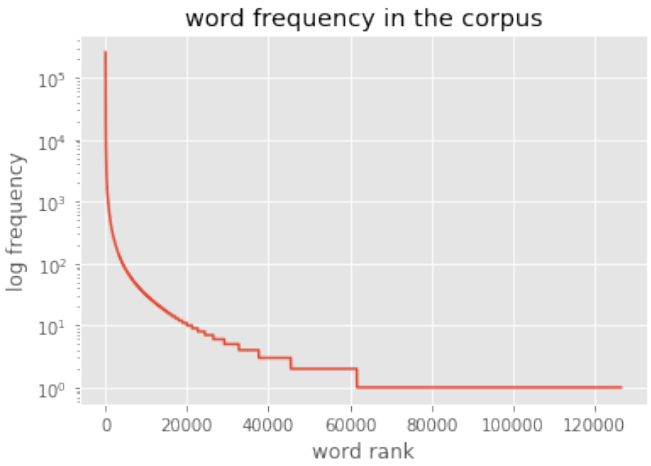


Figure 3.1.: Word frequency of the corpus after it is tokenized.

Sv	gång	dag	hand	man
Eng.	time	day	hand	man

Table 3.2.: Four nouns in our stopwords.

for topic models because they are specific to their context. Other than words that have a 10^0 frequency, we remove types that appear in fewer than ten books. This approach is the same approach used in Dahllöf and Berglund (2019). Additionally, we focus on nouns and exclude other parts of speech to reduce the size of the corpus further and to build topic models effectively. Some researches have focused on nouns (Jockers, 2013) or nouns and verbs (Dahllöf and Berglund, 2019). In any case, focusing on content words (i.e. nouns, verbs, adjectives, and adverbs) is essential to improve the quality of topic models. After removing uncommon words and the parts of speech except for nouns, the size of our corpus decreases 938,461 words comprising 6,395 types.

3.3. Topic Models

3.3.1. Model Parameter Decisions

When building topic models, researchers must define the number of topics in the model. In general, a model with many topics becomes a fine-grained model. Each topic in such a model tends to be composed of words from specific events or detailed descriptions of things. A model with a few topics becomes a coarse-grained model suitable for noticing broad themes (Dahllöf and Berglund, 2019), (Magnusson et al., 2018). Ultimately, researchers must decide on the number of topics (k hereafter) by themselves by conducting qualitative evaluation according to their tasks or by using quantitative criteria such as perplexity or coherence. We compute both to make decisions regarding our model. We build latent Dirichlet

allocation (LDA) topic models with k = {10, 20, 30, 40, 50, 60, 70, 80, 90, 100} topics on the training set and compute perplexity and coherence for each model to decide the model we use in experiments.

We use Gensim, a Python library for topic models to build the models. The parameters we use are in Table 3.3. The iteration in the table is the upper bound of the iterations in the E-step of the Algorithm 1, and passes are the number of iterations of the entire corpus, which is the iteration in the general meaning. We found a suitable iteration and passes by preliminary experiments. Alpha is the prior belief for each topic's probability. Eta is the prior belief of the word probability. We follow the recommendations of Steyvers and T. Griffiths (2007) for alpha and eta. They are updated in building models iteratively. It is the goal to find suitable alpha and eta given the corpus. The chank size denotes the number of documents which are used for each iteration. According to Hoffman et al. (2010) the chunk size should be large, and they recommend at least 256. Building models with bigger chunk size finishes quicker than with smaller chunk size, though big chunks require large memory. Thus, practitioners can decide the size according to their environment. We used default values for parameters other than ones described here.

# topics (k)	{10,20,30,40,50,60,70,80,90,100}
chunk size	2048
passes	2000
iterations	1024
alpha	50 / # topics
eta	0.01

Table 3.3.: Parameters for the experiment.

To obtain the best model, we build ten models on various topics on a training corpus composed of 90% of our corpus. After building all the models, we compute the perplexity by using a test set composed of the remainder (10%) of the corpus and coherence of each model by using the training corpus as a reference corpus. We use a method in Hoffman et al. (2010) to compute perplexity and a method in Röder et al. (2015) to compute coherence, both of which we discuss in Section 2.2.

3.3.2. Topic Representation

Researchers may choose representative words based on the probability distribution of words in topics, $p(word|topic)$ to represent their models. However, if representative words of a topic are chosen in that manner, the words tend to be popular among several topics because words that often appear in the corpus may have a high probability $p(word|topic)$ in several topics. Therefore, it is desirable to choose representatives not by computing $p(word|topic)$, but by how much words are dependent on their topics, which is intuitively described as $p(topic|word)$, although we do not compute the marginal probability of $p(topic|word)$ directly.

We use the χ^2 (chi-square) test (Manning et al., 2008) to choose the representatives of the topics. The method is a standard measure for computing the

association between two categorical variables. Specifically, the method is used to assess whether an instance A is dependent on class X, such as the relationship between sex and political party. In this thesis, we test whether a word type A is dependent on a specific topic X. We compute χ^2 as follows:

$$\chi^2(w_i, t_a) = \sum_{w_i \in 0,1} \sum_{t_a \in 0,1} \frac{(N_{w_i,t_a} - E_{w_i,t_a})^2}{E_{w_i,t_a}} \tag{3.1}$$

where $w_i = 1$ denotes a word is specific word for which we want to compute the score, $w_i = 0$ denotes a word that is not the target word, $t_a = 1$ denotes a word that belongs to a specific topic, and $t_a = 0$ is for a word that belongs to the other topics. We can write a confusion matrix based the four categories (Table 3.4). In Equation

	$t_a = 1$	$t_a = 0$
$w_i = 1$	N_{11}, E_{11}	N_{10}, E_{10}
$w_i = 0$	N_{01}, E_{01}	N_{00}, E_{00}

Table 3.4.: A confusion matrix for computing the χ^2 score for a word w_i

3.1, N_{w_i,t_a} is the observed number of words, and E_{w_i,t_a} is the expected number of words. For instance, $N_{1,1}$ denotes the actual number of times the word type w_i is observed in the corpus from the topic t_a, and $E_{1,0}$ is the expected number of times the word type w_i is observed in the corpus from the topics other than t_a. We can obtain N_{w_i,t_a} by counting them through the corpus and E_{w_i,t_a} is computed as follows assuming the type w_i and topic t_a are independent;

$$E_{w_i,t_a} = N * p(w_i) * p(t_a) \tag{3.2}$$

where N is the size of the corpus. Specifically, E_{11} is computed as follows:

$$\begin{aligned} E_{1,1} &= N * p(w_i) * p(t_a) \\ &= N * \frac{N_{1,1} + N_{1,0}}{N} * \frac{N_{0,1} + N_{0,0}}{N} \end{aligned}$$

By computing N_{w_i,t_a} and E_{w_i,t_a} and plugging them into Equation 3.1 we compute the χ^2 score for each type from a specific topic. We interpret the χ^2 score as a score showing how strongly a word type depends on a topic. That is, word types that have a higher score are good representatives of a topic.

3.3.3. Topic Labelling

We label topics both manually and automatically. We first carry out a manual evaluation. Some topics are difficult for us to interpret their theme because words in topics are not coherent semantically, and topics have several semantic themes. Hence, we label topics either by a term or by several terms.

Next, we implement an automatic labelling method based on Mei et al. (2007). To generate label candidates, we collect bigrams from our corpus with window size 10. Specifically, given a word sequences $w_1, w_2, ..., w_{10}$ we collect bigrams

$w_1w_2, w_1w_3, \ldots, w_9w_{10}$. Note that whereas Mei et al. (2007) used an external corpus to collect bigrams, we use the same corpus as we use to build our LDA topic model.

We apply mutual pointwise information (PMI) to the bigrams generated from the corpus to build a candidate label collection. Formally, for word pairs w_i, w_j we compute PMI as;

$$PMI(w_i, w_j) = \log \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{3.3}$$

The probabilities in the equation are computed by the maximum likelihood estimation.

$$PMI(w_i, w_j) = \frac{N_{w_i, w_j}/N_b}{N_{w_i}/N * N_{w_j}/N} \tag{3.4}$$

where N_{w_i, w_j} is the number of the occurrence of the bigram w_i, w_j, N_b is the number of bigrams in the corpus, N_{w_i} and N_{w_j} are the number of occurrence of the word types w_i and w_j, and N denotes the number of words in the corpus. Bigrams above a threshold score become label candidates.

We implement the zero-order relevance (Mei et al., 2007) to rank the label candidates and to decide labels for topics. The score of each candidate bigram w_iw_j given a topic θ is computed as;

$$Score(w_i, w_j) = \log \frac{P(w_i|\theta)}{P(w_i)} + \log \frac{P(w_j|\theta)}{P(w_j)} \tag{3.5}$$

where $P(w_i|\theta)$ and $P(w_j|\theta)$ are the marginal probability distribution of the word types w_i and w_j in a topic θ. A bigram with the highest score becomes a label for topic θ. We use the zero-order relevance though Mei et al. (2007) proposed a more robust method, the first-order relevance because it requires an external corpus to choose the best label from the candidates.

3.4. Experiment 1: Trends in Swedish Literature

Trends change over time in the literature. Movements such as modernism are explanations of trends by critics. What critics have described, however, is based on their close reading. Close reading is an aspect of a reading activity, and another method of reading is distant reading by digital methods such as LDA. The trends in LDA provide a new insight into literature because they are a mathematical expression of books written in the probability distribution, not by critics' observations. In this section, we introduce how to describe time trends in Swedish literature in our corpus by using the model we built in Section 3.3.1.

We follow the method proposed by Bastani et al. (2019), which we discuss in Section 2.3. For instance, assuming there are two documents in a specific time unit x which have topic distributions such as d_x^1 = {(1, 0.3), (2, 0.5), (3, 0.2)} and d_x^2 = {(1, 0.1), (2, 0.2), (3, 0.7)}. The first element in the parentheses is the assigned topic id, and the second element is its proportion. Hence the notation is interpreted such

that 30% of the d_x^1 comprises topic 1, 50% comprises topic 2 and so forth. Then, the topic popularity of topic 1 at time index x is computed as follows:

$$TP_{x,1} = \frac{0.3 + 0.1}{2} = 0.2 \tag{3.6}$$

In summary, $TP_{x,1}$ is an arithmetic mean of topic 1 at time x. By comparing $TP_{x,1}$ to $TP_{x+1,1}$, $TP_{x+2,1}$ ..., and $TP_{T,1}$, we obtain the time trend of the topic 1. The score increases over time if the topic becomes popular among authors but decreases when it becomes obsolete.

To compute time trends for all topics, we divide the corpus into subcorpora by time and apply the 5-year window to divide it. We slide the window by a year such that the first subcorpus is composed of documents written from 1879 to 1883, the second subcorpus is from 1880 to 1884, and so forth. We apply topic popularity to each subcorpus and plot the fluctuation of popularity for comparison.

3.5. Experiment 2: Author Characteristics

When reading a work of literature, sometimes the characteristics of the author are revealed. For instance, August Strindberg combines psychology and naturalism in a new type of European drama according to Encyclopedia Britannica. This type of explanation is a traditional means of describing the characteristics of authors. In this experiment, we carry out the method to describe similarities or differences among authors in our corpus. The advantage of an LDA model is that topics are statistical and accountable. For instance, an LDA model with five topics represents a document as a mixture of five topics, for example, w_1 = (1, 0.2), (2, 0.3), (3, 0.1), (4, 0.2), and (5, 0.2), where each parenthesis represents the proportion of each topic such that 20% of the document w_1 is composed of topic 1. In other words, LDA represents documents as K-dimensional vectors, where each feature of K is a topic in the model. Topic 1 may be words for chemistry, and topic 2 may be for statistics. By using these features we obtain by LDA, we describe authors statistically and visualise individuality or similarity among them in K dimensional vectors.

To describe authors by LDA, we divide the corpus into subcorpora by authors. For example, a subcorpus comprises 23 books by August Strindberg. To compute the topic preference of authors, we compute the topic distribution of each author, which is a variation of the topic trends in Section 3.4. Formally, we compute each topic's probability distribution of authors, the author topic (AT), as follows:

$$AT_{i,a} = \frac{\sum_{j=1}^{n_t} \theta_{i,d_a^j}}{n_t} \qquad (\forall i = 1, ..., K; \forall a = 1, ..., A) \tag{3.7}$$

Equation 3.7 is almost identical to Equation 2.12. In Equation 3.7, a document collection is $D_a = \{d_a^1, d_a^2, ..., d_a^{n_t}\}$ written by an author a, A is the total number of author indices (A = 35 in the corpus), and d_a^j is the jth document in the collection, where n_t represents the total number of documents in D_a. We compute Ecuation 3.7 for every topic, from $i = 1$ to $i = K$ to represent AT_a, which is the topic distribution of an author a. In summary, we compute the arithmetic mean of the topic distribution of all documents by each author.

We compare TP of each author to the topic distribution of the entire corpus to find characteristics of the author. Topics with quite higher or lower probability distribution in the TP than in the entire corpus are characteristic topics of the author.

4. Results and Discussion

In this chapter, we discuss our model used in the experiments in Section 4.1, describe our corpus by using the model in Section 4.2, and descuss results of our experiments in Section 4.3 and 4.4.

4.1. Model Parameters

We compute the perplexity and the coherence of the models with k = {10, 20, 30, 40, 50, 60, 70, 80, 90, 100} built on 90 % of the corpus. We use the implementation of the perplexity by Gensim and the reminder of the corpus (10%) for a test set in computing the perplexity. In Figure 4.1, a lower perplexity score indicates a better model. The score increases monotonically until k = 40 and increases again from k = 80. For coherence, we use Gensim's implementation of Röder et al. (2015) and the corpus used to build models as a reference corpus to compute the probability of the words. Figure 4.2 presents the coherence scores, where a higher score indicates a better model. As is shown, the model with 50 topics has the best coherence score. We propose that k = 20 (the lowest perplexity except for k = 10), k = 30 (the third-lowest perplexity and the fourth-best coherence), and k = 50 (the second-lowest perplexity and the highest coherence) are satisfactory options.

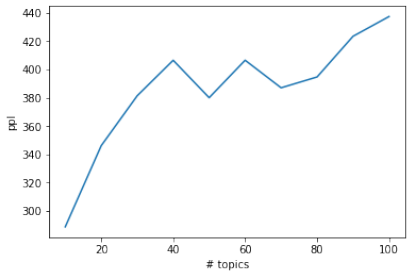


Figure 4.1.: Perplexity

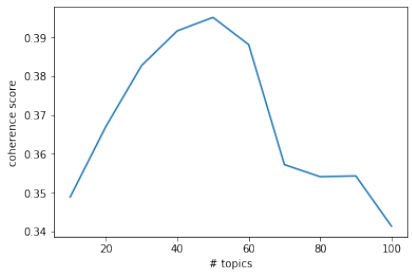


Figure 4.2.: Coherence

Next, we build models of k = {20, 30, 50} based on the entire corpus and compute coherence again to create our model for the experiments by using the entire corpus as a reference (Table 4.1). In the context of machine learning, researchers build their model on only a training set because they want their model to work well on unseen data rather than fitting the data they already have. By contrast, we build our model on the whole corpus for the model to fit the corpus as much as possible because the generalisation of our model is unnecessary. The coherence scores in Table 4.1 differ from Figure 4.2 because we use the entire corpus rather than training corpus. We use the k = 30 model for our experiments because the k

= 50 model requires a heavy workload compared with the other models though the gain of the score from k = 30 to k = 50 is small, and because the k = 20 model has a lower coherence score than the k = 30 model.

k	Coherence
20	0.376
30	0.401
50	0.408

Table 4.1.: Coherence score of the models trained on the basis of the whole corpus.

4.2. Description of the Corpus by Topic Model

Figure 4.3 presents the topic distribution of our corpus. Topic 6 has the highest distribution, followed by topics 21 and 26. Table 4.2 presents a brief description of the topics we discuss in this chapter, and Appendix A presents a detailed description of all the topics with 30 keywords.

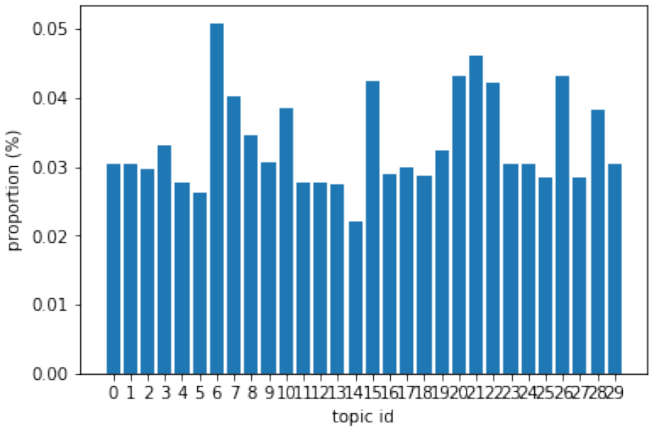


Figure 4.3.: Topic distribution of the corpus. Y-axis denotes the proportion in the entire corpus of each topic.

ID	Label	Keywords
0	Water boat water	båt, strand, vatten, sjö, ö, prinsessa [boat, shore, water, lake, island, princess]
1	Animal + nature + Person boy pair	pojke, hund, djur, fågel, vinge, unge [boy, dog, animal, bird, wing, young bird]
2	Relative + Time uncle mother	fröken, brev, tant, moster, morbror [unmarried woman, letter, aunt, aunt, uncle]

Continue to the next page

ID	Label	Keywords
3	Family + House mother father	fru, mamma, pappa, löjtnant [married woman, mom, papa, lieutenant]
4	Temple + Ruler temple town	gud, grav, jord, kors, tempel, död, frid [god, grave, earth, cross, temple, death, peace]
5	Town town people	stad, gata, herre, hus, port [town, street, gentleman, house, front door]
6	Work + Person shoemaker work	blick, arbete, kamrat, ansikte, kväll, morsa [look, work, comrade, face, evening, mamma]
7	Nature wave sea	sol, träd, luft, berg, himmel, sten, vind [sun, tree, air, mountain, sky, stone, wind]
8	Family daughter son	son, moder, fader, dotter, syster, hem [son, mother, father, daughter, sister, home]
9	Crime + Town inspector doctor	doktor, polis, domare, boll, bil, kommissarie [doctor, police, judge, ball, car, chief inspector]
10	Farm + person cart horse	gård, karl, väg, stuga, häst [yard, man, road, small house, horse]
11	Book lantern light	bok, slag, svar, papper, bild, tystnad [book, type, answer, paper, picture, silence]
12	King + War king country	kung, konung, slott, prins, svensk, soldat [king, king, palace, prince, Swede, soldier]
13	Master + Village master night	magister, jungfru, vagn, trädgård [master, virgin, carriage, garden]
15	Society + School class school	tidning, samhälle, författare, professor [newspaper, society, author, professor]
16	Title + Office Mr. gentleman	herr, kapten, överste, rektor [gentleman, captain, colonel, headmaster]
18	Art + Guest audience gentleman	gäst, dam, sällskap, scen, konst, publik [guest, lady, company, stage, art, audience]
19	Impression + Mankind miracle air	natur, intryck, mänsklighet, fruktan, brist [nature, impression, mankind, fear, lack]
20	Soul + Body life world	själ, hjärta, dröm, liv, lycka, blomma [soul, heart, dream, life, happiness, flower]
21	Girl + party + Occurrence rogue lady	flicka, fästman, rest, händelse, bankir [girl, fiancé, remainder, occurrence, banker]
22	Thought wrong right	människa, sätt, sak, fall, rätt, tanke [man, way, thing, case, right, thought]
24	Feeling + Punishment fate friend	vän, öde, liv, hat, ensamhet, lidande, fiol, skuld [friend, fate, life, hatred, solitude, violin, debt]
25	Christianity priest church	präst, kyrka, pastor, församling, bibel [priest, church, pastor, assembly, bible]
26	House lamp light	dörr, rum, fönster, golv, vägg, säng, steg [door, room, window, floor, wall, bed, step]
27	Family father son	far, mor, gumma, baron [father, mother, old woman, baron]

Continue to the next page

ID	Label	Keywords
28	Forest + Village forest people	skog, eld, björn, by, stig, mark, troll [forest, fire, bear, village, path, mark, troll]
29	Shop + Money money thing	peng, gubbe, mat, glas, klocka, affär, krona [coin, old man, food, glass, clock, affair, krona]

End of the table

Table 4.2.: Brief topic description and translations into English. Labels on the top row of the label column are manual labels and bigram labels on the bottom row are automatically generated labels.

Topic 6 has the highest proportion in Figure 4.3: thus, the word types in topics 6 occur most of all, although the topic is ambiguous. The ambiguity results from the fact that topics include semantically different types. For example, topic 6 includes kväll (evening in English) though it is mainly composed of word types on work and person. Given the inferring algorithm, hyperparameters and the corpus we use, types in a topic have similar characteristics among them although some are different semantically. In other words, kväll in topic 6 is related to the other types in topic 6 mathematically. We may observe how they are related among them when we evaluate books which include topic 6.

We label topics mainly comprised of a single theme with a term and topics comprised of several themes with several terms. For instance, we label topic 6 as work + people because we can divide the top 30 word types of topic 6 into two categories by semantic meaning, namely, person (e.g. blick, ansikte) and work (e.g. arbete, kamrat). The combination of the two semantically different groups in a topic probably occurs because the books in the corpus use the two groups in chunks at the same time. Specifically, words for person and work often co-occur in the corpus.

Ideally, each topic has a specific theme, and it does not include another. Many of our topics have several semantic groups of types, probably because our model only has 30 topics. Comparing our model to a model with fifty topics we built in Section 4.1, it has distinctive topics for each theme of topic 6 of our model (Table 4.3). The more topics a model has, the more distinctive each topic becomes. Therefore, the ideal number of topics in models vary depending on the research aim.

Model	Topic id	Keywords
30 topics	6	blick, arbete, kamrat, ansikte
50 topics	5	ansikte, öga, röst, blick
50 topics	25	arbete, peng, arbetare, krona

Table 4.3.: Comparison of topic 6 from the model with 30 topics to topics 5 and 25 from the model with 50 topics.

Some topics are more difficult to interpret than others because of semantically different types in them. For instance, it is challenging to evaluate the theme of

topic 21. The difficulty of evaluation roughly follows the coherence score of each topic. Topic 0, which has a clear theme, has the highest coherence score, and topic 21 has a low score. The coherence scores of all topics are in Appendix A.

The coherence score does not describe the quality of topics. It is one thing to evaluate topics mathematically by the coherence, and the quality of topics which is vital for labelling them is another thing. Although the quality of topics roughly relates to the coherence score, there are some exceptions. The labels of the topics in Table 4.2 are generated both manually and automatically. The labels on the top row are manually added to the topics, and bigram labels on the bottom row are automatically generated. It is a bit challenging for us to label topics with low coherence scores such as topic 13 (the coherence score 0.33), 21 (0.32) except for topic 5 (0.33). We label it with ease because of word types relating to town in the top 30 terms. Also, the types help the algorithm to generate the bigram label (town people).

Comparing manual labels to bigram labels generated automatically, we observe an advantage and a flaw of the labelling algorithm. The algorithm ranks candidate bigrams by the probability. It is the advantage for the algorithm to consider the probability. We label topic 8 and 27 as family, whereas the algorithm labels them as (daughter son) and (father son). The algorithm adds different labels to similar topics by using the probability. At the same time, it is a flaw for the algorithm to consider probability only. It generates less informative labels because it does not catch semantic information. Some labels such as topic 7 (wave sea) and topic 26 (lamp light) are redundant, and some are difficult to understand such as topic 21 (rogue lady)[1].

4.3. Experiment 1: Topic Popularity

In this section, we discuss the time trends of topics to answer the first two research questions (Section 1.2). We choose two types of topics from the topic description (Table A) to observe topic popularity. We discuss the first research question by using nine topics in Section 4.3.1 and the second research question by using 11 topics in Section 4.3.2

4.3.1. Popular Topics

In this section, we discuss popular topics in the corpus. Thresholds are necessary to choose popular topics and discuss what they have in common. We define popular topics as topics with more than a 1/30 proportion (more than the average probability distribution over 30 topics) in the topic distribution in Figure 4.3 and several peaks in time trends. Topics 7 (nature), 8 (family), 20 (soul + body), and 26 (house) fulfil the criteria (See Figures 4.4 to 4.7 for the topics' time trends). We use the topics to discuss the first research question.

We observe similarities of the four topics, both statistically and semantically. Regarding the statistical characteristics of them, the highest peaks are not high

[1] Originally, The topics are labelled in Swedish. For example, the label of topic 21 is (kanalije dam). We describe the original labels in Appendix A

compared with other topics such as topic 18, which has a peak of 0.14. This finding indicates that writers do not intensively use many of the words in topics 7, 8, 20, or 26 simultaneously but do use the words in many contexts. Additionally, many books by many authors favour the topics in different periods in the corpus, resulting in several spikes in the trends. Specifically, 59 books from 1881 to 1941 by 27 authors use topic 7 more than 0.034 (above the average). Similarly, 36 books from 1886 to 1940 by 12 authors use topic 8, 58 books from 1885 to 1941 by 29 authors use topic 20, and 69 books from 1879 to 1941 by 27 authors use topic 26 more than 0.034; whereas ten books by three authors use the least popular topic (topic 14) more than 0.034.

For a semantic aspect of the topics, all four topics relate to people's lives. Topics 7 and 26 comprise words that mainly represent environments or backgrounds such as nature or furniture, and topics 8 and 20 comprise words that mainly represent people. The terms for describing nature, backgrounds, people can be used in many scenes independently of storylines. Hence, authors use them in many books resulting in the popularity of the four topics.

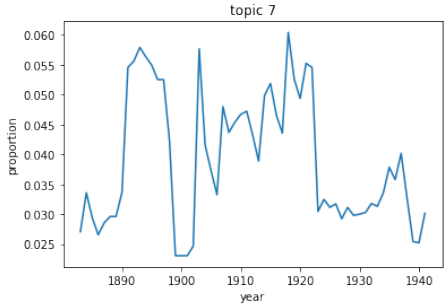


Figure 4.4.: Time trends of topic 7 (nature).

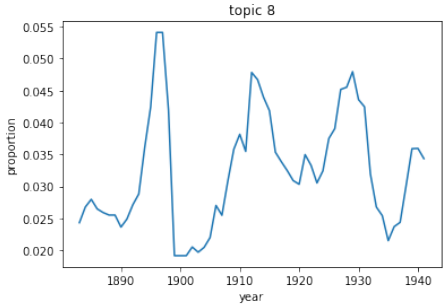


Figure 4.5.: Time trends of topic 8 (family).

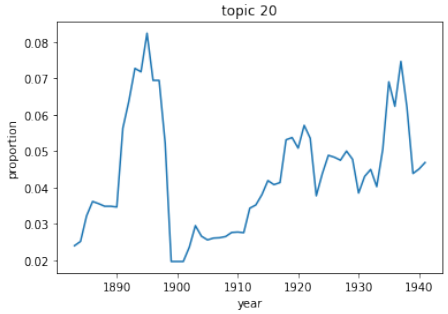


Figure 4.6.: Time trends of topic 20 (soul + body).

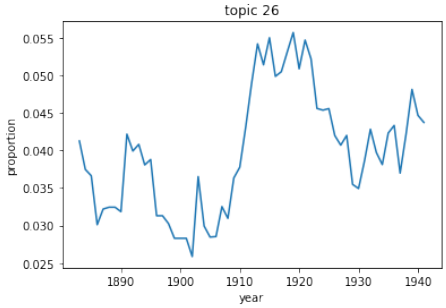


Figure 4.7.: Time trends of topic 26 (house).

Although we have examined topics above thresholds, some topics below the thresholds have a similar tendency. Topics 5 (town), 11 (book), 13 (master + village), 25 (Christianity), and 27 (family) have a distribution in the corpus lower than 0.034 but have several peaks and they are mainly for describing peoples' lives (See

Figures 4.8 to 4.12 for the topics' time trends.). The terms in these topics can also be used independently of storylines. Hence, it is understandable to have several peaks.

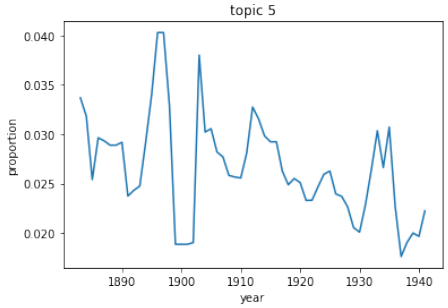


Figure 4.8.: Time trends of topic 5 (town).

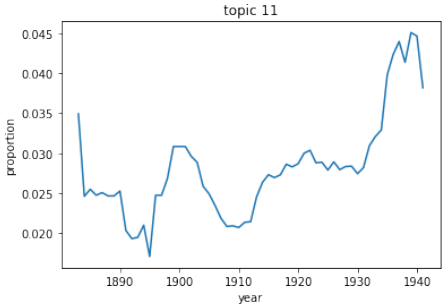


Figure 4.9.: Time trends of topic 11 (book).

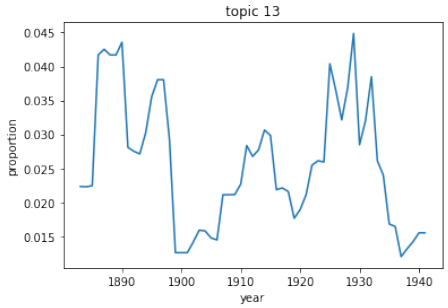


Figure 4.10.: Time trends of topic 13 (master + village).

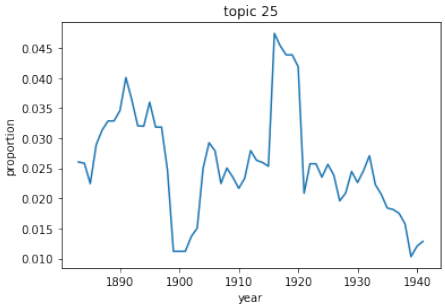


Figure 4.11.: Time trends of topic 25 (Christianity).

4.3.2. Time-specific Topics

To discuss time-specific topics, we use 11 topics with a spike in time trends (Figure 4.13 to 4.23). We focus topics with a peak and ignore topics with several peaks, such as topic 28 (Figure 4.24). We list books with a high proportion of the 11 topics in Table 4.4.

Topic id	Books
1	Nils Holgerssons underbara resa 1 (1907) by Lagerlöf (0.297) Nils Holgerssons underbara resa 2 (1907) by Lagerlöf (0.217) Torpare (1909) by Kämpe (0.046) Trälar (1907) by Kämpe (0.043) En saga om en saga och andra sagor (1908) by Lagerlöf (0.035)
2	Dagbok. Mårbacka III (1932) by Lagerlöf (0.267) Ett barns memoarer. Mårbacka II (1930) by Lagerlöf (0.114)

Continue to the next page

Topic id	Books
	Uppgörelser (1934) by Boye (0.042)
4	Samlade Verk 54 (1905) by Strindberg (0.248) Samlade Verk 56 (1906) by Strindberg (0.104) Svenskarna och deras hövdingar I (1908) by Heidenstam (0.040) En saga om en saga och andra sagor (1908) by Lagerlöf (0.038)
6	Man bygger ett hus (1938) by Värnlund (0.238) För lite (1936) by Boye (0.095) Kallocain (1940) by Boye (0.084) Ur funktion (1940) by Boye (0.082)
9	Blå spåret (1916) by Regis (0.476) Doktor Smirnos dagbok (1917) by Duse (0.412) Falska papper (1916) by Bergman (0.145) Spöksekretären (1919) by Berger (0.085) Guds vackra värld. II (1916) by Koch (0.035)
10	Gösta Berlings saga (1891) by Lagerlöf (0.130) Osynliga länkar (1894) by Lagerlöf (0.051)
12	Svenskarna och deras hövdingar II (1911) by Heidenstam (0.339) Svenskarna och deras hövdingar I (1908) by Heidenstam (0.158) Rustgården (1910) by Hansson (0.040) En saga om en saga och andra sagor (1908) by Lagerlöf (0.035)
16	Herr von Hancken (1920) by Bergman (0.178) Kvartetten, som sprängdes II (1924) by Sjöberg (0.082) Kvartetten, som sprängdes I (1924) by Sjöberg (0.070) Jag, Ljung och Medardus (1923) by Bergman (0.068) Farmor och Vår Herre(1921) by Bergman (0.061) Eros' begravning (1922) by Bergman (0.048) Chefen fru Ingeborg (1924) by Bergman (0.044)
18	Klostret (1898) by Strindberg (0.177) Fagervik och Skamsund (1902) Strindberg (0.124)
21	Lotten Brenners ferier (1928) by Bergman (0.369) Jonas och Helen (1926) by Bergman (0.208) Clownen Jac (1930) by Bergman (0.204) Kerrmans i Paradiset (1927) by Bergman (0.190) Mina dagböcker (1926) by Malling (0.092) Mannen som reste sig (1927) by Sandel (0.072) Vårdrömmar. Berättelser för ungdom (1927) by Beskow (0.062) Resan till Rom (1929) by Söderberg, H. (0.051) Den främmande staden (1928) by Söderberg, M. (0.050)
24	Klostret (1898) by Strindberg (0.096) Fagervik och Skamsund (1902) by Strindberg (0.087) Silverträsket (1898) by Strindberg (0.051)

Continue to the next page

Topic id	Books

End of the table

Table 4.4.: Books with high topic distribution for the topics we discuss. We chose books with more than 0.034 distribution of the topic. Parentheses following book titles denote the year the book was published, and parentheses following authors' names denote the topic distribution in the book.

We observe two findings from Table 4.4. The first finding is that when an author writes several books in a short time, these books tend to contribute to peaks in time trends. For instance, two books by Strindberg, *Klostret* and *Fagervik och Skamsund*, create peaks in topics 18 and 24. These peaks might occur because the vocabulary (word type) in the books are similar because they are written in a short period by a specific author. Her or his vocabulary does not change much over the short amount of time, resulting in the use of the same patterns, which LDA groups as one topic even if these words are not similar semantically. Specifically, topic 24 has a peak in 1900, although the topic has several semantic groups in it. The peak implies that August Strindberg used the terms in topic 24 to write the two books at that time much more than his earlier or later works in the corpus.

The other finding is that some of the topics were simultaneously used by several authors. For instance, topic 9 was used much by five authors from 1916 to 1919. It may be just a coincidence. Regis, Duse, Bergman, Berger, and Koch used many terms in topic 9 in the same period by accident because the topic is for the crime, which is a usual theme for writing books. Note that, the five books that use topic 9 much may have similar words in common, but it does not directly indicate that the theme is similar. Because such a group of books can convey entirely different ideas, close reading must be performed to evaluate them.

Additionally, other works or social events could affect topic trends. Books published around 1910 might encourage the aforementioned 5 authors to write books by using the words in topic 9, or they wanted to write about crime in the state of the society at that time. It requires close reading or methods other than topic models to examine the relations among works or between social situation and books. Hence, we cannot discuss the topic trends from this aspect any further.

In this section, we have discussed topic trends. For popular topics, the authors use topics 7, 8, 20,26 frequently in the corpus. More specifically, words or sentences for expressing everyday life are abundant in the corpus. This finding shows that topics related to normal life are standard tools in books. The characters in many books may live with their family in a house surrounded by nature. The finding may be intuitively self-evident but has been reached without reading any books. In other words, we support the intuition by using LDA.

For time-specific topics, we have discussed 11 topics which have a sudden peak in the time trends. The peaks result from two reasons: first, when an author writes several books in a short period, the time trends of a topic increase because the vocabulary used is similar, and second, when several authors write books using the same theme (e.g., crime) in a short time, sudden peak in the time trends of

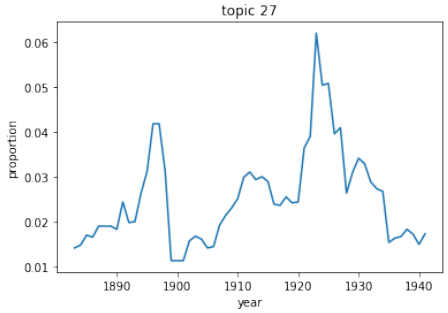


Figure 4.12.: Time trends of topic 27 (family).

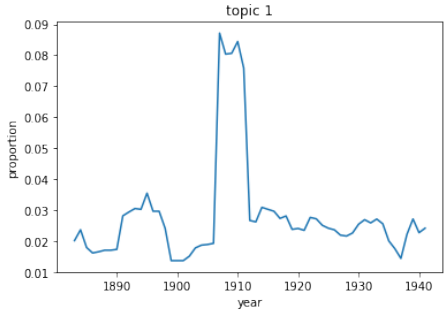


Figure 4.13.: Time Trends of Topic 1 (animal + nature + person)

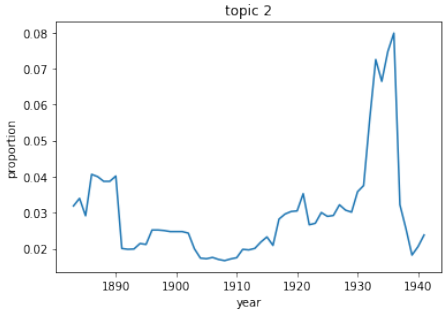


Figure 4.14.: Time Trends of Topic 2 (relative + time)

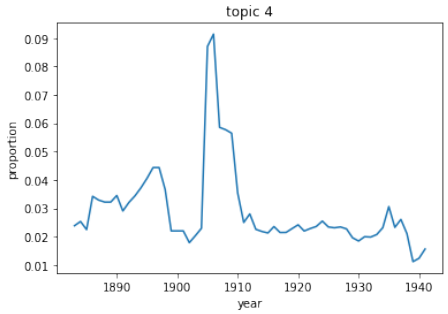


Figure 4.15.: Time Trends of Topic 4 (temple + ruler)

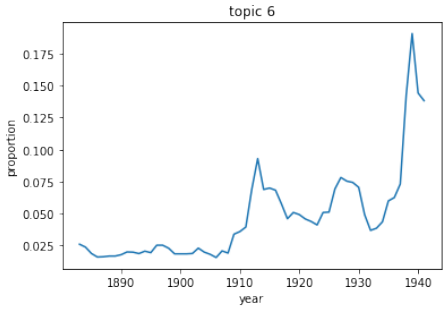


Figure 4.16.: Time Trends of Topic 6 (work + person)

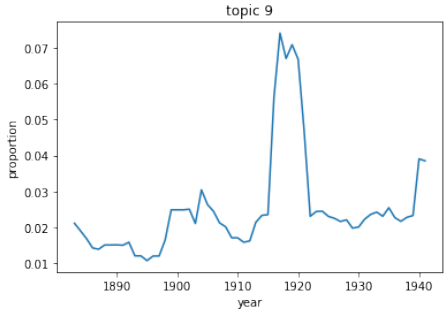


Figure 4.17.: Time Trends of Topic 9 (crime + town)

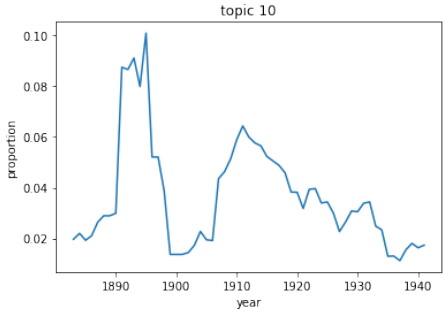


Figure 4.18.: Time Trends of Topic 10 (farm + person)

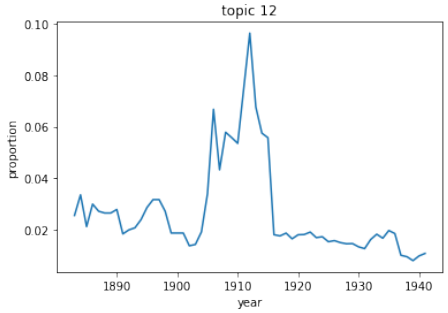


Figure 4.19.: Time Trends of Topic 12 (king + war)

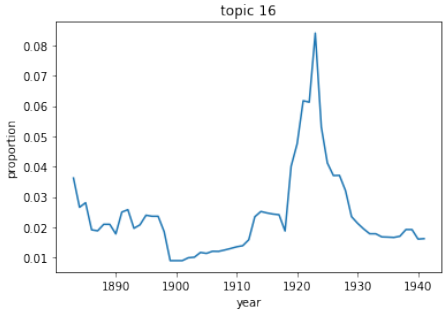


Figure 4.20.: Time Trends of Topic 16 (title + office)

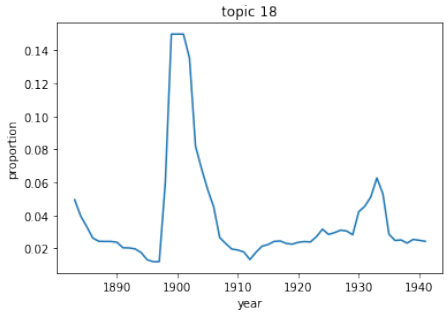


Figure 4.21.: Time Trends of Topic 18 (art + guest)

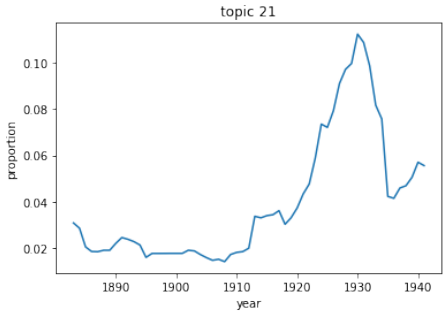


Figure 4.22.: Time Trends of Topic 21 (girl + party + occurrence)

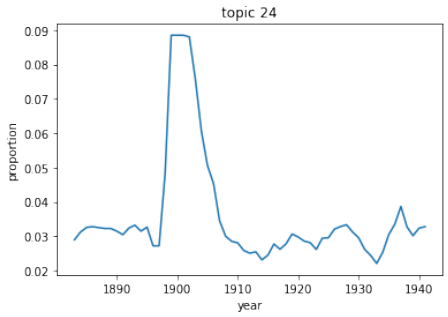


Figure 4.23.: Time Trends of Topic 24 (feeling + punishment)

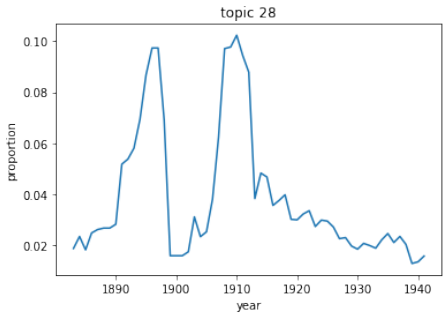


Figure 4.24.: Time Trends of Topic 28 (forest + village). They are not discussed here because two peaks are observed.

the topic is observed. We cannot discuss the reason why the authors wrote books using the same theme in the same period by using topic models.

4.4. Experiment 2: Author Topic Preference

In this section, we discuss the characteristics of several authors. To mitigate the bias to a small amount of data, we discuss nine authors who have more than three books in the corpus; Andersson, Bergman, Boye, Koch, Kämpe, Lagerlöf, Sandel, Strindberg and Värnlund. We compare the topic distribution of the nine authors to the topic distribution of the entire corpus and discuss which topics characterise each author compared with the entire corpus. We use the characteristic topics for each author in Table 4.5. We use both topics with high proportions and low proportions to demonstrate the topics an author uses frequently and infrequently.

Table 4.5 shows that topic 29 (shop + money) is a characteristic topic for Andersson. He uses the words in the topic often in his books in the corpus. Similarly, eight other authors have at least a topic they use frequently. Bergman uses topics 16 (title + office), 21 (girl + party + occurrence), and 27 (family); Boye and Koch prefer topic 22 (thought); Kämpe uses topics 6 (work + person) and 10 (farm + person); Lagerlöf uses topics 1 (animal + nature + person), 10, and 13 (master + village); Sandel uses topics 6 and 8 (family); Strindberg uses topics 15 (society + school) and 19 (impression + mankind), and Värnlund uses topic 6.

By contrast, some topics are rarely used by the authors compared with the distribution of the entire corpus. Andersson infrequently uses topics 3, 16 and 21; Bergman seldom uses topics 0, 7 and 28; Boye seldom uses topics 0, 10 and 12; Koch seldom uses topics 2, 3 and 21; Kämpe seldom uses topics 3 and 6; Lagerlöf seldom uses topics 6, 9, 15, 18, 19 and 21; Sandel seldom uses topics 7, 12 and 28; Strindberg seldom uses topics 6 and 21; and Värnlund seldom uses topics 0, 2, 10, 12, 13, 25 and 28.

Notably, the characteristics do not show that a specific author uses those topics rarely or frequently in all of their books. That is, Andersson may use many words

Author	Topic id
Andersson, Dan	3, 16, 21, 29*
Bergman, Hjalmar	0, 7, 16*, 21*, 27*, 28
Boye, Karin	0, 10, 12, 22*
Koch, Martin	2, 3, 21, 22*
Kämpe, Alfred	3, 6*, 10*,12
Lagerlöf, Selma	1*, 6, 9, 10*, 13*, 15, 18, 19, 21
Sandel, Maria	6*, 7, 8*, 12, 28
Strindberg, August	6, 15*, 19*, 21
Värnlund, Rudolf	0, 2, 6*, 10, 12, 13, 25, 28

Table 4.5.: Nine authors and their corresponding characteristic topics. A topic is characteristic if it is more than twice or less than half in proportion compared with the topic distribution of the corpus. Topic ids with asterisks denote the topics with more than twice the distribution of the whole corpus. Topics without asterisks denote topics with less than half the proportion of the topic distribution of the entire corpus. The comparisons of the topic distribution of the authors and the whole corpus in bar charts are in Appendix C.

in topic 29 in one book but then write another book that uses words in topic 3, even though he has rarely written on the latter topic.

The characteristics of an author are defined among other authors. We select specific topics on the basis of a comparison of the authors' books in the corpus and the topic distribution of the entire corpus. Thus, the result should be interpreted such that topic 29 is a particular topic for Andersson as long as his book in the corpus is compared with the whole corpus. More specifically, Andersson uses the word types from topic 29 more than the average of the corpus. If we compare two authors' topic distributions, the characteristics will differ from what we have now. For instance, the characteristic topics of Boye are topics 1, 2*, 10, 11*, 21*, and 29, and differ from what we present in Table 4.5 when we compare her topic distribution to that of Koch. In this case, we interpret the result as she uses topics 2, 11 and 21 more than Koch. Additionally, the books we use to compute the topic distribution of each author are the books in the corpus, and many other publications are not in the corpus. Characteristic topics might differ if books not listed in the corpus are used to build models.

We can also discuss the characteristics of authors from the semantic point of view of topics. For instance, Strindberg writes on society, impression and humankind (topics 15 and 19) much. His topic preference seems to correspond to the description of Encyclopedia Britannica, that says he "combined psychology and Naturalism in a new kind of European drama that evolved into Expressionist drama." For the other eight authors, we can also interpret their preference as characteristic themes such that Boye writes on thought (topic 22) often in her books. However, it requires careful discussion by methods other than distant reading such as close reading to interpret the topics which are built on the basis of statistics as semantic themes. Hence, it is beyond the scope of our research.

In this section, we have observed the particular topics for several authors on

the basis of the comparison to the corpus. All authors have at least one topic they prefer and one topic they seldom use. The distribution shows the average of how much an author uses the topics per book. The topic distribution is representative, and each book must be examined to discuss their content in detail.

5. Conclusion

5.1. Conclusion

We apply LDA to a corpus of Swedish literature comprising 118 Swedish novels and prose from 1879 to 1941 to capture time trends in Swedish literature and to find characteristics of authors. Topic models such as LDA provide a means of distant reading. Although we do not read any Swedish literature closely, we observe time trends of topics by using our model. We also observe the characteristics of the authors in comparison to the entire corpus.

We build the LDA model on Swedish literature with 30 topics by using Gensim, a Python library for topic models. We compute the perplexity and the coherence to decide the number of topics of our model, and we implement χ^2 (chi-square) test to choose representative words for each topic and the zero-order relevance method to label the topics automatically. We also label the topics manually. The difficulty of labelling follows the coherence score roughly. However, some topics with low coherence score have semantically coherent words enough to label them both manually and automatically.

We discuss the time trends of topics in the first experiment. Four topics for daily life (i.e., nature, family, emotion, and house) are frequently observed in the entire corpus because authors can use the topics to describe scenes or environments independent of storylines. By contrast, eleven topics peak in a period because of two reasons. One reason is that several authors write stories by using the word types in the topics such as topic 9 (crime). The other reason is that an author writes several books in a short time by using a specific topic heavily, such as topic 24 (friend). An author may use similar patterns consciously or unconsciously in her or his books in a short period, which results in a peak of a topic.

The second experiment aims to determine the characteristics of authors by using topic models. Nine authors we observed have at least one topic that they use more than the average of the corpus, and at least one topic that they rarely use compared with the entire corpus. These topics are statistical characteristics of the authors, although close reading may be required to discuss the characteristics semantically.

Methods for distant reading, such as LDA enable us to discuss literature more objectively based on the results than we analyse books closely. However, our subjectivity can enter our experiments in some steps. For example, the model parameter relies on our decision, and the manual labelling requires our evaluation. Besides, the analysis of the results is also a subjective activity. In other words, the way we build models and the way we discuss the results rely on our subjectivity. That said, our model supports our discussion mathematically and gives the perspective which we can never obtain by close reading. We propose further research relating to topic modelling in the next section.

5.2. Further Research

The model we use can be improved because the parameter space we searched for was limited. Changing the hyperparameters α and η affects the inference directly. Preprocessing methods also affect the model. We carry out typical preprocessing methods, namely tokenisation, lemmatisation, stopword removal and lowercase conversion, and we remove part-of-speeches other than nouns. As Uysal and Günal (2014) discussed, appropriate combinations of preprocessing methods improve outcomes of downstream tasks.

The representation of topics is another research field, especially methods for automatic labelling. Manual labelling is subjective and a time-consuming task, and it is impossible to reproduce the result. Therefore, automatic labelling methods will improve the usability of topic models. Most of the automatic labelling methods require an external corpus to generate labels. Methods which use the same data as the topic model is built on is preferable for researchers who do not have a large amount of data like us.

Researchers can apply topic models in a different manner in literary analysis. First, researchers could relate the time trends of topics to events, for example, an author's experience of a severe social event (e.g. war) may affect their emotions such that they write on themes that they have never written on. Therefore, practitioners can observe the effect of a catastrophe on literature by comparing time trends before and after the event. Magnusson et al. (2018) observed changes in time trends for the mention of immigration in Swedish Parliamentary debates by using topic models on speeches by the parliament. Similar methods may be applicable to topics in literature. Our corpus comprises documents from 1879 to 1941. Footprints of social changes or important events in Sweden can be observed if a model is built appropriately for the target event.

Second, researchers could use topic models to compare books. They can interpret the topic distribution of each book as a vector of the book. In other words, LDA is a method for dimensionality reduction. We convert the documents in our corpus into 30-dimensional vectors in our experiments, and the vectors facilitate the mathematical comparison of the authors' similarity. Notably, the computation of similarity based on mathematics would differ from similarity based on meaning. Therefore, further research could verify whether books or authors with a similar topic distribution are similar semantically.

Topic models are also applicable in research to determine the nationality of a corpus, as discussed in Jockers (2013). Topic 25 in our model is about Christianity. If we build a topic model on literature from other countries where Christianity is not commonly practised, topics about Christianity might not be observed. Hence, a comparison of corpora from different countries would provide a mathematical perspective on the nationality of literature.

A. Topic Description

ID	Label	30 keywords
0	Water båt vatten (boat water) 0.6	båt, strand, vatten, sjö, ö, prinsessa, brygga, skär, fiskare, segel, fisk, land, vik, udde, ångbåt, hamn, holme, hav, fjärd, åra, eka, duva, sjöman, däck, lots, fiske, kurs, fartyg, veranda, skärgård (boat, shore, water, lake, island, princess, jetty, rocky islet, fisherman, sail, fish, shore, bay, cape, steamboat, port, islet, sea, fjord, oar, rowing-boat, pigeon, sailor, deck, pilot, fishing, course, vessel, veranda, archipelago)
1	Animal + Nature + Person pojke par (boy pair) 0.34	pojke, hund, djur, fågel, vinge, unge, damm, räv, å, gås, slätt, ägg, håll, mark, järn, ben, älv, kråka, råtta, människa, land, dal, åker, bo, korp, uggla, bom, svans, trakt, får (boy, dog, animal, bird, wing, young bird, pond, fox, stream, goose, plain, egg, direction, ground, iron, leg, river, crow, rat, person, shore, valley, field, nest, raven, owl, bar, tail, district, sheep)
2	Relative + Time morbror mamma (uncle mother) 0.34	fröken, brev, tant, moster, morbror, farbror, student, faster, lördag, släkting, dagbok, mod, skull, vänskap, eftermiddag, kort, morgon, augusti, resa, maj, förmiddag, måndag, tillgivenhet, hälsning, fredag, skrivelse, tisdag, anförvant, kväll, morse (unmarried woman, letter, aunt, aunt, uncle, uncle, student, aunt, Saturday, relative, diary, courage, sake, friendship, afternoon, card, morning, August, journey, May, morning, Monday, attachment, greeting, Friday, official letter, Tuesday, relation, evening, morning)

Continue to the next page

ID	Label	30 keywords
3	Family + House mamma pappa (mother father) 0.42	fru, mamma, pappa, löjtnant, kök, mamsell, hushållerska, mormor, råd, dörrn, klänning, piga, frukost, tallrik, strumpa, lakan, födelsedag, smörgås, sal, kaffe, sängkammare, förmak, matsal, svärfar, lärarinna, brud, svärmor, soffa, pall, piano (married woman, mom, papa, lieutenant, kitchen, a title for middle class women, housekeeper, grandmother, advice, door, dress, maid, breakfast, plate, sock, sheet, birthday, open sandwich, dining room, coffee, bedroom, lounge, dining room, father-in-law, teacher, bride, mothe-in-law, sofa, stool, piano)
4	Temple + Ruler tempel stad (temple town) 0.44	gud, grav, jord, kors, tempel, död, frid, kejsare, synd, kloster, munk, folk, altare, kapell, mästare, helgon, påve, rike, åsna, vin, julafton, högmod, pest, hedning, helvete, härskare, himmel, palats, mur, syndare (god, grave, earth, cross, temple, death, peace, emperor, sin, monastery, monk, people, altar, chapel, master, saint, pope, state, donkey, wine, Christmas Eve, pride, plague, heathen, hell, ruler, sky, palace, wall, sinner)
5	Town stad folk (town people) 0.33	stad, gata, herre, hus, port, hatt, namn, torg, kyrkogård, kista, major, bostad, vakt, kläder, klang, kanna, borgmästare, köpman, tåg, bänk, trottoar, gesäll, smuts, portgång, hop, engelsman, folkhop, vandring, mängd, krog (town, street, gentleman, house, front door, hat, name, square, cemetery, coffin, major, house, guard, clothes, ring, pot, mayor, merchant, train, bench, pavement, craft worker, dirt, gateway, crowd, Englishman, group, wandering, quantity, restaurant)

Continue to the next page

ID	Label	30 keywords
6	Work + Person skomakare arbete (shoemaker work) 0.33	blick, arbete, kamrat, ansikte, kväll, morsa, uttryck, tag, leende, rörelse, fabrik, sekund, medvetande, arbetare, plats, bygge, läpp, öre, axel, glimt, grabb, tonfall, förnimmelse, stämma, slut, smula, sko, kyla, skomakare, drag (look, work, comrade, face, evening, mamma, expression, hold, smile, motion, factory, second, consciousness, worker, place, building under construction, lip, penny, shoulder, gleam, boy, intonation, feeling, meeting, end,bit, shoe, cold, shoemaker, pull)
7	Nature våg hav (wave sea) 0.52	sol, träd, luft, berg, himmel, sten, vind, storm, moln, hav, regn, våg, jord, dimma, snö, stjärna, vatten, rök, gren, gräs, sommar, sky, mörker, väder, löv, vår, buske, höjd, måne, ro (sun, tree, air, mountain, sky, stone, wind, storm, cloud, sea, rain, wave, earth, fog, snow, star, water, smoke, branch, grass, summer, cloud, dark, weather, leaf, spring, bush, height, moon, rest)
8	Family dotter son (daughter son) 0.41	son, moder, fader, dotter, syster, hem, gosse, broder, flicka, förälder, bror, tår, barn, syskon, familj, hjälp, vård, nyhet, svåger, bud, änka, begravning, vrede, snille, gråt, löfte, sjukhus, utväg, lock, korg (son, mother, father, daughter, sister, home, boy, brother, girl, parent, brother, tear, child, brother/sister, family, help, care, news, brother-in-law, bid, widow, burial, wrath, genius, crying, promise, hospital, means, lid, basket)
9	Crime + Town kommissarie doktor (inspector doctor) 0.48	doktor, polis, domare, boll, bil, kommissarie, mord, mördare, telefon, konstapel, vittne, spår, betjänt, dur, villa, journalist, revolver, detektiv, konsul, undersökning, madame, monsieur, brottsling, polisman, papper, skott, arbetsrum, advokat, bibliotek, adress (doctor, police, judge, ball, car, chief inspector, murder, murderer, telephone, police, witness, mark, manservant, major (music), house, journalist, revolver, detective, consul, examination, madam, gentleman, criminal, police, paper, shot, workroom, lawyer, library, adress)

Continue to the next page

ID	Label	30 keywords
10	Farm + Person kärra häst (cart horse) 0.41	gård, karl, väg, stuga, häst, dräng, ko, backe, bonde, grevinna, kavaljer, patron, släde, kärra, bruk, landsväg, folk, husbonde, stall, käring, ved, gruva, äpple, varg, socken, spis, torpare, kreatur, torp, marknad (yard, man, road, small house, horse, farmhand, cow, hill, farmer, countess, cavalier, refill, sleigh, cart, usage, main road, people, gentleman, stall, wife, wood, mine, apple, wolf, parish, cooker, crofter (farmer), cattle, crofter's holding, market)
11	Book lykta ljus (lantern light) 0.30	bok, slag, svar, papper, bild, tystnad, skrivbord, lykta, stol, penna, fråga, figur, hylla, bokstav, plan, te, slut, porträtt, förhoppning, linje, hälft, spöke, anda, blad, tecken, samtal, rad, siffra, gåta, promenad (book, type, answer, paper, picture, silence, desk, lantern, chair, pen, question, figure, shelf, letter, plan, tea, end, portrait, hope, line, half, ghost, breath, leaf, sign, conversation, row, figure, riddle, walk)
12	King + War konung land (king country) 0.58	kung, konung, slott, prins, svensk, soldat, general, hertig, mäster, krig, knekt, krigare, rike, ryttare, hov, ärkebiskop, fogde, dansk, kanon, värja, här, torn, fana, vall, officer, fänrik, nunna, turk, ryss, kardinal (king, king, palace, prince, Swede, soldier, general, duke, master, war, jack, warrior, state, rider, hoof, archbishop, public official, Danish, cannon, rapier, army, tower, flag, bank, officer, a title for officer, nun, Turk, Russian, cardinal)

Continue to the next page

ID	Label	30 keywords
13	Master + Village magister natt (master night) 0.34	magister, jungfru, vagn, trädgård, brukspatron, prost, prostinnan, kusk, prästfru, förvaltare, ingenjör, lov, prostgård, åkdon, allé, gästgivargård, friherrinnan, trädgårdsmästare, bock, kyrkby, grind, tjuv, väv, adjunkt, stalldräng, förstubro, lusthus, besked, sängkläder, spinnrock (master, virgin, carriage, garden, owner of a factory, dean, wife of dean, coast, wife of priest, administer, engineer, holiday, yard (of a dean), vehicle, avenue, inn, wife of friherr (a title), master of a garden, horse, a village, gate, thief, web, teacher, worker for stall, staircase, gazebo, answer, bedclothes, spinning wheel)
14	Authorisation greve herre (count gentleman) 0.36	greve, ära, majestät, ers, drottning, stat, regering, fiende, talare, förbund, närvaro, namn, nation, armé, fosterland, tal, beskydd, vän, narr, frånvaro, assessor, fädernesland, go, fullmakt, uppträde, maka, häxa, gudinna, monark (count, honor, majesty, your, queen, state, government, enemy, speaker, alliance, presence, name, nation, army, native country, number, protection, friend, fool, absence, deputy judge, native country, power, authorization, scene, wife, witch, goddess, monarch)
15	Society + School klass skola (class school) 0.46	tidning, samhälle, författare, professor, intresse, utveckling, klass, bolag, litteratur, överklass, redaktör, parti, kapital, ämbetsman, skola, kritik, bildning, million, individ, skald, lärare, omdöme, bana, termin, system, filosofi, universitet, underklass, åsikt, ideal (newspaper, society, author, professor, interest, development, class, company, literature, upper class, editor, part, capital, public official, school, criticism, education, million, individual, poet, teacher, opinion, path, term, system, philosophy, university, lower class, view, ideal)

Continue to the next page

ID	Label	30 keywords
16	Title + Office herr herre (Mr. gentleman) 0.40	herr, kapten, överste, rektor, häradshövding, landshövding, lektor, aktie, kupa, ekipage, krubba, kabinett, portfölj, herre, pincené, börs, trupp, kontor, skrubb, juni, alm, styrelse, kyckling, fönstersmyg, gäspning, tabell, hedersman, karamell, hyllning (gentleman, captain, colonel, headmaster, chief of district, chief of country, lecturer, share, globe, horse and carriage, manger, cabinet, briefcase, gentleman, glasses, purse, troop, office, boxroom, June, elm, committee, chicken, window (embrasure), yawn, table, man of honor, sweet, applause)
17	Marriage hustru vän (wife friend) 0.41	kvinna, barn, hustru, kärlek, äktenskap, frun-timmer, kön, skilsmässa, väninna, make, egendom, älskare, älskarinna, förhållande, läkare, svartsjuka, boja, amma, äganderätt, maka, trohet, sällhet, hona, vigsel, aktning, saknad, arvinge, usling, gift, svägerska (woman, child, wife, love, marriage, female, sex, divorce, girlfriend, husband, property, lover, mistress, state of things, doctor, jealousy, fetter, breast-feed, ownership, wife, fidelity, peace, female, marriage, respect, loss, heir, a bad guy, married , sister-in-law)
18	Art + Guest publik herre (audience gentleman) 0.49	gäst, dam, sällskap, scen, konst, publik, hotell, salong, direktör, värdinna, teater, afton, sekreterare, roll, tavla, pensionat, sir, situation, telegram, krets, gentleman, skådespelare, konstnär, lokal, målare, apa, kypare, arkitekt, kafé, värd (guest, lady, company, stage, art, audience, hotell, exhibition, director, hostess, theatre, evening, secretary, part, picture, boarding house, sir, situation, telegram, circle, gentleman, actor, artist, room, painter, ape, waiter, architect, café, host)

Continue to the next page

ID	Label	30 keywords
19	Impression + Mankind underverk luft (miracle air) 0.34	natur, intryck, mänsklighet, fruktan, brist, drift, religion, form, umgänge, underverk, art, släkte, kultur, beröring, last, hjärna, förbindelse, fiende, personlighet, växt, patient, tankarne, metall, modren, tro, inflytande, näring, landskap, skapelse, vilde (nature, impression, mankind, fear, lack, instinct, religion, form, relation, miracle, species, generation, culture, contact, vice, brain, connection, enemy, personality, growth, patient, thoughts, metal, mother, belief, influence, nourishment, province, creation, savage)
20	Soul + Body liv värld (life world) 0.35	själ, hjärta, dröm, liv, lycka, blomma, öga, väsen, glädje, kärlek, sorg, längtan, kropp, blod, död, smärta, doft, kraft, sång, ros, ande, varelse, ångest, ljus, verklighet, gestalt, glans, anlete, syn, skönhet (soul, heart, dream, life, happiness, flower, eye, essence, pleasure, love, sorrow, longing, body, blood, death, pain, scent, force, song, rose, mind, being, anxiety, light, reality, figure, shine, countenance, sight, beauty)
21	Girl + party + Occurrence kanalje dam (rogue lady) 0.32	flicka, fästman, rest, händelse, bankir, orsak, fest, chef, firma, betydelse, smula, yngling, nöje, mån, ring, karaktär, smak, likhet, ursäkt, förslag, omständighet, örfil, uppträdande, tillfälle, kanalje, värdighet, frack, paradis, kamrer, uppdrag (girl, fiancé, remainder, occurrence, banker, cause, party, head, firm, meaning, bit, youth, pleasure, extent, ring, character, taste, resemblance, excuse, proposal, circumstance, a thick ear, appearance, occasion, rogue, dignity, tail coat, paradise, accountant, commision)

Continue to the next page

ID	Label	30 keywords
22	Thought orätt rätt (wrong right) 0.35	människa, sätt, sak, fall, rätt, tanke, allvar, mening, vilja, ord, handling, del, känsla, orätt, möjlighet, värld, mål, ting, gräns, makt, ögonblick, visshet, skull, förnuft, samvete, grund, värde, säkerhet, merit, oro (man, way, thing, case, right, thought, seriousness, opinion, will, word, action, part, feeling, incorrect, possibility, world, goal, thing, boundary, power, moment, certainty, sake, reason, conscience, foundation, value, certainty, qualification, anxiety)
23	Time + World år sak (year thing) 0.29	tid, år, ungdom, minne, historia, sida, månad, början, ålder, tal, liv, dikt, avstånd, saga, del, vis, barndom, förändring, bi, vana, värld, område, längd, skillnad, ställe, diktare, ersättning, avseende, varning, medelpunkt (time, year, youth, memory, history, side, month, beginning, age, number, life, poem, distance, fairy tale, part, wise, childhood, change, bee, habit, world, territory, length, difference, place, writer, compensation, reference, warning, centre)
24	Feeling + Punishment öde vän (fate friend) 0.33	vän, öde, liv, hat, ensamhet, lidande, fiol, skuld, lugn, förtvivlan, fängelse, musik, hopp, elände, brott, fasa, morgon, natt, olycka, begär, medlidande, ånger, timme, sömn, straff, bitterhet, samvetskval, beslut, afton, våld (friend, fate, life, hatred, solitude, suffering, violin, debt, calm, despair, prison, music, jump, mistery, crime, horror, morning, night, misfortune, desire, pity, regret, hour, sleep, punishment, bitterness,remorse, decision, evening, violence)

Continue to the next page

ID	Label	30 keywords
25	Christianity präst kyrka (priest church) 0.44	präst, kyrka, pastor, församling, bibel, kyrkoherde, gärning, nåd, klockare, predikan, länsman, lära, söndag, folk, lag, gudstjänst, skola, tro, predikant, psalm, predikstol, nattvard, djävul, synd, rättfärdighet, apostel, möte, predikning, garn, orgel (priest, church, pastor, assembly, bible, rector, deed, be pardoned, sexton, sermon, police, faith, Sunday, people, law, worship, school, belief, preacher, psalm, pulpit, the Holy Communion, devil, sin, righteousness, apostle, meeting, sermon, yarn, organ)
26	House lampa ljus (lamp light) 0.43	dörr, rum, fönster, golv, vägg, säng, steg, trappa, huvud, hår, tak, ljud, arm, lampa, röst, ansikte, ruta, våning, gardin, ljus, mörker, hörn, stol, matta, bord, vrå, kudde, soffa, fot, spegel (door, room, window, floor, wall, bed, step, stairs, head, hair, roof, sound, arm, lamp, voice, face, screen, flat, curtain, light, darkness, corner, chair, carpet, table, corner, cushion, sofa, foot, mirror)
27	Family far son (father son) 0.39	far, mor, gumma, baron, farmor, farfar, kusin, släkt, onkel, dosa, katt, pojke, barnbarn, bur, vedbod, smörj, sedel, mage, sonson, huvut, betyg, snusdosa, misshandel, lärare, mognad, lår, skåra, pys, spökeri, väckarklocka (father, mother, old woman, baron, grandmother, grandfather, cousin, family, uncle, box, cat, boy, grandchild, cage, woodshed, beating, bill, stomach, grandson, head, certificate, snuffbox, maltreatment, teacher, maturity, leg, cut, little boy, ghost, alarm clock)
28	Forest + Village skog folk (forest people) 0.38	skog, eld, björn, by, stig, mark, troll, örn, orm, horn, träl, snår, yxa, spelman, borg, fjäll, bäck, kulle, kvist, guld, brudgum, stam, koja, älg, hövding, båge, svärd, hök, kämpe, spjut (forest, fire, bear, village, path, mark, troll, eagle, snake, horn, slave, brush, axe, musician, castle, mountain, brook, hill, twig, gold, bridegroom, stem, cabin, moose, chief, bow, sword, hawk, fighter, spear)

Continue to the next page

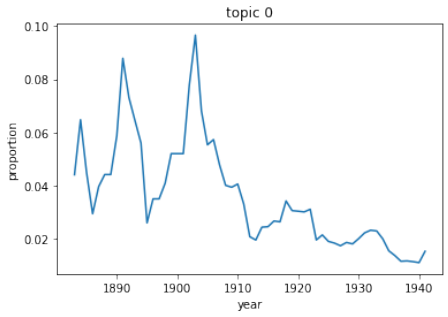
ID	Label	30 keywords
29	Shop + Money peng sak (money thing) 0.34	peng, gubbe, mat, glas, klocka, affär, krona, bord, pipa, middag, kaffe, öl, flaska, disk, bröd, kund, ficka, bank, potatis, svett, brännvin, summa, bit, sup, kniv, vin, tobak, pris, vecka, stryk (coin, old man, food, glass, clock, shop, Swedish krona, table, pipe, noon, coffee, beer, bottle, washing-up, bread, customer, pocket, bank, potato, sweat, snaps, sum, bit, snifter, knife, wine, tobacco, price, week, beating)

End of the table

Table A.1.: Detailed topic description and English translations. We added labels manually on the top rows of the label column and bigram labels automatically on the middle rows in Swedish and in English translation. The numbers on the bottom row of the label column are coherence score of each topic (0 < score < 1). Topics with higher scores are better topics in respect of the coherence.

B. Topic Trends over 5 Years

In this appendix, we represent topic trends of the topics we do not refer to in the thesis.



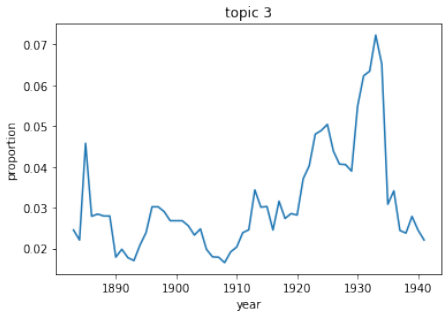
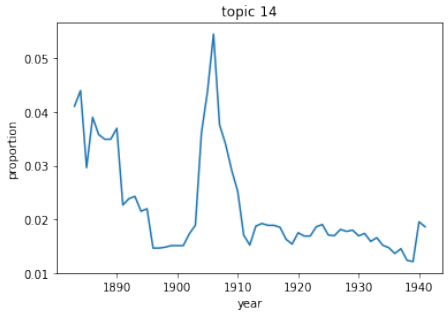


Figure B.1.: Time trends of topic 0 (water). **Figure B.2.:** Time trends of topic 3 (family).



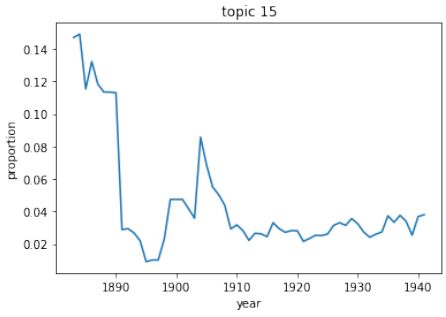


Figure B.3.: Time trends of topic 14 (authorisation). **Figure B.4.:** Time trends of topic 15 (society).

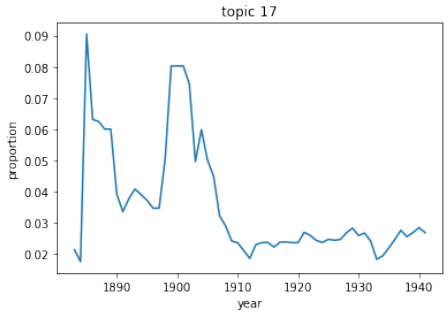


Figure B.5.: Time trends of topic 17 (marriage).

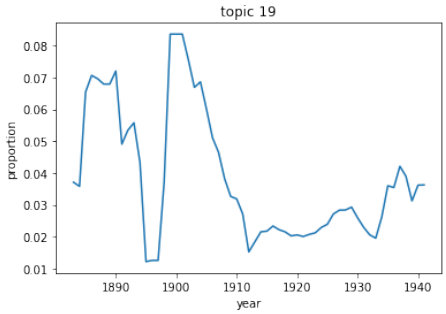


Figure B.6.: Time trends of topic 19 (nature).

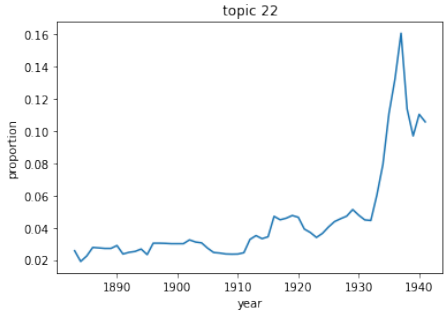


Figure B.7.: Time trends of topic 22 (thought).

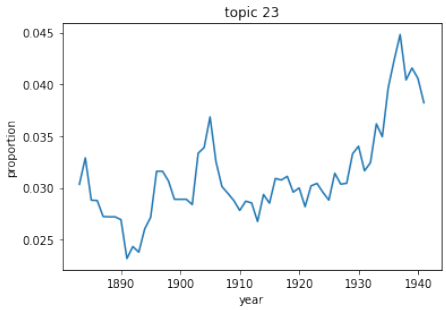


Figure B.8.: Time trends of topic 23 (time).

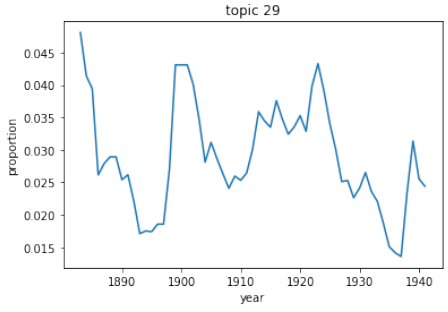


Figure B.9.: Time trends of topic 29 (restaurant).

C. The Comparison of Topic Distribution of Authors to That of the Entire Corpus.

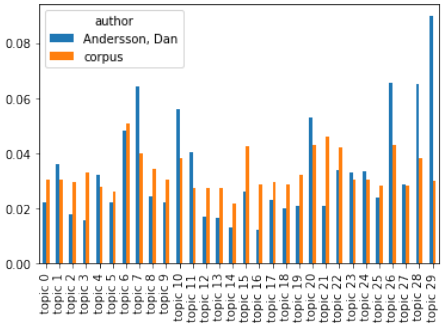


Figure C.1.: Andersson, Dan

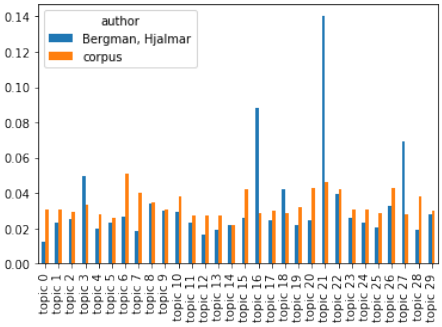


Figure C.2.: Bergman, Hjalmar

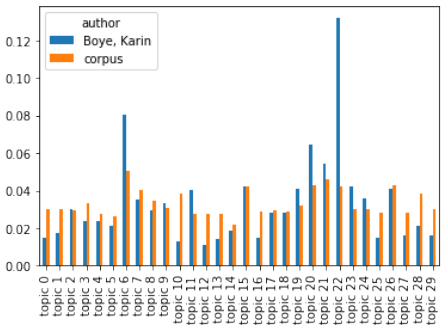


Figure C.3.: Boye, Karin

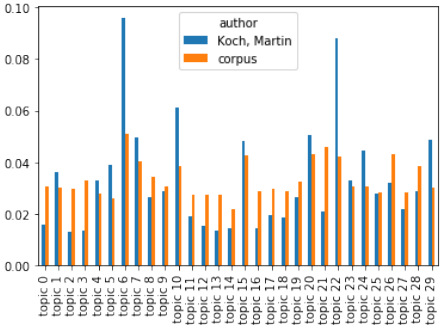


Figure C.4.: Koch, Martin

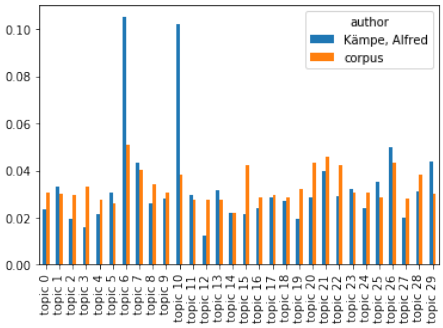


Figure C.5.: Kämpe, Alfred

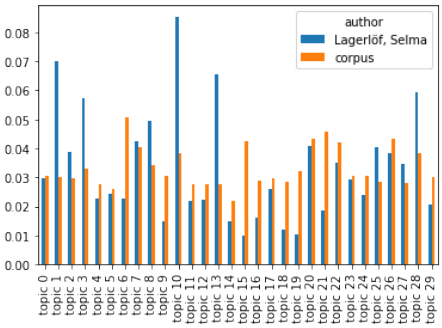


Figure C.6.: Lagerlöf, Selma

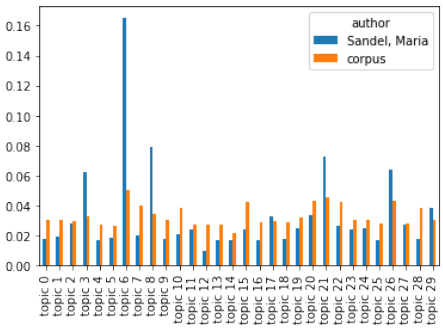


Figure C.7.: Sandel, Maria

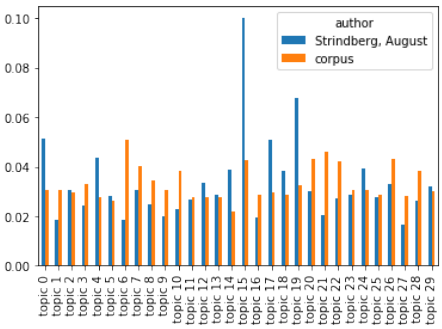


Figure C.8.: Strindberg, August

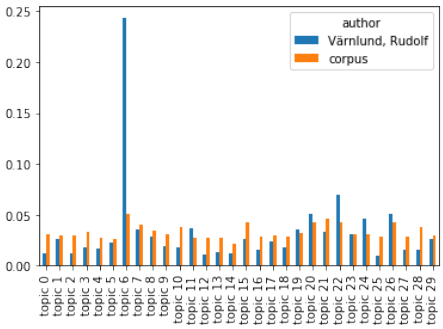


Figure C.9.: Värnlund, Rudolf

Bibliography

Barakat, A (2018). "What makes an (audio)book popular?" MA thesis. Linköping University, Department of Computer, Information Science, The Division of Statistics, and Machine Learning.

Bastani, Kaveh, Hamed Namavari, and Jeffrey Shaffer (2019). "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints". *Expert Systems with Applications* 127, pp. 256–271.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research* 3, pp. 993–1022.

Cano Basave, Amparo Elizabeth, Yulan He, and Ruifeng Xu (2014). "Automatic Labelling of Topic Models Learned from Twitter by Summarisation". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 618–624.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei (2009). "Reading Tea Leaves: How Humans Interpret Topic Models". In: *Advances in Neural Information Processing Systems 22*, pp. 288–296.

Chen, J., J. Yan, B. Zhang, Q. Yang, and Z. Chen (2006). "Diverse Topic Phrase Extraction through Latent Semantic Analysis". In: *Sixth International Conference on Data Mining*, pp. 834–838.

Dahllöf, Mats and Karl Berglund (2019). "Faces, Fights, and Families: Topic Modeling and Gendered Themes in Two Corpora of Swedish Prose Fiction". In: *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference*, pp. 92–111.

Griffiths, Thomas L. and Mark Steyvers (2004). "Finding scientific topics." *Proceedings of the National Academy of Sciences* 101 Suppl 1, pp. 5228–35.

Hoffman, Matthew, Francis R Bach, and David M Blei (2010). "Online learning for latent dirichlet allocation". In: *Advances in Neural Information Processing Systems*, pp. 856–864.

Jockers, Matthew L. (2013). *Macroanalysis. digital methods & literary history*. University of Illinois Press.

Lau, Jey Han, Karl Grieser, David Newman, and Timothy Baldwin (2011). "Automatic Labelling of Topic Models". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pp. 1536–1545.

Lau, Jey Han, David Newman, Sarvnaz Karimi, and Timothy Baldwin (2010). "Best Topic Word Selection for Topic Labelling". In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 605–613.

Magnusson, Måns, Richard Öhrvall, Katarina Barrling, and David Mimno (2018). *Voices from the far right: a text analysis of Swedish parliamentary debates*.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Maskeri, Girish, Santonu Sarkar, and Kenneth Heafield (2008). "Mining Business Topics in Source Code Using Latent Dirichlet Allocation". In: *Proceedings of the 1st India Software Engineering Conference*, pp. 113–120.

Mei, Qiaozhu, Xuehua Shen, and ChengXiang Zhai (2007). "Automatic Labeling of Multinomial Topic Models". In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 490–499.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 262–272.

Moretti, Franco (2000). "Conjectures on World Literature". *New Left Review* 1, pp. 54–68.

Navarro-Colorado, Borja (2018). "On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry". *Front. Digital Humanities*.

Norén, Fredrik and Pelle Snickars (2017). "Distant reading the history of Swedish film politics in 4500 governmental SOU reports". *Journal of Scandinavian Cinema*, pp. 155–175.

Östling, Robert (2018). "Part of Speech Tagging: Shallow or Deep Learning?" *The Northern European Journal of Language Technology*, pp. 1–15.

Ostrowski, D. A. (2015). "Using latent dirichlet allocation for topic modelling in twitter". *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing*, pp. 493–497.

Piantadosi, Steven T. (2014). "Zipf's word frequency law in natural language: A critical review and future directions". *Psychonomic Bulletin & Review* 21, pp. 1112–1130.

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*, pp. 399–408.

Roe, Glenn, Clovis Gladstone, and Robert Morrissey (2014). "Discourses and Disciplines in the Enlightenment: Topic Modeling the French Encyclopédie". *Front. Digital Humanities*.

Steyvers, Mark and Tom Griffiths (2007). "Probabilistic topic models". *Handbook of latent semantic analysis* 427, pp. 424–440.

Tangherlini, Timothy R. and Peter Leonard (2013). "Trawling in the Sea of the Great Unread: Sub-corpus topic modeling and Humanities research". In: pp. 725–749.

Uysal, Alper Kursat and Serkan Günal (2014). "The impact of preprocessing on text classification". *Information Processing & Management*, pp. 104–112.