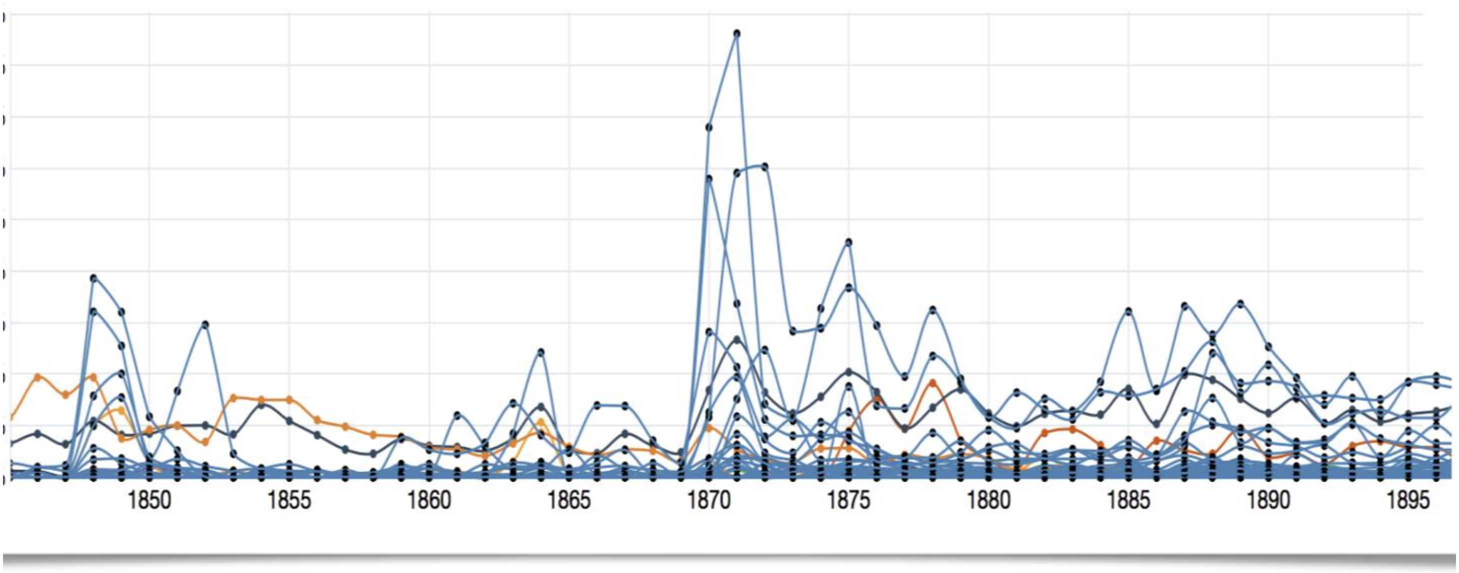
Tutorial IIb: Digital Text Analysis

RMA History Research Methods

GKRMV19015

2023/24 blok 2



Teacher

Dr. Pim Huijnen
Drift 6, room 0.22
p.huijnen@uu.nl

Utrecht University

Schedule

Tuesday 11:00 – 12:45, Drift 23, 1.07
Thursday 15:15 – 17:00, Drift 23, 107 (week 1) / DH Workspace Drift 27 (weeks 2-6)

Description

The purpose of this module is to familiarize students of the Research MA programme History with computational techniques to analyse textual data. More and more historians are using large sets (or corpora) of digital texts – both digitised and born-digital – as well as a growing variety of digital techniques to extract information from these texts. Scholars are discussing the disruptive nature of these increasingly common practices for historical scholarship under the header of Digital Humanities and Digital History. It is timely to become acquainted with a number of widely used techniques and skills, and to learn how to use them for the study of history in a critical and conscientious way. A second goal of this tutorial is to provide students with sufficient information to form their own opinion of the extent to which digital resources and tools change the heuristics of historical enquiry.

In this module we will work with the Python programming language in the Jupyter Notebook environment. Off-the-shelf software for text analysis does exist (AntConc, Voyant Tools, Sketch Engine), but working with Python gives you endlessly more flexibility. You are trained in the techniques we will use in this tutorial in a 'hands on' setting: you will work on your own research projects in groups of 2, based on a topic of your interest. You will start by gathering a dataset that represents your topic, after which you will try out different techniques to extract information from your data during the lab sessions. While these lab sessions are the place to work on your portfolio (see below), we will use the Tuesday seminar sessions to reflect on the implications of digitization for the study of history.

Assessment: Portfolio

In this tutorial, you will study a topic of your choice in groups of two. Collaboration is a core-element of digital history, since 'it is rare for a single individual (be it instructor or student) to have all of the technical and interpretive skills that are needed to reach a learning outcome' (Guiliano 2022). As a group, you will describe what you did in a research portfolio (ca. 5.000 words) to be handed in at the end of the tutorial (22 December 2023, midnight).

On data

There is no restriction on the topic that you study in this tutorial, but there are a few prerequisites for the **data** that you will use. Obviously, you should be able to argue how the data represents the topic of your interest. In addition to that, some formal requirements are:

- Your data should encompass some **element of comparison**. You might compare, for example, handbooks on psychiatry in 1900 with those from 2000. You might study diplomatic papers on the downfall of the Soviet Union from the Netherlands with those from Germany. Or you might take Thomas Mann's complete oeuvre to see how his wordings and topics change over time. These (random) examples show that the comparison may be over time as well as between domains or geographical regions, and the may be between single or multiple texts.

- You should be able to **download the data** as readable text files onto your own computers – ideally legally. The above examples might sound great, but might prove unfeasible because of lack of readily available digitized data (in the second example) and/or because of copyright restrictions (in the first and third example). Getting your hands on relevant or interesting data can be difficult, because it can be hidden, literally, anywhere on the webosphere. If you have no idea, you can find some suggestions at the end of this manual. As a last resort, newspaper articles always work well for the goals of this module (Nexis Uni, dating back to the mid-1990s).
- This leads us to the third requirement: getting a workable **amount of data**. A general rule of thumb is that digital text analysis has an added value first and foremost when you have more data than you can reasonably read for your research. What does that mean? If your data consists of books, two text files (representing two books) might already do the trick. If you plan to study newspapers, you should rather think in terms of hundreds or thousands of articles. Generally, digital text analysis works better the more data you have. In this respect, there is no upper limit. Your laptop or pc will disagree, however. Therefore, it is important that you balance between having sufficient data to yield robust results, but still small enough to not have to wait hours or days to generate them. Since this tutorial is about learning skills and methods rather than answering research questions, I advise you to start small. What works for small amounts of data will also work for larger amounts, so you can always go for it if time allows—although a minimum of 6 documents (texts, files) would be good to have.
- Lastly, what **format** should your data be in? The notebooks that we will use in class generally work with the plain text format (.txt) or the plain spreadsheet format (.csv). Online, you will not come across the latter very often. The former all the more. This is the standard format that most platforms that allow downloading digital or digitized data offer. If so, always choose .txt, because the data format is small and, again, the notebooks work with them, anyway. Pdf's of text or other text formats (like .doc or .rtf) are also viable formats, because we can convert these to .txt or .csv. If your research project involves comparison over time, the standard format to store your data locally is in **files-per-year** – ideally, with the year as filename. Alternatively, you might (also) put more metadata in the filename (like author or country). If you are familiar with spreadsheet formats, you are, of course, welcome to put all metadata in the column structure.

Form

You are free to choose the form of this portfolio, but it must contain the following elements (not necessarily in this order):

1) a description and justification of your topic (including some state-of-the-art) and the data that represents it;
2) the weekly assignments aimed at exploring and comparing your data and works/topics within it
3) a reflection on the results from the digital text analyses in terms of what this information adds to the existing literature, as well as a general reflection on the opportunities, but also the risks and pitfalls of quantitative text analysis for this type of historical research.

You will have noticed that this description speaks of ‘research topic’ rather than ‘research question’. Naturally, you are allowed to answer a research question in your paper, but this formulation is to signal that this is not a prerequisite. The aim of this setup is to offer room for experiment, in a reasoned manner, with different techniques.

Assessment

The portfolios are assessed equally on the doing (the quality of the work, specifically the level of application of the methods at hand) as well as on the reflection on what you have been doing (how suitable are data and methodology to study this research topic? What perspectives do they add to the existing scholarship? What do you gain but also lose in a digital environment?).

Notebooks

Are to be found on GitHub: https://github.com/PimHuijnen/QTA2023.

Weekly schedule

Week	Seminar	Lab	Portfolio
14-16 Nov	What is digital text analysis for historians?	Python introduction	Team, topic, data
21-23 Nov	Between statistics & hermeneutics	Preprocessing & counting	General statistics on the data
28-30 Nov	Text in numbers, text as language	Collocations, keywords, n-grams, POS, NER, tf-idf	Words, keywords and their relations
5-7 Dec	Data criticism in times of AI	Word embeddings	Semantic relations
12-14 Dec	Presentations	Presentations	
19-21 Dec		Deadline portfolio: **December 22, 2022, 11:59pm**	

Week 1 – Getting started: What is digital text analysis for historians?

Seminar (Tuesday 14 November)

In this first seminar we will zoom in on the particularities of quantitative text analysis. What sets computational techniques in historical research apart from conventional scholarship? (See, for example, Huijnen this week, Rockwell & Sinclair next week). As a way of introducing this, we will look at menus (http://menus.nypl.org/) as a source for both ‘traditional’ and digital history.

Readings:

- Huijnen, Pim, ‘Digital History and the Study of Modernity’, *International Journal for History, Culture and Modernity* 7 (2019): http://doi.org/10.18352/hcm.591.

- Lässig, Simone. “Digital History: Challenges and Opportunities for the Profession.” *Geschichte Und Gesellschaft* 47, no. 1 (June 10, 2021): 5–34. https://doi.org/10.13109/gege.2021.47.1.5.
- Nguyen, D. et.al., ‘How we do things with words: analyzing text as social and cultural data’, *Frontiers in Artificial Intelligence* 3 (2020), DOI: https://doi.org/10.3389/frai.2020.00062.

Further reading (not mandatory):

- Arguing with Digital History working group, ‘Digital History and Argument’, White Paper (Roy Rosenzweig Center for History and New Media, November 13, 2017).
- Fickers, Andreas, 'Towards A New Digital Historicism? Doing History In The Age Of Abundance', *VIEW Journal of European Television History and Culture* 1:1 (2012) 19–26.
- Hitchcock, Tim, ‘Confronting the Digital: Or How Academic History Writing Lost the Plot’, *Cultural and Social History* 10:1 (2013) 9–23.
- Jurafsky, Dan, Victor Chahuneau, Bryan R. Routledge, and Noah A. Smith, 'Linguistic Markers of Status in Food Culture: Bourdieu's Distinction in a Menu Corpus', *Journal of Cultural Analytics*, October 18 (2016).
- Madrigal, Alexis C. “The Way We Write History Has Changed.” The Atlantic, January 21, 2020. https://www.theatlantic.com/technology/archive/2020/01/smartphone-archives-history-photography/605284/.
- Nicholson, Bob ‘The Digital Turn’, *Media History* 19:1 (2013) 59–73.

Preparation:

Start thinking about a research topic and, if possible, a research question. To make your life easier, pick something that you have worked on in other classes. Try and find an appropriate dataset that represents this topic/question. Finding data is something that I can help you with in class, but you must make sure yourself not to wait for too long to decide what data you would like to use. Ideally, you have a dataset by next week. Start downloading!

Lab (Thursday 16 November) – Introduction to Python

In the first lab session we will start with the basics of Python programming. Before we can use Python code for text analysis, you will have to become familiar with the logic and most common commands of Python programming. You will do so by doing chapter 1 of ‘computational humanist’ Folgert Karsdorps Python Programming for the Humanities course. In 2021, Karsdorp has overhauled his course in the open access book Humanities Data Analysis, together with Mike Kestemont and Alan Lidell. This book, however, presupposes basic knowledge of Python. For the aims of this tutorial, therefore, his old course is better suited. You can find chapter 1 here (click on ‘download raw file’ on the right top) and open it in Jupyter Notebooks.

Preparation:

Please get yourself acquainted with Jupyter Notebooks before class. It is an interactive format that allows you to code and type in your browser. This is included in the Anaconda distribution. The easiest way to use Jupyter Notebooks, therefore, is to download the

Anaconda app to your computer. After installation, you will have an app called Anaconda Navigator installed. You can create and manage enviromnents and use Jupyter Notebooks from it. Open the app and click on Jupyter Notebooks, after which it will launch in your default browser. See https://programminghistorian.org/en/lessons/jupyter-notebooks if you run into any problems or if you're looking for a gentle introduction to Jupyter Notebooks.

Week 2 – Between statistics and hermeneutics

Seminar (Tuesday 21 November)

In this seminar we will look at some of the underpinnings of computational techniques for digital history on a methodological level (next week we'll zoom in on tools & techniques). Central to today's class will be the debate between those who think that sound use of quantitative techniques means adhering to the scientific method (as in hypothesis testing on the basis of statistical significance) and those who believe that quantitative techniques are yet another way of opening up our source material, regardless how we come to our conclusions. In 2019, this debate reached a high point when Nan Z. Da criticized her own field (English literature) for using computational tools way too naively (see the first item in the Further Reading list). The pamphlets by Da and Underwood in the Chronicle of Higher Education can be read as summaries of this debate. Literary scholars dominate the digital humanities in the United States, but it is not difficult to see the implications of this debate for History, as Rockwell and Sinclair (both historians) demonstrate.

Readings:

- Da, Nan Z., 'The Digital Humanities Debacle. Computational Methods Repeatedly Come up Short', *Chronicle of Higher Education*, March 27 (2019).
- Gavin, Michael. "Is There a Text in My Data? (Part 1): On Counting Words." *Journal of Cultural Analytics*, January 25, 2020. https://doi.org/10.22148/001c.11830.
- Rockwell, Geoffrey and Stéfan Sinclair, *Hermeneutica. Computer-Assisted Interpretation in the Humanities* (Cambridge and London: The MIT Press, 2016), chapter 6: 'A Model Theory: Thinking Through Hermeneutical Things'.
- Underwood, Ted, 'Dear Humanists – Fear Not the Digital Revolution', *Chronicle of Higher Education*, March 27 (2019)

Further reading:

- Da, Nan Z., 'The Computational Case against Computational Literary Studies', *Critical Inquiry* 45:3 (2019) 601–39, https://doi.org/10.1086/702594.
- Rieder, Bernhard and Theo Röhle, 'Digital Methods: Five Challenges', in David M. Berry (ed), *Understanding Digital Humanities* (Basingstoke & New York: Palgrave Macmillan, 2012) 86–103.
- Online discussion forum on Da's article on the *Critical Inquiry* website (with contributions from, among others, Ted Underwood and Stanley Fish)
- Debates section of the *Journal of Cultural Analytics*, with responses to Da's article.
- Special issue on digital history of *BMGN-LCHR* from 2013, including a forum section with responses to Rens Bod's thesis of the humanities as a pattern-searching domain of knowledge.

Preparation

Think of keywords that might be of interest in the light of your research question. Words that are key to your topic, words that might denote a shift in focus or mentality, and/or words that have strong associations in your dataset that you are interested in. Particularly interesting are words that you can base the comparison within your data (see above, ‘On Data’) on. You will use these words to interpret the analyses that you will do in this and next weeks’ lab sessions.

Lab (Thursday 23 November) – Getting to know your data: preprocessing and counting

This week we will start with the notebooks for digital text analysis that I provide in this GitHub repository. ‘Doing things with text 1’ is all about one of the most important aspects of text mining: preprocessing. This cleans and restructures your data so that it allows you to find what you are looking for. ‘Doing things with text 2’ adds to this several approaches to getting to know your data: counting the total number of words in your texts (before and after preprocessing) and identifying and counting the most frequent words in your texts (in total and by word length). This notebook also contains some basic visualizations of these calculations, which you can use for your portfolio.

N.B.1 If you have not yet managed to find relevant data by today, please use the example text files ‘1977.txt’ and ‘2015.txt’ (the screenplays of Star Wars IV and VII, respectively) from the GitHub repository to practice with the notebooks.

N.B.2 The notebook ‘Doing things with text 2’ works for a single text file. The notebooks 3a and 3b allow you to do the same with multiple files (text or csv files, respectively). Do note that most of the code in these notebooks will treat all files that you load simultaneously as one single corpus, so make sure not to load any files that you would like to compare with one another.

Portfolio assignment:

You cannot publish anything based on digital text analysis without being absolutely clear on your data, so this is what we will start with. The question what your data is about should be answered both in a qualitative way—how representative is it for your topic of interest or, in other words: what’s in there, what’s not?—and in a quantitative one. With regard to the latter, think about the following questions:

- What does your data look like in number of documents (and what do they represent: books? Articles?)
- How coherent is your data: does it cover an entire period or archive, is it a sample?
- How coherent is your data in terms of words-per-document (and, possibly, *unique* words per document, specific types of words per document, etc.)?
- What is the OCR quality (and how can you say something about this)?

You can answer some of these questions based on looking at your data, others on the calculations in the notebook ‘Doing things with text 2’.

Commonly this information is provided in the form of tables, so make sure you add these to your portfolio. If you wish, you can also add some visualizations with the basic statistics to your portfolio: number of documents, number of words/tokens per document, type token ratio, etc. etc.

Week 3 – Text in numbers, text as language

Seminar (Tuesday 28 November)

Today, we will talk about the different approaches towards digital text analysis. Broadly speaking, there are two different philosophies: some try and turn the unstructured data that texts usually are into structured data formats that can be counted. Others aim to find ways to study text as language. We will explore both and look at the ways in which basic computational operations like word or n-gram counts and collocations on the one hand (see Moretti & Pestre), and part-of-speech tagging on the other (see Baker et.al.), can provide historians with new and meaningful insights.

The long list of further readings are mostly examples of empirical studies that use some of the techniques – or at least approaches – that are central to today's class. These are primarily meant for inspiration.

Readings:

- Baker, P., C. Gabrielatos, and T. McEnery, 'Sketching Muslims: A Corpus Driven Analysis of Representations Around the Word "Muslim" in the British Press, 1998-2009', *Applied Linguistics* 34 (2013) 255-278.
- Moretti, Franco and Dominique Pestre, 'Bankspeak. The language of World Bank Reports', *New Left Review* 92 (2015) 75-99.

Further reading:

- Blaxill, Luke, 'Quantifying the Language of British Politics, 1880-1910', *Historical Research* 86:232 (May 2013) 313–41, https://doi.org/10.1111/1468-2281.12011.
- De Bolla, Peter, *The Architecture of Concepts: The Historical Formation of Human Rights* (New York: Fordham University Press, 2013).
- van Eijnatten, Joris and Pim Huijnen (2021), 'Something Happened to the Future. Reconstructing Temporalities in Dutch Parliamentary Debate, 1814-2018', *Contributions to the History of Concepts* 16, pp. 52-82, DOI: 10.3167/choc.2021.160204.
- Guldi, Joanna, 'The History of Walking and the Digital Turn: Stride and Lounge in London, 1808–1851', *The Journal of Modern History* 84:1 (March 2012) 116–44, https://doi.org/10.1086/663350.
- Kestemont, Mike, Folgert Karsdorp, and Marten Düring, 'Mining the Twentieth Century's History from the Time Magazine Corpus', *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (2014) 62.
- Klingenstein, S., T. Hitchcock, and S. DeDeo. "The Civilizing Process in London's Old Bailey." *Proceedings of the National Academy of Sciences*, no. Published ahead of print (June 16, 2014). https://doi.org/10.1073/pnas.1405984111.
- McGillivray, Barbara, 'Computational Methods for Semantic Analysis of Historical Texts', in Kristen Schuster and Stuart Dunn (eds.), *Routledge International Handbook of Research Methods in Digital Humanities* (London: Routledge, 2020), pp. 261-274.
- Michel, Jean-Baptiste et al., 'Quantitative Analysis of Culture Using Millions of Digitized Books', *Science* 331 (2011) 176 -182

- Murdock, Jaimie, Colin Allen, and Simon DeDeo, ‘Quantitative and Qualitative Approaches to the Development of Darwin’s Origin of Species’, *Current Research in Digital History* 1 (2018), https://doi.org/10.31835/crdh.2018.14
- Scherrer, Marten et.al., ‘The rise and fall of rationality in language’, *Proceedings of the National Academy of Sciences* 118, 51 (2021). DOI: https://pnas.org/doi/full/10.1073/pnas.2107848118.
- Twenge, Jean M., W. Keith Campbell, and Brittany Gentile, ‘Changes in Pronoun Use in American Books and the Rise of Individualism, 1960-2008’, *Journal of Cross-Cultural Psychology* 44: 3 (2013) 406–15.

Lab (Thursday 30 November) – Approaches to digital text analysis

We will continue our exploration of our data with the notebooks Doing things with words 4, 5 & 6. These add some proven functionalities for text analysis, both on the keyword and on the corpus level:

- **Keyword searches** allow you to plot the relative frequency of one or more keywords in a graph (scatter plot). If your data is diachronic, you can see trends in word usage in this way.
- **N-grams** print the top n of phrases (word combinations) of 2, 3, or more words in each of your texts along with their frequency. For some questions this might be more relevant than single words.
- **Collocations** find statistical relations between words within a particular window of words. The meaning of a word can often be deduced from the words that appear strikingly often in its vicinity in a text. There are different measures to compute collocations and our notebook works with three of them (log likelihood, pointwise mutual information and raw frequency). The code prints the top n-collocations for one or more keywords within one or more windows based on each of the algorithms.
- Term frequency – inversed document frequency (**tf-idf**) is a basic statistic that sets texts apart from one another. The tf-idf function in notebook 5 highlights the top n-words that are most unique, specific or distinguishable for a single text (document, file) in your corpus compared to all texts (documents, files) in bar charts. The **cosine similarity** function yields a heat map that shows the similarity between each of the texts in your corpus (the greener, the more alike).
- The **part-of-speech** (POS) tags that the Spacy package adds to each word in the data you load allows you to search for specific types of words (verbs, nouns, proper nouns, adjectives – more are possible). The notebook prints the top n of each of these word types along with their frequency
- The **Named Entity Recognition** (NER) function creates lists of words that are recognized as particular entities: geographical locations, persons, product names, etc. Do realize that the quality of NER is heavily dependent on the language of your data (English always works best).

Portfolio assignment:

Experiment with each of the functionalities in Doing things with words 4, 5 and 6. Decide which of these make most sense for your topic of interest. Zoom in on these functionalities:

- Reflect on your choice for particular functionalities: how can these help studying your topic of interest? How do they relate to each other?

- Make reflected choices for your input: what keywords do you use? What texts do your contrast or combine? Make sure to experiment with different combinations and not necessarily stick to your first choice.
- Fiddle around with different settings: algorithms, window sizes, POS- or NER-categories, top n, etc. Demonstrate that your choices have not been informed by what happened to be default or easiest, but by your interest and expertise.

Week 4 – Data criticism in times of AI

Seminar (Tuesday 5 December)

In this seminar, we will go a step further when it comes to computational techniques and explore the particularities of AI and machine learning. This approach to data science has huge potential (see Gavin et.al.), but implies a major shift in thinking about how science is done that has important consequences (see Kissinger in the Further Reading). Machine learning, therefore, underlines the need for critical reflection in the use of computational techniques – particularly for whom the use of these techniques aren't native to their disciplines, like historians. This is what we will do in this seminar (see Koolen et.al.).

Readings:

- Gavin, Michael et.al., 'Spaces of Meaning: Conceptual History, Vector Semantics, and Close Reading', Matthew K. Gold en Lauren F. Klein (eds.), *Debates in the Digital Humanities* (Minneapolis 2019), 243-267.
- Koolen, Marijn, Jasmijn van Gorp and Jacco van Ossenbruggen, 'Toward a model for digital tool criticism: Reflection as integrative practice', *Digital Scholarship in the Humanities* 34 (2019) 368-385.

Further readings:

- Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proceedings of the National Academy of Sciences* 115, no. 16 (April 17, 2018): E3635–44. https://doi.org/10.1073/pnas.1720347115.
- Goldberg, Yoav, and Omer Levy. "Word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method." *arXiv:1402.3722 [Cs, Stat]*, February 15, 2014. http://arxiv.org/abs/1402.3722.
- Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *Proceedings of ACL*, 2016.
- Kenter, Tom, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. "Ad Hoc Monitoring of Vocabulary Shifts over Time." In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, 1191–1200. New York: ACM, 2015.
- Kissinger, Henry A. 'How the Enlightenment Ends', *The Atlantic*, June (2018).
- Tahmasebi, Nina, Lars Borin, and Adam Jatowt. "Survey of Computational Approaches to Lexical Semantic Change." *arXiv:1811.06278 [Cs]*, March 13, 2019. http://arxiv.org/abs/1811.06278.
- Wevers, Melvin and Marijn Koolen, 'Digital Begriffsgeschichte: Tracing semantic change using word embeddings', *Historical Methods: A Journal of Quantitative and*

Interdisciplinary History 53 (2020) 226-243, DOI: https://doi.org/10.1080/01615440.2020.1760157.

Preparation:

Listen to the item ‘First Contact’ in episode 803 of the famous podcast series This American Life. In the item, TAL editor David Kestenbaum talks to Microsoft scientists about the question whether AI in times of ChatGPT has got to the point that it is starting to have something like human intelligence? Form your own opinion on the conclusions the item offers.

Lab (Thursday 7 December) – Word embeddings

The final notebook Doing things with words 7 introduces you to machine learning by way of word embeddings (or vector semantics – depending on your field). The notebook lets you model your corpus in your own word embedding space. This allows you to find words with similar embeddings (vectors) to one or more keywords, which indicates semantic similarity. In other words: word embeddings let you find word meanings by showing you synonyms and other similarly used words within a given corpus of texts. Keep adhering to the principle of starting with a small data set (because the model can take long to train), but know that word embeddings work better the more data you put in—so you might want to try that at some point.

Portfolio assignment

Experiment with the functionalities in Doing things with words 7:

- Make reflected choices for your input: what data did you build your model on? What keywords do you use? Make sure to experiment with different combinations and not necessarily stick to your first choice.
- Reflect on how the clustering visualization works for you: what insights do you get from it? How can you make sure you see what you think you see?

Week 5 - Presentations

Seminar (Tuesday 12 December) + Lab (Thursday 14 December)

Student presentations (max. 20 minutes) on the progress of your research. The focus in the presentations should be less on the results than on the process of the project from topic or question to finding data to iterating between close and distant reading, as well as any challenges that you may encounter in the process. Use a Powerpoint/Keynote presentation, but restrict yourself to max 4 slides, the last of which with questions for the rest of the group. These presentations, after all, are among other things meant as an opportunity to use the brainpower of the group to collect suggestions, ideas, and best practices. To this end, every presentation should entice a discussion with the group, for which you – as presenters – are responsible.

Week 6 - Portfolio

Friday 22 December, midnight: portfolio deadline

Data Repositories

English & other languages

- Hathi Trust: https://www.hathitrust.org/ ("The largest digitized collection of knowledge allowable by copyright law", 18 miljoen items)
- Europeana: https://www.europeana.eu/en
- Project Gutenberg: www.gutenberg.org
- UU Library: https://cdh.uu.nl/2021/09/historical-data-available-via-yoda/
    - Eighteenth Century Collections Online
    - Guardian (1791-1909) & Observer (1910-2003)
    - Nineteenth Century U.K. Periodicals, Module 1
    - Times Digital Archives (1785-2011)
    - Times Literary Supplement (1902-2014)
    - (N.B. most of these datasets are very large and very badly OCR'd)
- Princeton's Curated List of Humanities Datasets: https://cdh.princeton.edu/research/resources/humanities-datasets/
- Historical magazines: https://centurypast.org/magazine-directory/
- Distributed Email of Secrets: https://ddosecrets.substack.com/ (very *very* large datasets from document leaks by transparency advocates)
- Movie scripts: https://imsdb.com/, https://www.simplyscripts.com/, https://gointothestory.blcklst.com/script-download-links-9313356d361c, http://www.script-o-rama.com/snazzy/table.html, http://www.awesomefilm.com/, https://sfy.ru/, https://www.dailyscript.com/movie.html, http://www.moviescriptsandscreenplays.com/
- Pop music lyrics: https://genius.com/

Dutch

- Architectuurtijdschriften: https://tresor.tudelft.nl/tijdschrift/architectuurtijdschriften/
- Delpher Tijdschriften: https://www.delpher.nl/nl/tijdschriften (o.a. Libelle, Margriet, Vrij Nederland, e.v.a.)
- Delpher Boeken: https://www.delpher.nl/nl/boeken (o.a. veel (academische) leerboeken)
- Digitale Bibliotheek der Nederlandse Letteren: www.dbnl.nl
- (Digitale) bronnencollecties Huygens Instituut: https://resources.huygens.knaw.nl/
- Natuurtijdschriften: www.natuurtijdschriften.nl
- Nederlands Tijdschrift voor Geneeskunde: www.ntvg.nl (via UU-proxy, historisch archief raadpleegbaar via zoekfunctie)
- Bij je docent: digitale edities van o.a.
    - Allerhande (1954-1973, 1979-2022)
    - BMGN (ca. 1970-2020)
    - jaarverslagen BVD/AIVD (1946-2019)
    - DSM I-V
    - Economisch Statistische Berichten (1916-2016)
    - De Groene Amsterdammer (1877-1940)
    - Militaire Spectator (1832-2007)
    - Staten Generaal Digitaal (1814-2018)