Topic Modeling for Swedish Historical Data

Ellinor Lindqvist, Eva Pettersson and Joakim Nivre
Department of Linguistics and Philology
Uppsala University
firstname.lastname@lingfil.uu.se

Abstract

In this project, we apply Latent Dirichlet Allocation (LDA) to a corpus of historical Swedish texts and explore whether different amounts of pre-processing affect the performance of topic modeling. The result shows that our topic model generally provides incoherent topics based on coherence scores and a qualitative evaluation. We conclude that the model's performance is affected by the size of our relatively small data set, and that the settings of our model, in addition, may have influenced the result. However, we also find that normalisation and lemmatisation had a modest positive effect by providing topics with the most interpretable words and less redundant topics.

1 Introduction

Topic models – statistical algorithms that automatically derive the general content from documents – have shown utility for large collections of modern data, as well as promising results for research in the field of digital humanities. In this project, we explore the use of a commonly used topic modeling (TM) method, Latent Dirichlet Allocation (LDA), for historical Swedish texts. In particular, we want to test if different pre-processing steps, in this case spelling normalisation, part-of-speech (POS) tagging and lemmatisation, lead to essential differences in the results compared to the use of raw data. When working with historical data, it can be desirable to narrow the amount of pre-processing steps, especially due to the general limitation of annotated data. Ideally though, we do not want to compromise on the quality of the results. Within the scope of this project, we study the performance of TM through standard evaluation in the form of coherence scores as well as a more qualitative assessment.

2 Background

TM has its roots in information retrieval techniques, such as Latent Semantic Analysis (LSA) (Deerwester et al., 1990), and the later Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999). Perhaps the most common topic model currently in use is Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a generative probabilistic model that uses prior Dirichlet distributions to generate document-topic and word-topic distributions. TM has also been widely used in the field of digital humanities. Newman and Block (2006) assessed several topic decomposition techniques to study a collection of colonial U.S. newspapers from 1728-1800. Yang et al. (2011) likewise explored the task of TM applied to collections of newspapers published in Texas from 1829 to 2008 using LDA. Dahllöf and Berglund (2019) applied LDA as a tool for 'distant reading' of two Swedish literary corpora by combining TM with a study of how the topics relate to gender in characters and authors.

While TM enables content-analysis of texts in a fairly quick and efficient manner, the application of these models provides some challenges. Firstly, there is a lack of developed practice guidance for methodological decisions. Maier et al. (2018) discuss important methodological challenges for topic modeling, and LDA in particular, and they point out that it is important to consider the specifics of the pre-processing steps, as well as the prior parameter setting (especially the number of topics as well as the alpha value used in LDA).

Secondly, assessing the performance of a topic model is not an easy task, since a gold standard list of topics is usually missing due to the unsupervised nature of these algorithms. For automatic evaluation, one of the most used measures is *coherence*. Röder et al. (2015) describes that the best performing coherence measure, CV, averages the normalised Pointwise Mutual Information (PMI) for every pair of words within a sliding window over an external corpus. A widely used implementation of the coherence measure in the Gensim library (Řehůřek and Sojka, 2010) is computed on the LDA model itself. Another way to assess the performance of a model is through human evaluation, or 'eye balling' method, by looking at the provided topics and determine whether the the topics are interpretable and of good quality. This type of evaluation could be performed either by experts with implicit knowledge of the data (e.g. Newman and Block, 2006; Yang et al., 2011), or by non-experts judging whether a topic has human-identifiable semantic coherence (Chang et al., 2009).

3 Methodology

The main purpose of this project is to evaluate how various pre-processing steps affect the performance of TM applied on a historical corpus. The lack of standardised orthography and often limited amount of annotated data pose challenges for automatic pre-processing of historical texts. Nevertheless, if pre-processing is crucial to maintain quality of a model's output, it would be a motivated step.

3.1 Data

The Gender and Work (GaW) research project, conducted at the Department of History, Uppsala University, studies how women and men sustained and provided for themselves in Sweden in the period from 1550 to 1800 (Fiebranz et al., 2011). Pettersson (2016) used parts of the GaW corpus to examine different normalisation approaches to tackle the problem of inconsistent spelling in historical documents, and found that an approach based on statistical machine translation generally performed the best.

In this project, we perform our experiments on a subset of the GaW corpus where we use a raw, unnormalised version of the data as well as normalised versions of the same texts. The texts are normalised using the same approach as Pettersson (2016), which is available as an online tool.[1] The documents in our corpus, Stora Malm, are protocols of parish meetings dating between the years 1728 and 1812.[2] The original documents of Stora Malm are divided into five different year spans (1728-1741, 1742-1760, 1761-1783, 1784-1795 and 1796-1812). In order to have more flexible division of time periods, and be able to feed the TM model with more (though shorter) documents, we split the provided documents into shorter documents based on the individual parish protocols, ending up with 234 documents (where most documents contain between a couple of hundred to a couple of thousand tokens). The size of our data set is 281120 tokens for the raw corpus and 295933 tokens for the normalised version of the corpus.

3.2 Pre-processing

We use standard procedures – spelling normalisation, part-of-speech (POS) tagging and lemmatisation – before feeding the texts into our TM model. The result is compared with the use of a raw version of our corpus.

The corpus is annotated using Efficient Sequence Labeling (EFSELAB), including the tasks of tokenisation, POS tagging and lemmatisation (Östling, 2018). EFSELAB is implemented in Python/C and available online.[3] The Swedish annotation pipeline is joint work with Aaron Smith, Jesper Näsman, Joakim Nivre, Filip Salomonsson and Emil Stenström. For POS tagging, EFSELAB uses a model trained on Universal Dependencies data (Nivre et al., 2017), while the lemmatisation is performed using a lexicon-based lemmatiser.

Based on common practice, we perform relative pruning by filtering out highly frequent and infrequent terms in order to reduce terms that are either too general or too specific to describe the content. More specifically, we remove terms that are among the 100 most frequent ones, terms that only appear once in the corpus, and terms that appear in more than 99% of all the documents in the corpus. We here include information about the POS, if the tokens are tagged. In line with the work of Dahllöf and Berglund (2019), we also use POS information to select only nouns and verbs, with the aim of capturing terms that are more informative. However, this is not done for the raw corpus version, due to the lack of POS tagging in this particular corpus.

After the pre-processing steps, TM is performed on each version of the corpus. An overview of all the steps in our method, and the order in which they are performed, can be seen in Figure 1.

[1] https://cl.lingfil.uu.se/histcorp/tools.html

[2] https://gaw.hist.uu.se/vad-kan-jag-hitta-i-gaw/kallunderlag/stora-malm—sockenstamman/

[3] https://github.com/robertostling/efselab

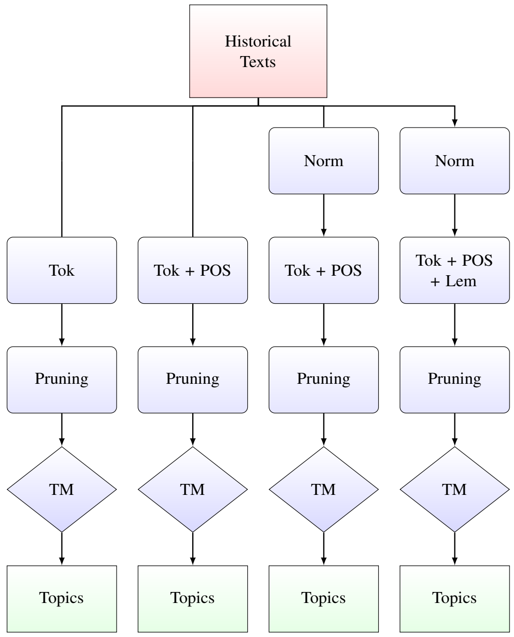

Figure 1: Methodological overview. Norm = normalisation, Tok = tokenisation, Lem = lemmatisation, Pruning = removal of high and low frequent terms, and selection of terms that are nouns or verbs.

3.3 Topic Modeling and Parameter Selection

We perform TM by applying LDA to all versions of our data set. The toolkit MALLET is an open source software (Java-based) for statistical natural language processing, including models for TM (McCallum, 2002).[4] Gensim offers a Python wrapper for LDA (Řehůřek and Sojka, 2010) that uses an (optimized version of) collapsed gibbs sampling from MALLET, which we use in this project.

Based on work of Maier et al. (2018), we focus the tuning on the number of topics (K) as well as the alpha value (α). A pre-experiment is conducted where the parameter settings are tested on a randomly selected subset of our raw corpus, containing 90 documents. Following the parameter values used in Maier et al. (2018), Dahllöf and Berglund (2019), or the default values, our setting is as follow:

$$K \in \{10, 20, 30, 40\}$$
$$\alpha \in \{0.1, 1.0, 10.0, 2.0 * K\}$$
$$\textit{iterations} = 1000 \text{ (default)}$$
$$\textit{random\_seed}^{5} = 1234$$

The three best performing models, in terms of coherence scores, are qualitatively evaluated, where the model with the most interpretable topics is selected (see evaluation methodology in section 3.4). In this experiment, we select the model with $K = 20$ and $\alpha = 10$. Though, the quality of the topics turned out to be generally low for all the models that were evaluated in our pre-experiment.

3.4 Evaluation Methodology

The provided topics are evaluated through automatic coherence scoring and a human assessment in terms of interpretability and coherence quality. For coherence scoring, the Gensim library (Řehůřek and Sojka, 2010) offers implementations of the coherence measures based on the work of Röder et al. (2015), where we use the CV measure. Each topic is represented by its most probable words, from where we select the top 20 keywords (based on the parameter selection from our pre-experiment, see section 3.3). Also, a human evaluation is conducted (by the main author of this paper), where each topic and its provided terms is evaluated and scored by using a simple grading system (see Table 1).

Score	Interpretation
0	no apparent theme, (almost) completely incoherent terms
1	a weak tendency of an interpretable theme, a few coherent terms
2	a tendency of a interpretable theme, several coherent terms
3	a pretty interpretable theme, most terms are coherent
4	a very clear theme, (almost) all terms are coherent

Table 1: Grading system to score topic quality.

4 Results and Discussion

4.1 Comparison of various Pre-processing Steps for Topic Modeling

Our main experiment compares how the topic model performs on different pre-processed versions of our corpus. The results from the automatic coherence scoring are shown in Figure 2. The coherence scores are similar between the different corpus versions, though indicating slightly decreasing scores when the degree of pre-processing increases. These scores suggests that the topic model, at least with our settings and data, do not benefit from more pre-processing steps in terms of coherence scores. However, it is likely that the coherence scores decrease due to reduction of spelling variations and

[4] http://mallet.cs.umass.edu/

[5] Fixed to enable reproducibility.

inflected forms of words. By not having spelling normalisation and lemmatisation, as in the raw corpus, we have more similar words within the same topic. An example of this phenomenon is present in the third topic from the raw corpus that has the words *dödt* and *död*, which could be two different spellings of the same word *död* (*death*).

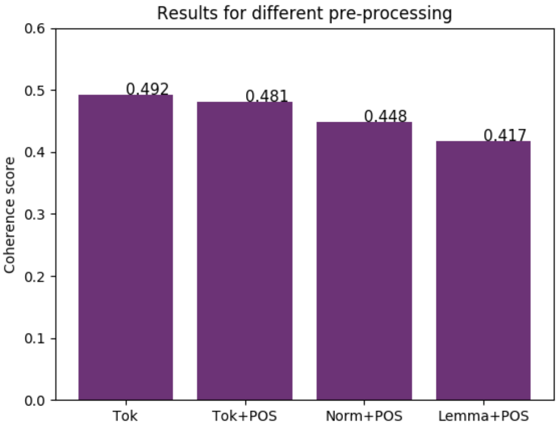

Figure 2: Coherence scores for different pre-processing versions of the data set.

A human assessment is also performed by evaluating the top words of each topic. The sum for each corpus (based on the scoring system presented in Section 3.4) was as follow:

Tok: 12
Tok+POS: 12
Norm+POS: 11
Lemma+POS: 14

Although the human evaluation cannot exclude subjectivity (being performed by only one annotator), it suggests that the topics from the lemmatised corpus are slightly more interpretable and coherent. However, most of the provided topics, in all of the corpora, are found to be incoherent and difficult to interpret. Some expected weaknesses can be observed within the different data set types. The raw, untagged version of the data set results in words that are sometimes difficult to recognise due to too different spelling compared to contemporary spelling. The output of the tagged, but unnormalised, version of the data set contains several words that have the wrong tag, presumably due older spelling (for example, in one topic the tokens *then, the, thet, thetta* and *thess* are tagged as nouns, though they are more likely to be articles or pronouns). This leads to a topic model that includes unwanted terms, with other POS than nouns and verbs. The normalised and tagged version of the corpus has topics that contain the same lemma with different inflections, e.g. *barnet, barn, barnets* (*the child, children, the child's*), and this is a tendency that can be seen in all corpus versions except the lemmatised corpus. The normalised, lemmatised and tagged corpus avoids all the mentioned weaknesses from the other corpus types, though most of the provided topics are still not interpretable due to too diverse words. The TM performs generally poorly on our data set, where we believe that the size of our corpus is the main reason. Compared to other related work, the corpus in our experiment is small, which likely affects how well a probabilistic model generates representative topics. Another potential explanation is simply that sensible topics does not exist in the data set. It is also possible that our parameter setting, or the chosen method for our TM (LDA), is a weak match for our data set.

5 Conclusion

In this project, we apply LDA to a historical Swedish corpus and explore how various pre-processing steps influence the result. We conclude that different pre-processing steps do not substantially affect the performance of the topic model of our settings. The results, based on coherence scores and a human evaluation, show that our TM generally provides incoherent topics that are difficult to interpret. We can, however, see that the normalised, lemmatised and tagged corpus provide topics with the most interpretable words, though most of the topics themselves are not cohesive. It is possible that our relatively small data set is less suited for TM, at least with the chosen settings.

For future work, our first (and perhaps most challenging) suggestion would be to train the topic model on additional data. Since historical data sets often are limited in size, this task would be more achievable by adding similar data from other domains or other time periods. Furthermore, it would likely be beneficial to further explore the tuning of parameters. It is also possible that the pruning step could be more aggressive by removing a larger proportion of frequent terms, and perhaps filtering out weaker terms as well within the targeted part-of-speech (e.g. help verbs). Finally, given that we would have a more functional topic model, it would be interesting to further explore TM for different historical time periods, and to provide a fuller analysis and evaluation with the help of historians.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296.

Mats Dahllöf and Karl Berglund. 2019. Faces, fights, and families: Topic modeling and gendered themes in two corpora of Swedish prose fiction. In *DHN 2019, 4th Digital Humanities in the Nordic Countries 2019, University of Copenhagen, Copenhagen, Denmark, March 6–8, 2019*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Rosemarie Fiebranz, Erik Lindberg, Jonas Lindström, and Maria Ågren. 2011. Making verbs count: the research project 'Gender and Work' and its methodology. *Scandinavian Economic History Review*, 59(3):273–293.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, Hannah Schmid-Petri, and Silke Adam. 2018. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2-3):93–118.

Andrew K. McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

David J. Newman and Sharon Block. 2006. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology*, 57(6):753–767.

Joakim Nivre, Željko Agic, Lars Ahrenberg, et al. 2017. Universal dependencies 2.0–conll 2017 shared task development and test data. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Robert Östling. 2018. Part of speech tagging: Shallow or deep learning? *Northern European Journal of Language Technology (NEJLT)*, 5:1–15.

Eva Pettersson. 2016. *Spelling normalisation and linguistic analysis of historical text for information extraction*. Diss., Uppsala Universitet.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. Citeseer.

Michael Röder, 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*, pages 399–408. ACM.

Tze-I Yang, Andrew J. Torget, and Rada Mihalcea. 2011. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104.