ML for a Spatial History Project

Jonathan Skjøtt
Minerva Schools at KGI

The Dataset:

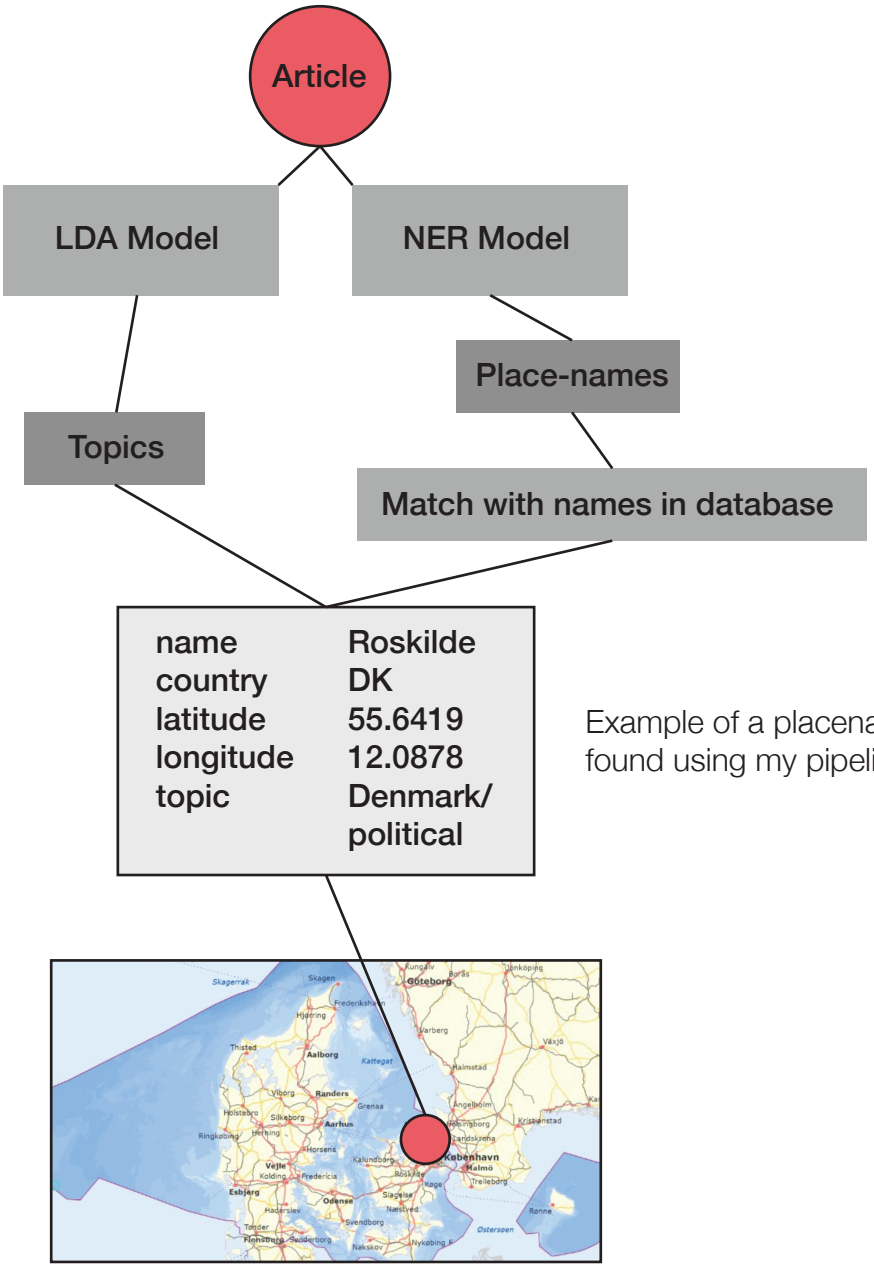
I created the dataset I am using for this project by scraping the website of the Danish Royal Library which has many resources publically available online. I scraped the article data for all journals of the Danish history journal Historisk Tidsskrift.

The Aim:

For each article in the journal I want to find the mentions of place-names in Denmark, Norway, Schleswig-Holstein, Greenland and the Danish colonies in India, the West Indies, as well as the Faroe Isands. Furthermore, I want to find how the themes discussed in the journal and the places mentioned in an article correlate with each other.

Approach:

I propose, implement and evaluate a pipeline for modelling the topics occuring in the journal articles using Latent Dirichlet Allocation(LDA) as implemented in the Gensim python package and using Named Entity Recognition(NER) as implemented in the SpaCy package for finding place-mentions.



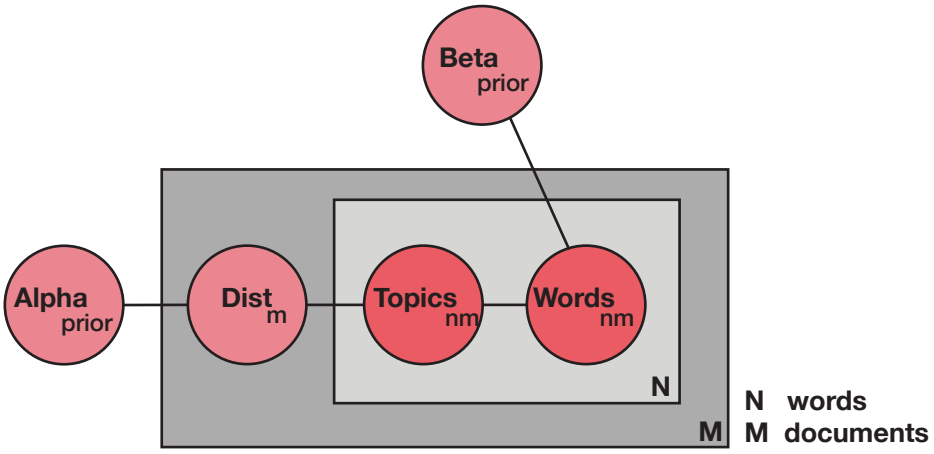
Example of a placename I found using my pipeline.

Dataset cleaning:

The script for cleaning the scraped data can be accessed here.

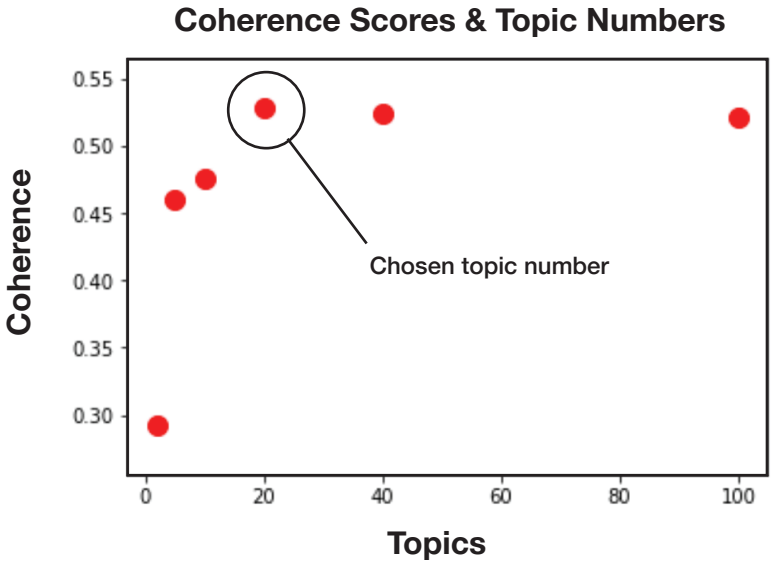
Topic Modelling:

I used LDA for my topic modelling. The model is considered to be part of natural language processing The model is based on the assumption that a document is a collection of different topics and that a a body of documents from a similar source can be modelled by finding the distributions of the underlying topics. The model is represented below using plate notation:

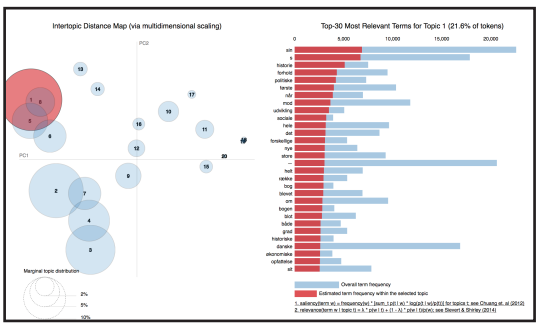


Finding the optimal model

I undertook an extensive data preparation and model evaluation in order to find the optimal model to use for my data set. The script for cleaning and evaluating the model can be found here. There were 750000 different words mentioned in the full corpus of journal. It is importnat decrease this, removing to common or uncommon words. I choose to remove any words mentioned more than 2000 times as well as any words mentioned less than 10 times. After this cleaning the corpus contained slightly more than 50.000 words. I also removed very short articles from the training set. In order to find the optimal lda model to use I generated multiple ones using different numbers of topics. I then used the coherence score metric to evaluate which number of topics produced the most distinct topics. I found that ~20 topics was the optimal number.



Go explore the interactive vizualisation I made for the 20 topic LDA model I used:



Named Entity Recognition:

NER is a kind of information extraction which aims at locating entities within text. There is only limited information about the NER algorithm used by the SpaCy library. Most likely the NER model used by SpaCy is of a statistical nature; using an ensemble of linear models whose weights have been learned by using an averaged perceptron.

Finding the Optimal Model:

Accuracy is inappropriate as a metric for evaluationg NER because of the relatively rare occurance of named entities in text. If every 20th word is a named entity and we classify half of those as named entities and all other words as regular words we would get an accuracy of 0.975 even though our model only finds half of the entities in the text. Therefore I decided to use recall as my metric. Recall here is defined as the ratio of entities actually located by the algorithm.

There are currently no Danish language named entity recognition solutions. Therefore, I decided to evaluate whether the German NER or multi-language NER model would perform best on the Danish language corpus. Danish is similar to German, which is why I figured the German model might perform better than the generic multi-language model.

	Multi-language NER	German NER
Recall:	**0.2044**	**0.0855**

The performance of both models should be considered mediocre at best. There is clearly a need for a Danish NER library. It is worth noting that I am working with a historic data set which decreases the performance of modern natural language processing tools which are based upon current language.

Discussion and Further Development

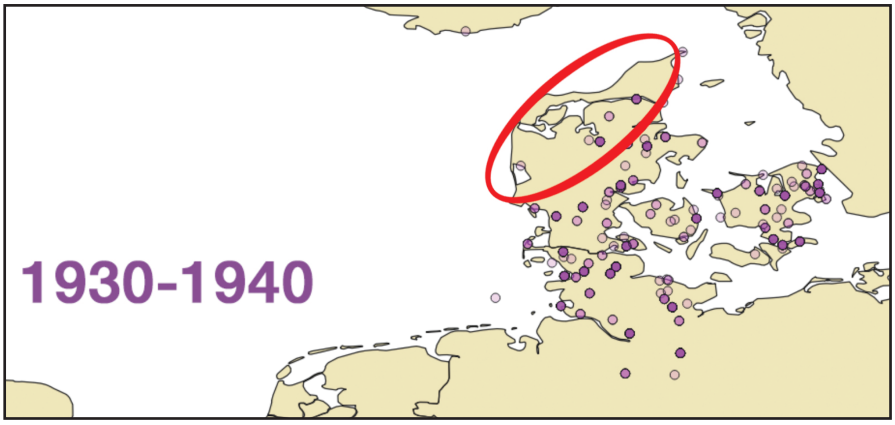
A full spatial analysis is beyond the scope of this final. However, at the moment I have suceeded in implementing the tools necessary for conducting a basic analysis. For a range of articles I am able to find the their topics as well as the placenames that occur in the articles.

Going forward with this project I must improve the performance of the NER portion of my pipeline. The argument made using the spatial analysis will only be valid if the recall of the NER model can be pushed above 80%; which is about 3 times better perfomance than the current 20%.

The topic model can be improved significantly in two ways. First, there are many one and two letter nonsense words which make their way into the topics. It would be straightforward to filter out words of 2 or 3 letters of less from the corpus when training the LDA model. I would have done this, however the model training takes about a day, which is too long for me to incorporate in this final. Secondly, I should have evaluated the alpha and beta priors more rigourously. It is insuficient to use just the coherence score when finding the optimal parameters of a topic model.

Going forward it would be highly interesting to develop my own named entity recogniser for Danish. SpaCy has an infrastructure for training NERs which would make it possible for me to develop the model. There are also other models than LDA which might be helpful in describing the content of the articles. Several friends of mine have trained and used word2vec models in attempts to understand textual data.

You can find an animated version displaying the placenames found in the period between 1860 and 1940 by following this link. Below you can inspect an initial vizualisation of the placenames found for the period 1930-1940:



Appendix:

1. LDA Implementation
2. Named Entity Recognition & Spatial Analysis
3. Article Dataset Cleaning
4. Cities Dataset Cleaning
5. History paper written based on this analysis