The Technical University of Denmark

42186 Model-Based Machine Learning

Project on speeches in *Folketinget*

Authors:

Anders Thuelund Jensen - s164057

Christian Dandanell Glissov - s146996

Mikkel Grønning - s144968

Toke Bøgelund-Andersen - s164202

May 30th 2020, Kongens Lyngby

1 Introduction and Data

In this project, we will investigate and build a graphical model based on transcripts of parliamentary speeches from the sittings in the Chamber of the Danish Parliament (Folketinget). This will be done to look into what topics are discussed using Latent Dirichlet Allocation (LDA) and Dynamic Topic Modelling (DTM). To do this we will use a dataset that has been made available by The Danish Parliament Corpus 2009 – 2017. The dataset includes transcripts of what was said in Parliament, who said, and more additional info to each speech. For descriptive statistics and prepossessing of the data please see the notebook.
More information about the dataset can be found at (Hansen, 2018).

2 Goals and Issues of the Project

In this project we have two goals. Firstly we would like to investigate what topics are discussed using a Latent Dirichlet Allocation. This should allow us to split the words into a set amount of topics. Hopefully, we would see some topics that fits subjects that we know Parliament discusses. Building on top of this we will carry out Dynamic Topic Modelling to investigate how topics are evolving over time.
With the Pyro-implementations we quickly realised that we have far from sufficient computer power to carry out the actual optimizations of the Pyro-models when parsing out complete dataset. To still obtain results on our full data we therefore used built-in Python libraries that are much more efficient.
We refer to out notebook for PGMs, generative stories and Pyro-implementations of the models.

3 Latent Dirichlet Allocation

We started with the Latent Dirichlet Allocation (LDA), (Hui, 2019; Murphy, 2012), where we used the `Gensim`-library to carry out the calculations so computations could be run in a reasonable time. We have built an LDA and Amortized LDA in Pyro (Bingham et al., 2018) but the computations are far too slow and take up too much memory. Everything including the theory behind LDA can be found in the notebook.
When we ran the LDA in Gensim we opted for 25 topics. This number was simply based on the fact that we wanted enough topics to cover the wide range of topics discussed in Parliament but also keeping the output interpretable. The resulting 25-word clouds are shown in figure 1. The size of the words corresponds to the

importance of the words for that given topic.

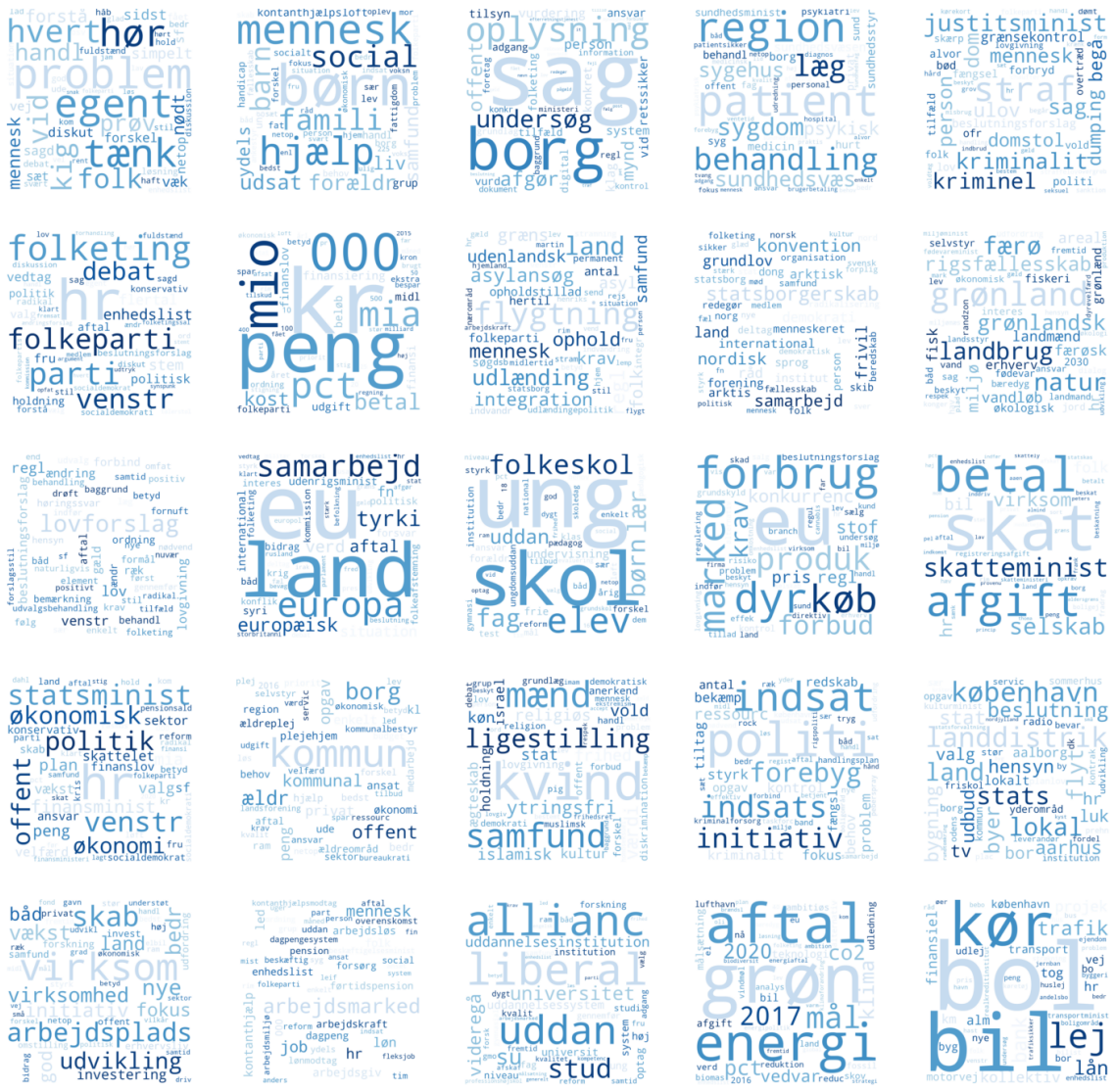


Figure 1: Would clouds of the 25 topics obtained using the Latent Dirichlet Allocation.

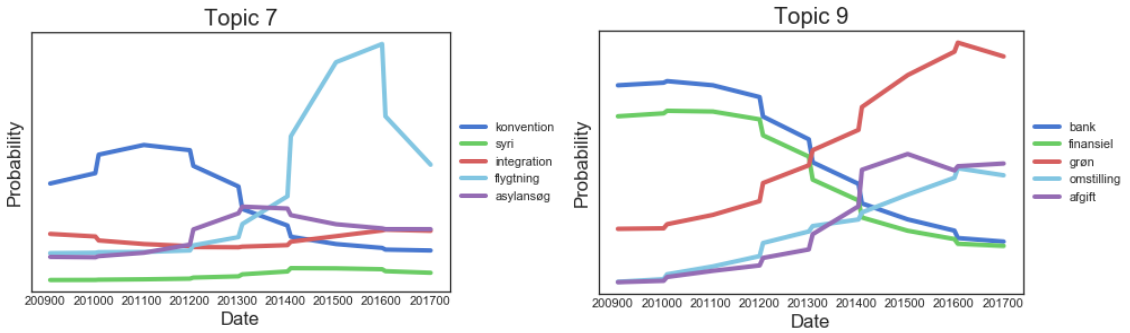
With knowledge of the Danish language, one should clearly be able to see word-clouds that relates to political subjects in Denmark. As some examples, we see in the middle word-cloud words that relate to primary education. The one to the left of it has a lot of words related to the European Union. In general, relevant topics are seen without perfect separation if you are into politics. We did not prioritize to do a lot of tweaking to the number of topics due to the time it took to run the model.

4 Dynamic Topic Modelling

As an extension to the LDA, we used Dynamic Topic Modelling (DTM), (Blei and Gerrish, 2006; Blei and Lafferty, 2011), to look for temporal evolutions within the topics. Once again we refer to the notebook for the theory behind it and the PGM. In the end, we had to see us defeated by Pyro in the attempt to make a fully working implementation. In the notebook the implementation can be seen, the model itself works, but the guide is not functional.

We still opted to have some results to look at. Therefore we once again turned to `Gensim`. After about 6 hours of training, we obtained our results, and figure 2 shows the evolution of two of the topics. Looking at topic 7 in figure 2a it mainly contains words related to refugees and asylum. It can be seen how the word *flygting* meaning refugees have a massive spike from 2015-2017. This makes very good sense as the refugee crisis was greatly discussed in that period in Denmark and the Parliament. Topic 9 seems to be a mixture of some financial words and words related to the environment. The words *bank* and *finansiel* (bank and financial) has consistently dropped since 2009 while *grøn* and *omstilling* (green and transition) has steadily grown. Even though this topic combines two slightly different political aspects, it is still interesting how the focus has shifted towards the green transition compared with the focus on banks after the financial crises. We have shown the evolution of selected words in six more topics in our notebook.

Once again with more optimization of the number of topics (which required a lot of computations) even more interesting outcomes should be found!



(a) Evolution of five words within topic 7 from 2009-2017.

(b) Evolution of five words within topic 9 from 2009-2017.

Figure 2: The evolution from 2009-2017 for five selected words within topic 7 and 9.

5 Conclusion

In this project we have investigated speeches in the Danish Parlaiment using LDA and DTM. Due to computational speeds we had to resort to the `Gensim`-library to obtain results on our real data. The results were though still interesting. In the notebook a longer discussion is found.
With the available data more analysis can be carried out with the available time. A few suggestions for future work could be:

- **External Data**: Incorporate data from e.g. the stock market, unemployment etc. and to see if correlations between topics and other external factors can be found.
- **Consistency within parties**: One could also look into how consistent parties are and do the analysis on a party level. Parties might change their focus over time.
- **Sentimental analysis**: Even though parties discuss the same topics in Parliament they probably *feel* different about the topics. Adding sentiment scores to the speeches, if possible, would allow us to discover the sentiment evolution towards different topics.

6 References

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., & Goodman, N. D. (2018). Pyro: Deep Universal Probabilistic Programming. Journal of Machine Learning Research.

Blei, D. M., & Gerrish, S. M. (2006). Dynamic topic models. ICML '06: Proceedings of the 23rd international conference on Machine learning. Retrieved May 30, 2020, from https://mimno.infosci.cornell.edu/info6150/readings/dynamic_topic_models.pdf

Blei, D. M., & Lafferty, J. D. (2011). Machine learning — latent dirichlet allocation lda. Retrieved May 30, 2020, from https://radimrehurek.com/gensim/models/wrappers/dtmmodel.html

Hansen, D. H. (2018). The danish parliament corpus 2009 - 2017,v1. CLARIN-DK-UCPH Centre Repository. Retrieved May 30, 2020, from https://repository.clarin.dk/repository/xmlui/handle/20.500.12115/8

Hui, J. (2019). Machine learning — latent dirichlet allocation lda. Medium. Retrieved May 30, 2020, from https://medium.com/@jonathan_hui/machine-learning-latent-dirichlet-allocation-lda-1d9d148f13a4

Murphy, K. P. (2012). Machine learning a probabilistic perspective. Massachusetts Institute of Technology, The MIT Press.