Visualization and usage of Topic Modelling
(MICS2-13 – Knowledge Discovery and Data Mining)

Keller Patrick, Meder Jeff, Olszewski Maya

D, Month 2018

1 Introduction

With the exponentially growing amounts of data accumulating in our society, it is very easy to lose the overview in a collection of data. This is where the field of topic modelling comes in handy. Topic modelling is a technique that finds topics in a document, or in a collection of documents. But even with a topic model it still can be hard to imagine what the data set represents and how the components are linked with one another. This is why visualization is so important. It allows, if done correctly, to get an intuitive overview of the data set.
The task of this project was to think about different kinds of visualization possibilities and to implement the one that seemed the best to us.
We decided to go with an interactive visualisation of the data set. Starting at one document, the visualisation shows some statistics about the topics in the document compared to the whole data set in a pie chart, shows interactively which words from which topic are present to what extent in the given documents, and shows the most similar documents, to which one can also navigate.

2 Related Work

In 2012, A.J.B Chaney and David M. Blei [2] implemented a visualisation of topic models with a user-friendly interface as output, that should be easy to understand for everyone. The interface permits it to navigate through the data set in the following way: one starts with a selection of topics in the whole data set, ordered by which topics appear most often. One then has the possibility to select a topic, which then shows a list of words that constitute this topic, the related documents to this topic and also related topics. From here one can select another topic, or select a document, which in turn will show a view of the document with the related topics of that documents and related documents. Then one can select one of the related documents or topics and so forth. A preliminary user study has found that users found this approach very intuitive and easy to comprehend [2].
More recently, a tool called *LDAExplore* has been presented [1]. This is also a tool for the visualisation of topic models. It works in a different way than the first visualization. First of all, it has several components. One would be parallel coordinates, that show the topic distribution in all the documents. A treemap is used to show the percentage of each topic by assigning a rectangle to each topic where the size of the rectangle is proportional to the importance of the topic in the collection.

If a topic is selected, one gets a similar treemap with the 10 most frequent words in the document. Besides the treemap, one also has the possibility to search for top words and other methods for document filtering. The preliminary study conducted on this tool [1] showed that if no filtering is applied to the data set, the parallel coordinated look cluttered and hard to read. Overall, not all study participants were able to determine which topics were the most/least important.

3 Methodology

Topic modelling is an unsupervised machine learning technique that looks for relations in document collections by finding themes in them. For our visualization, we first apply a topic model called latend Dirichlet allocation (LDA) on our data set. It is a topic model introduced in 2003, which is a probabilistic model over the whole data set, where topics have distributions over words that are present in the documents [3].

4 Implementation

Introduce the data used in the experiments, the setup of the experiments, and the results and/or comparisons.

5 Conclusion

Conclude the whole project in short text.

References

[1] Ashwinkumar Ganesan and Kianté Brantley and Shimei Pan and Jian Chen *LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation*, CoRR, abs/1507.06593, 2015

[2] Allison June-Barlow Chaney and David M. Blei, *Visualizing Topic Models*, ICWSM, 2012

[3] David M. Blei and Andrew Y. Ng and Michael I. Jordan and John Lafferty, *Latent Dirichlet Allocation*, Journal of Machine Learning Research, volume 3, 2003