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

Keller Patrick, Meder Jeff, Olszewski Maya

May 30, 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. A topic is a collection of words which occur together and form a theme. However, even with a topic model it still can be hard to imagine what the data set is actually representing or how the components are linked with each other. A potential aid to tackle this problem is visualization. It represents an important help as it allows, if done correctly, to get an intuitive overview of the data set as well as extracting new information and knowledge.
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 topic, the visualisation shows some statistics about the distribution of that topic in the documents of the whole data set in the form of a pie char. Furthermore, it also interactively displays the frequencies of the words constituting the topic and, upon selection of one of the documents lists the most similar documents.

2 Related Work

In 2012, A.J.B Chaney and David M. Blei implemented a visualisation of topic models with a user-friendly interface as output, that should be easy to understand for everyone [1]. The interface permits 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 [1].
More recently, a tool called *LDAExplore* has been presented [2]. 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 are 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 [2] showed that if no filtering is applied to the data set, the parallel coordinates look cluttered and hard to read. Overall, not all study participants were able to determine which topics were the most or least important ones.

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].
We implemented an euclidean distance measure that we applied to the topic distributions over the documents of the collection in order to find the most similar documents.
We decided to visualise the information in the manner described below, which has a similar navigation idea as the Chaney and Blei tool, as their user study has shown that this approach is very user-friendly and intuitive [1]. It also allows us to give a global perspective of the data collection, yet still providing enough information on the individual documents.

4 Implementation

For our implementation we used the LDA topic model from the Java library MALLET on our data set. In order to test our program, we made use of the sample data set provided by MALLET, from which we generated 20 topics, after doing some preprocessing, including the removal of stopwords and stemming of tokens. We developed an interactive user interface with Java Swing and JFreeChart, which allows the user to select a directory containing the data set to be visualised. The resulting view shown in Fig. 1 displays on the left a list containing all the topics extracted from the data set. The user is able to select the topic which he wishes to visualize from that list. The remaining parts of the interface will adapt accordingly. In the center of the interface, a pie chart shows the five documents where the selected topic is most present. By hovering over one of the slices of the chart, the percentage of the chosen topic that constitutes the document is displayed. Upon a click on a slice, the most similar documents to the one selected, determined by euclidean distance on the topic composition vectors, are displayed. Another information being displayed is the frequency throughout the data set of the words defining the topic.

5 Conclusion

This project has allowed us to gain deeper understanding of the subject of topic modelling. It has challenged us to evaluate different data visualisation techniques in order to find one that is easy to understand and quite complete in introducing the user to the underlying relations and structures of the data set.

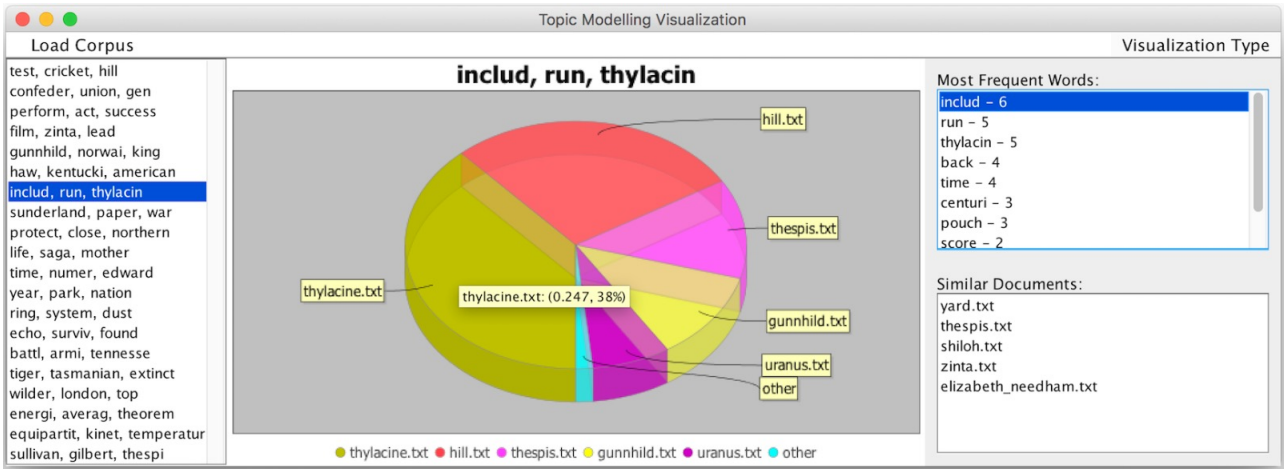

Figure 1: User interface with the generated topics on the left, the pie chart representing the distribution of the selected topic over all documents in the center, the words constituting the topic and their number of occurrences on the top right and the similar documents on the bottom right that appear upon clicking on a slice of the pie chart. On the pie chart one can see a mouse-over of the slice thylacine.txt, where the first position, 0.274 represents the fraction of the document constituted by the current topic among all topics. The second position, 38% represents the percentage of the topic distribution belonging to document thylacine.txt.

References

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

[2] 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

[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