Visualization project learning diary 2 - Visualizing Reuters Corpus

Heli Huhtilainen - 19.4.2021

Domain problem characterization

After reading about the nested model by Munzner I was wondering about the user and the purpose of the visualization that was considered in the outer layer of the model (Figure 1). I think the main user group could be data scientists or researchers or someone who is planning to use the Reuters corpus in for example NLP. In explorative data analysis, it is important to find the relevant aspects of the data that can be utilized in for example when building a machine learning model with the data in question. Another user type could be the users who are interested in news articles and their characteristics in general, and how for example different topics have arisen in different places. I guess in my situation the project is a bit more data-driven than domain and problem-driven.

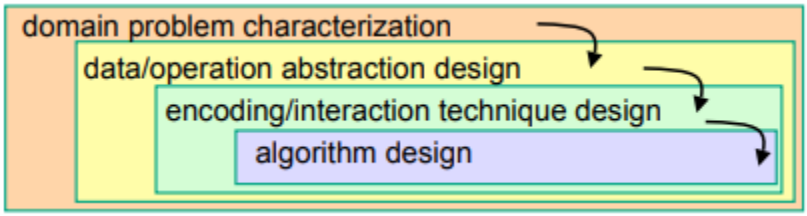

Figure 1: The layers of the nested model show different stages of the visualization design process.

Data cleaning challenges

I started analysing the data frame from last week a bit more carefully. The number of rows was smaller than in our deep learning project data. Also, the locations seemed a bit weird, and as I went through some more documents, the earliest ones did not have this record at all. The text parsing also did not work as it should have. I did corrections to these in the load script and got the row count and locations to match. The data now has 299773 rows as it should. I removed extra white spaces from the headlines but did not touch the texts as they might contain some useful structures. I also replaced all null values with empty strings.

Transformation

I decided to replace the index with the id column. The date would have been an interesting index option, but there are several rows with the same date and all the documents are dated between 4.6.1997 - 31.7.1997. As the dates are from such a short period, they are only mildly interesting. I consider if it would be useful to have a feature for each topic for each document, but this would add 103 more dimensions to the data frame, which seems a bit unhandy. I decided to make this change later if it proves to be necessary. As the data I have is text, it only

has some numeric information if I process it, for example, do some word counts. The topic columns, locations and countries are categorical.

Exploration

I explored frequencies of topic combinations and single topics, these are visualized in Figure 2. The distribution in both is very polarised: there are only few topic combinations/topics that occur often, and most are rare. The word cloud made of the topic distribution looks interesting (Figure 3). How ever, word clouds are quite slow to create and there are many texts. There probably should only be few word clouds, maybe only if the user wants more details about a text.

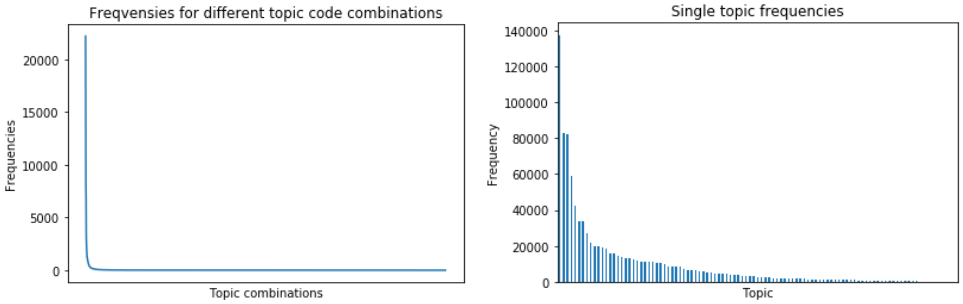

Figure 2: Topic combination and topic distributions might set some challenges to make interesting visualizations.

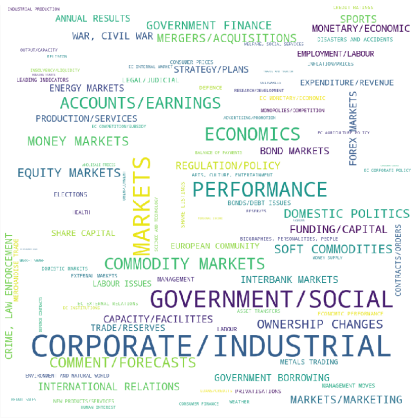

Figure 3: Topic distribution as a word cloud looks quite promising and visually pleasing.

When I explored the locations, I found out that there were 225 countries and two rows without the country. The locations had different formats: some had both city and country, so this column needed cleaning. After removing the redundant country info from locations, 4517 locations were remaining. There still are some weirdness, like four rows have “Africa” as their country. Other anomalies are concepts that are not really counties. There are 427 rows that state EU as their country, and 500 rows that have United Nations. Also, according to Google, there are now 195 countries in the world and the Geopanda world data frame has 177 countries. I did a thorough cleaning and comparisons with Geopandas world map countries and the Reuters country

column. Finally, I ended up with 207 countries after cleaning all the doubles and errors that I could find.

Plotting with maps

I decided to try mapping the document frequencies on a world map. For this, I tried Geopandas as it was familiar to me. I managed to combine most of the Reuter counties with Geopandas data, but 44 countries (or concepts like UN) still need to be mapped somehow. There would also be a possibility to use cities, but not all rows have that data. I tested the rows that I have and again saw some polarization: most of the documents are from the USA.

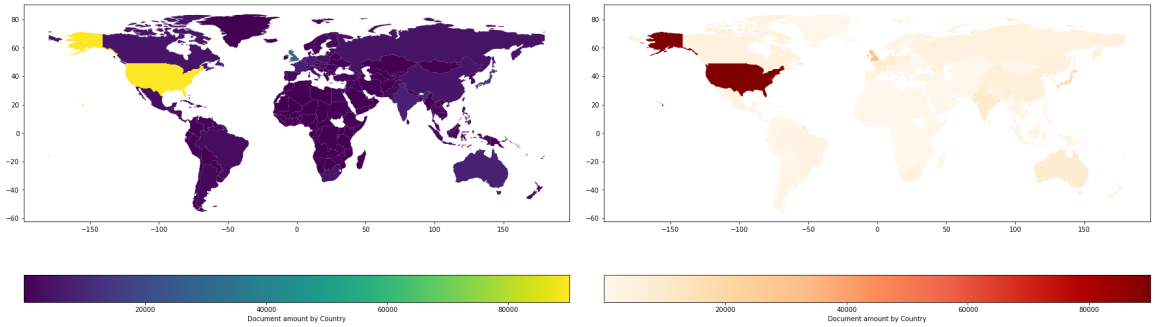

Figure 4: Most of the documents are located in the USA, and that dominates the map view.

I am not sure if this mapping approach is working too well. At least the results that are mapped in Figure 4 seem a bit uninteresting. Maybe some other colour scheme might be better, or maybe the document locations could be shown as points on the map. Perhaps the map could be a starting point, where the user can zoom in and find more details after zooming.

Graph layout

As an alternative approach, I wonder if the data could be represented in a graph form to show relationships between different topics and countries. I will think about this a bit more and do some testing if the idea seems to be worthwhile.

Next week

Next week I want to design and try out the interactive environment, and how the user can decide what to see. I will try combining Geopandas with Plotly and start the web application.

References

1. Source code for the project: https://github.com/apndx/ReutersVisualizer
2. A Nested Model for Visualization Design and Validation: https://www.cs.ubc.ca/labs/imager/tr/2009/NestedModel/NestedModel.pdf