Visualization project learning diary 3 - Visualizing Reuters Corpus

Heli Huhtilainen - 26.4.2021

The idea for the first interactive visualization view

The Reuters Corpus documents are written in many places around the world, and I wanted to visualize how the topic distributions differ in different locations. One way to visualize this could be with word clouds, though more conventional methods like bar charts or maybe parallel coordinates might also provide interesting and probably more informative and numerically correct viewpoints. First I however wanted to try out what word clouds have to offer.

Exploring with WordClouds and Geopandas

I decided that first I would like to try plotting word clouds that have topics of the reports of each country or area in the Reuters corpus. If the data had the country area geometry details, then the word cloud would take the shape of the country. Otherwise, the shape would be a square.

I had previously used WordCloud for Python (2) and wanted to use it here as well. The world cloud receives the cloud shape as an image, where the white areas of the image are ignored and other areas are filled with words.

First I thought that to have the topic counts for each country I would need to have the topics in separate columns in the data frame. Then I thought that I could create a string for each country that has all the topic lists concatenated as a string. This could then be given to the word cloud. I tried concatenating the topics to one string, but as some of the topics contain many words, the word cloud interpreted these wrong. Next, I realised I could have the topic frequencies in a dictionary for each country, and this solution seemed to work quite well.

I made the country-shaped masks using Geopanda's (3) plotting functionality and saved the resulting image files to be used in the word clouds. I also saved the data frames that I had created, so I could more easily use them later. When I was saving data frames that had the geometry details and loading them again, I realised I needed to use Geopanda's own file format for saving the geo data frame to reserve the geometry data in the right format. I made the needed changes to have the data in the right format.

In Figure 1 there are some examples of the word clouds I created. I think the clouds without borders look more stylish, but as the words are too sparse with the topic data I use at the moment, the shapes are not recognizable without the borders. I decided to plot the borders, but I am not totally satisfied with the result as the borderlines seem to be plotted in a non-continuos and almost sketchy manner. At the moment the colours used in the word clouds are defaults from the library and as nice as they are, they do not convey any meaning. As we can see from Figure 1, they can also change each time a word cloud is plotted. I think it would be beneficial to

map a certain colour for each topic, so the users could more easily recognize common features between the word clouds. Another option would be to use separate colours for each country, for example using colours taken from a flag. This way the topics would be more strongly connected to the country in question.

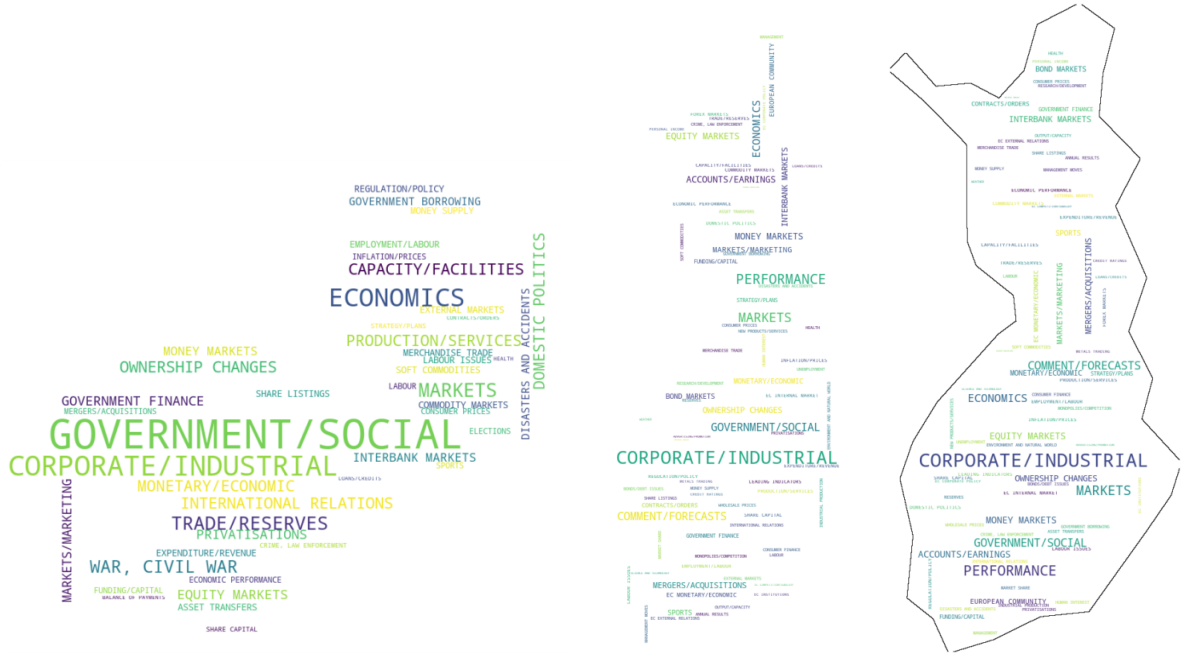

Figure 1: The word cloud on the left is of Zambia, and the other two are of Finland. The masked word cloud countries are too vague without borders, but the borders are plotted in a sketchy manner.

Exploration and prototyping with Dash

As I now had a working solution for creating topic word clouds, I decided to start building my web application. I had previously tried Plotly and Dash sounded promising as it used Plotly and enabled web development. I read a couple of Dash tutorials to get going and used the tips they provided to first build a view that had one hard coded example of a word cloud. When that worked out ok, I added a dropdown where a country could be chosen, and after that, the correct word cloud would be built. Figure 2 shows a screenshot of the MVP solution.

At the moment the word cloud is just an image, and quite static in nature. I would at least want some sort of zooming to be enabled. However, as Dash does not provide word clouds as a ready solution, I need to see how nicely I can implement this with the resources that I have available. I think I would like to have some meaningful statistics attached to this view like how many reports were there, and what is the percentage of the total report amount for this country.

Figure 2: The MVP view of the word cloud visualizer looks promising, but still has some problems regarding the right size of the image, lack of numerical details and interactivity.

I tried improving my solution so that the country shapes would not be created if they already existed. I could not yet find a solution for this, as the changes I tried did not yet work out. I think I might try to resolve this, but it is not a big priority.

Plans for the next week

Next week I will deploy my MVP solution to Heroku, so it will be easier to reach. I will create an overview world-shaped word cloud with the whole corpus distribution as a start view. I will see if I can change my word clouds to be more interactive or at least zoomable.

I will also try presenting the same date with bar charts, and parallel coordinates, or other solutions that I find suitable.

References

1. Source code for the project: https://github.com/apndx/ReutersVisualizer
2. Word Cloud library by Andreas Mueller: https://amueller.github.io/word_cloud/index.html
3. Geopandas: https://geopandas.org/
4. Dash: https://plotly.com/dash/