P2 – ODS

By Sara Liv Broe and Camilla Ejby Pedersen

Research question:

Which topics was mostly used to discuss Donald Trump?

First of all, we visited the site open-platform.theguardian.com/explore/ and found out what we wanted. On this site there was different parameters we could use such as date, tags and so on. By using the parameters, we got an instant look at which articles we would get. We could hereby quickly find the different filters that we needed in our code.

We search for ‘Trump’ and so that the relevant articles had the show-tags ‘usnews’ and ‘worldnews’. Then we made another search where we used the said show-tags. Furthermore, we chose to narrow the articles to a year. So, we asked for all articles related to ‘Trump’, ‘usnews’ and worldnews’ which were publicated between October the 30th 2018 and October the 30th 2019. Since Trump is a highly popular topic, we felt quite confident that we would get enough articles within the time frame.   
  
We searched on ‘content’ and got 8810 articles.

These articles got imported to Python. Since we narrowed it down to a year, we got 356 days worth of articles. The code we used imported the articles pr. day. Which means that we could have one article in day one, but 50 articles in day two.

The articles was saved as Json-files.

We then asked to get all of the Json files – however, we ran into a problem. It seems that there were some files (that the computer put in between the Json files) that couldn’t be read in Python. We therefore asked to get all the files, but, if Python couldn’t read the files, then we asked it to exclude these.

By now we had only limited the files through the site, mentioned earlier in this document. So the next step was to use code to limit the articles further. We used stopwords – we imported this part and therefore couldn’t be sure what exactly these stopwords were.

Furthermore, we asked not to include comma and dots and that the words should be separated by white space.

After this we chose the 10 most common words for each article and to see these in a list. However, it took way to long, so to find out if our code was working or not and to get an indicator as to which words we would get, we chose to only take a small part of the articles and test our code on these.

We asked for the list in reverse, since we wanted the numbers that occurred most frequently on top of the list.

It turned out that our code was working, however, we ran into a problem, since we both got relevant words like ‘Trump’, ‘US’, but we also got words like ‘said’ and ‘would’.

This led us to limit the words even further. Also, we had only searched the articles for ‘Trump’ which means that we could end up with articles written about ‘Ivanka Trump’ or ‘Melanie Trump’.