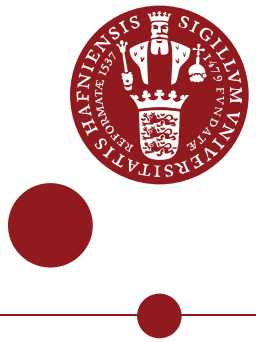
UNIVERSITY OF COPENHAGEN



Portfolio 2

Open Data Science

Eksaminand: Stephanie Rose Acampado Soelmark, PZG932

Gruppe:

Martine Ingemann Jørgensen, JPR328

Natacha Rylander Bech, TGZ940

Stephanie Rose Acampado Soelmark, PZG932

6 januar 2020

Introduction

For this assignment we will be working with the "The Guardian" dataset. To define a research question, we looked through the dataset and found it to be news articles. We then limited the data, meaning the articles, to a month, here from September 1 to November 1, 2019. From this we got 1669 articles. In these we looked for repetitions and found that Boris Johnson's name came up several times. This led us to the following research question:

Under which circumstances were 'Boris Johnson' discussed in relation to the 'Politics' category, in the Guardian from September 1 to November 1, 2019.

Our key findings

Task 1

For this task we printed the number of id's/text's/fields' which are 1669, symbolizing the number articles in our chosen dataframe.

Task 2

To derive a document matrix, we used CountVectorizer to remove stopwords, added a token pattern with a regular expression and subsequently transformed the data to a matrix in our "vecfit" variable. Then, we can illustrate the amount of unique words that were tokenized, namely, 43713 words from our textcorpus.

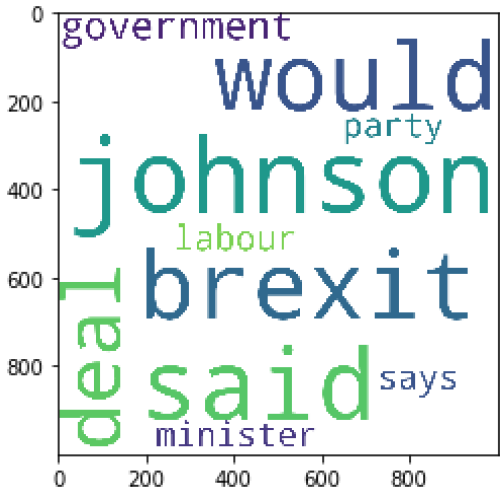
We calculate the sparsity of our matrix to be 0.99, which signifies that out matrix is sparse. As the next step, we calculate the TF-IDF weighting of our matrix.
A key finding in this task is the previous and the tokenized word count. The previous is 2.330.313, while the tokenized is 1.055.035. We have used CountVectorizer dictionary stopwords and the regular expression for this.
Another key finding is the average length of a document. We calculated this to be 632,1 after removing the stopwords from the text.

Task 3

For this task we narrowed down our search and generated a query, focusing on ‘Politics’ in relation to Boris Johnson. We choose this topic in order to answer our research question. Hereafter, we created a matrix over our query to use for the next task.

Task 4

In relation to topic modelling, we made the following word cloud. These illustrate the top 10 words in four different components. Thus, it becomes evident that topics regarding brexit, government and the deal have been on the agenda during the two months, September and November. These comply with reality as they do in fact reflect current and well-debated topics in the UK. Therefore, by further assessing and interpreting the situation in the UK based on news, we can use our model to identity and reveal interesting topics that may set the scene for further research.



Conclusion

To answer our research question, from our topic modelling we can conclude that Boris Johnson is discussed in relation to politics, Brexit, the Labour party, government, minister and England. These topics makes sense as Boris Johnson is the newly appointed prime minister of England, and thereby a part of the government and the ongoing discussion of Brexit.