UNIVERSITY OF COPENHAGEN

Portfolio 2

Open Data Science

Martine Ingemann Jørgensen, JPR328

Natacha Rylander Bech, TGZ940

Stephanie Rose Acampado Soelmark, PZG932

November 4, 2019

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 September 7, 2019. From this we got 1467 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 the Guardian from September 1 to September 7, 2019?

Our key findings

Task 1

For this task we print the number of id’s/text’s/fields’ which is 1467, symbolizing the number articles in our chosen dataframe.

Task 2

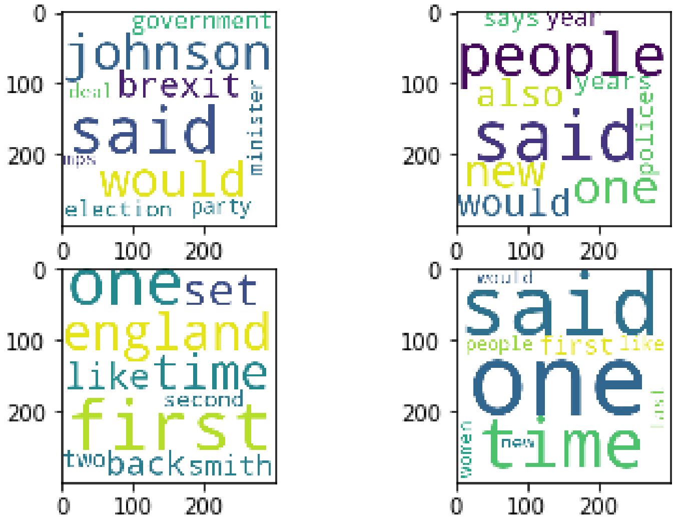
A key finding in this task is the previous and the tokenized word count. The previous is 1.528.838, while the tokenized is 49.907. We have used CountVectorizer dictionary stopwords and the regular expression for this.

Task 3

Here we narrowed down our dataframe from 1467 document to 100 documents, to make the dataframe simpler.
We then used Groupby and a random sample of 100 documents to find out how much Boris Johnson is mentioned in the different section and found that his name occurred several times.

Task 4

In relation to topic modelling we made the following word clouds. These illustrate the top 10 words in four different components.



Conclusion

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