**Assignments - BIG DATA ANALYTICS**

**text analysis**

**Greta Kurpicz & Dominik Walter**

**FS 2021**

**University of Lucerne**

**Professor**

**Luigi Curini**

Table of contents

[1 Assignment - Wordfish & Wordscore 1](#_Toc69985239)

[2 Twitter and STM 2](#_Toc69985240)

[3 Semi-supervised classification 3](#_Toc69985241)

[4 Naïve Bayes - SVM - Random forest - Gradient Boost 3](#_Toc69985242)

[5 Word embedding - WE-estimates and ML 5](#_Toc69985243)

List of Figures

[Figure 1: Most frequent words and their relationship in terms of cosine. 5](file:///C:\Users\labor4\Desktop\Dominik\R_Studio\QUANT%20II\BigDataAnalystics\BigDataAnalytics.docx#_Toc69985246)

# Assignment - Wordfish & Wordscore

# Twitter and STM

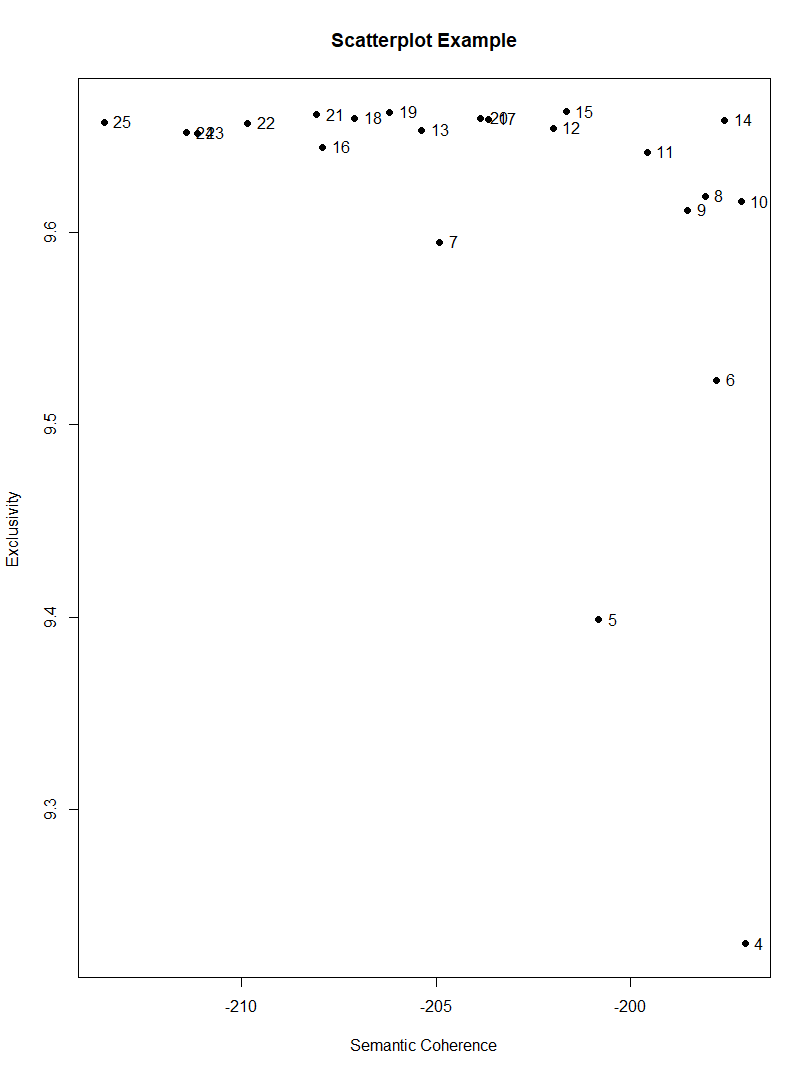
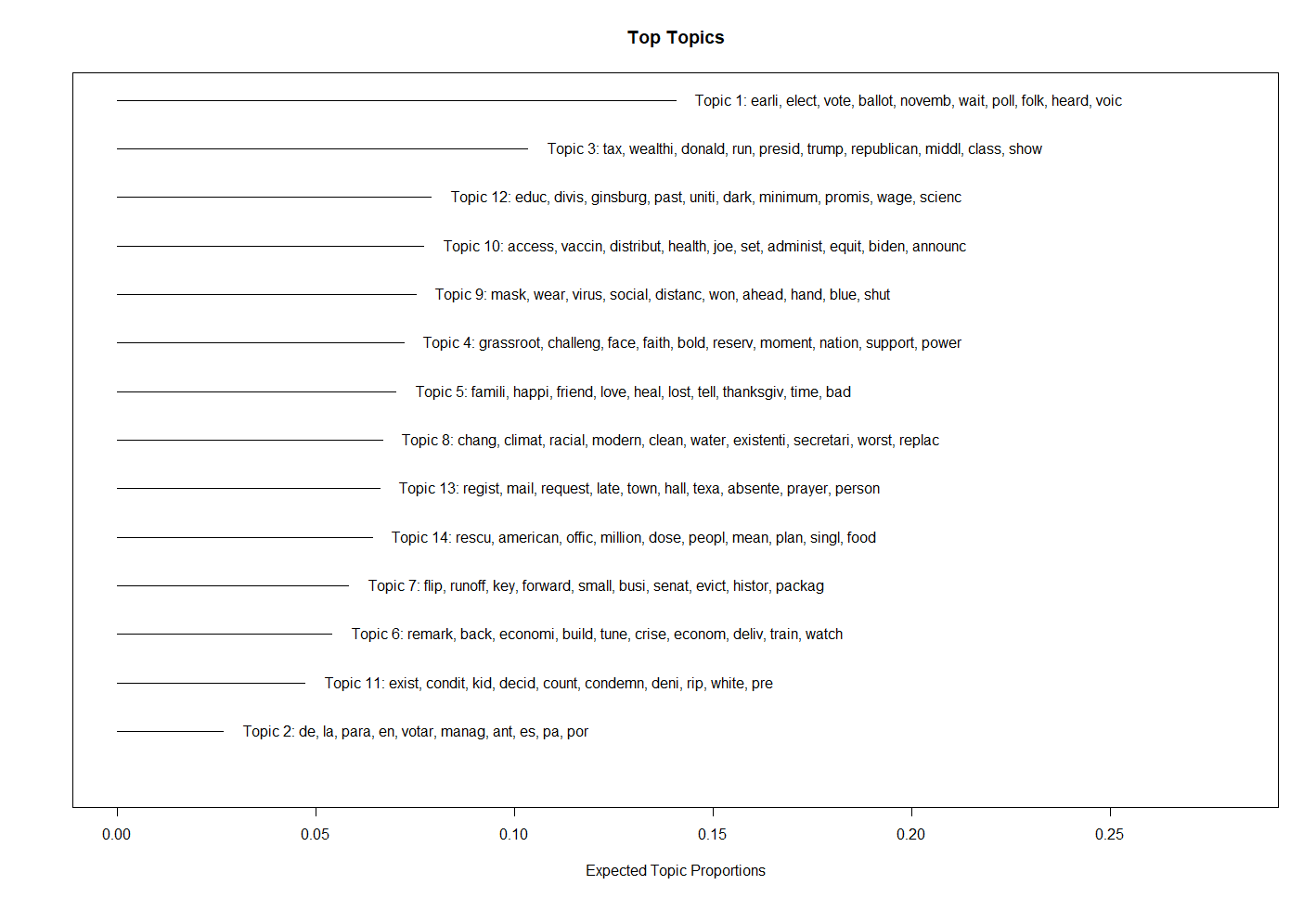
First, we ran an LDA and after estimated the optimal number of K for further STM calculations. As plotted below, K = 14 indicates the high exclusivity while still showing high values on the semantic coherence.

Figure 1

## Results with and without topic prevalence covariates



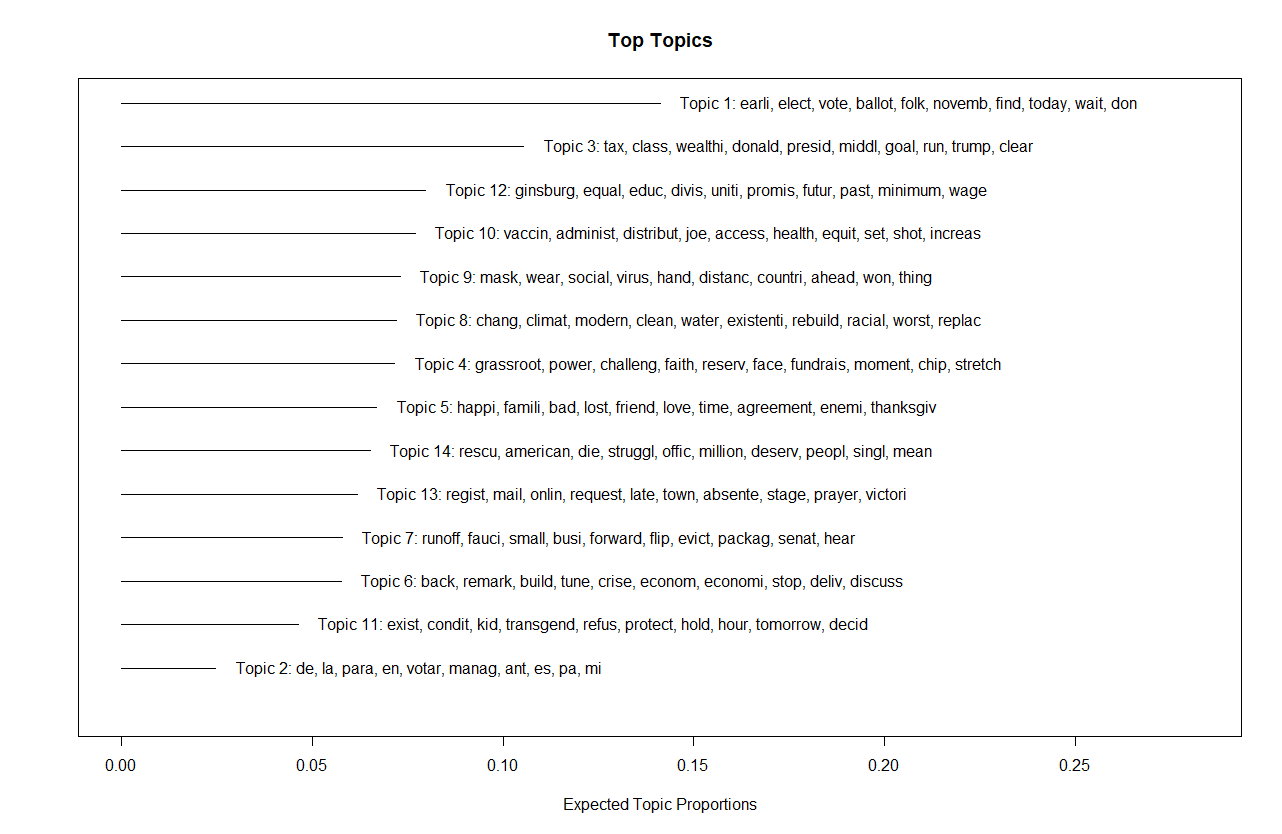
Figure 2:

Figure 3:

# Semi-supervised classification

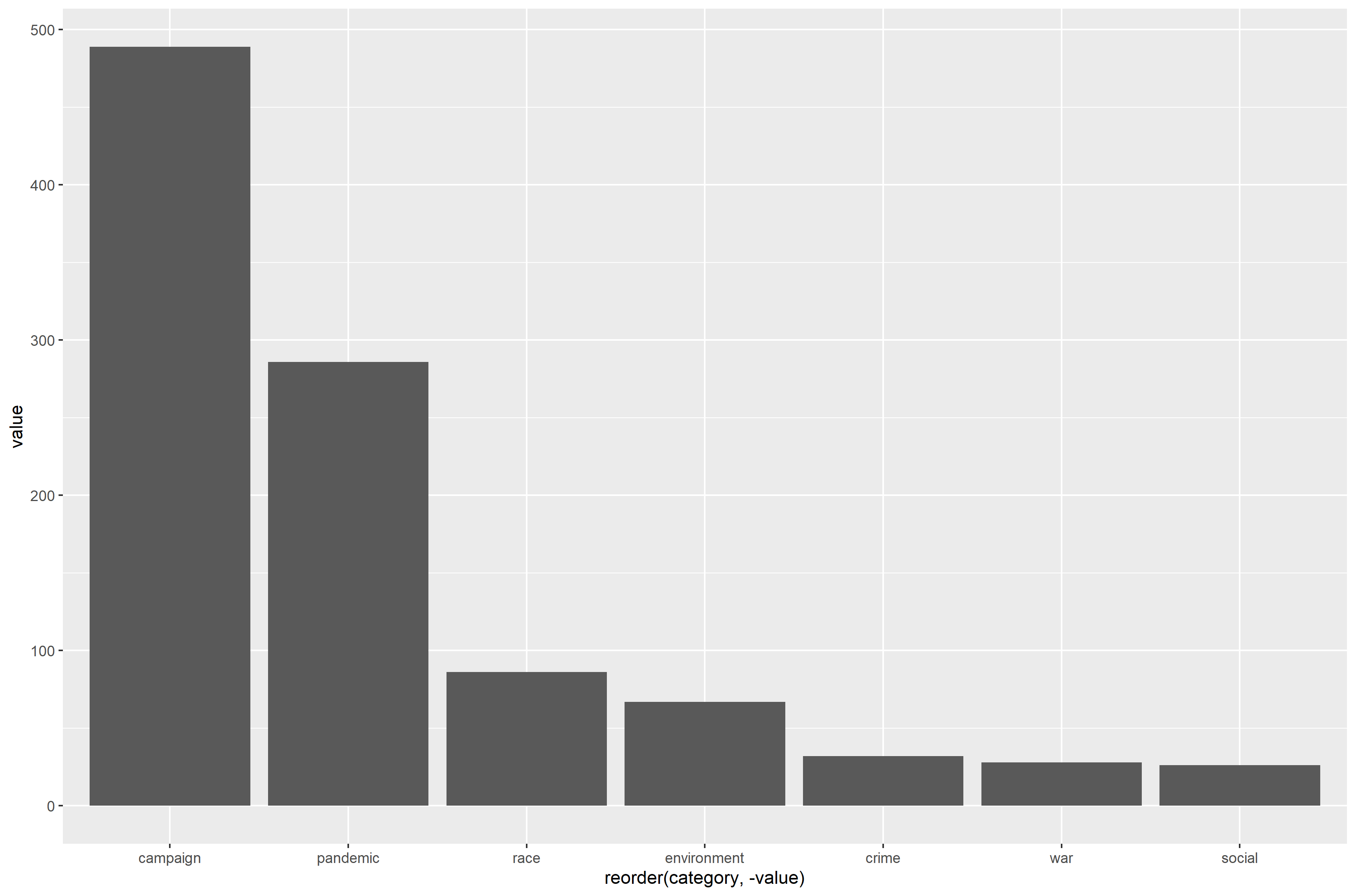


Figure 4

We analyzed the latest 1’500 tweets from Joe Biden. For semi-supervised classification we created a dictionary as stated in the code with the following categories: campaign, pandemic, race, environment, crime, war and social.

In total 990 tweets were assigned to the suggested categories, therefore in 510 tweets none of the words from our dictionary were found.

The tweets cover the time period from September 2020 until April 2021. Taking this into account it makes sense, that campaigning was a very prevalent topic. Obviously, the pandemic is also a key topic in general politics these days, therefore also in the tweets from Joe Biden.

As a critic of our dictionary would be our composition of the social dictionary words. Due to a lack of knowledge of specific key words in this topic, this category may be more prevalent that depicted in our graph.

# Naïve Bayes - SVM - Random forest - Gradient Boost

# Word embedding - WE-estimates and ML

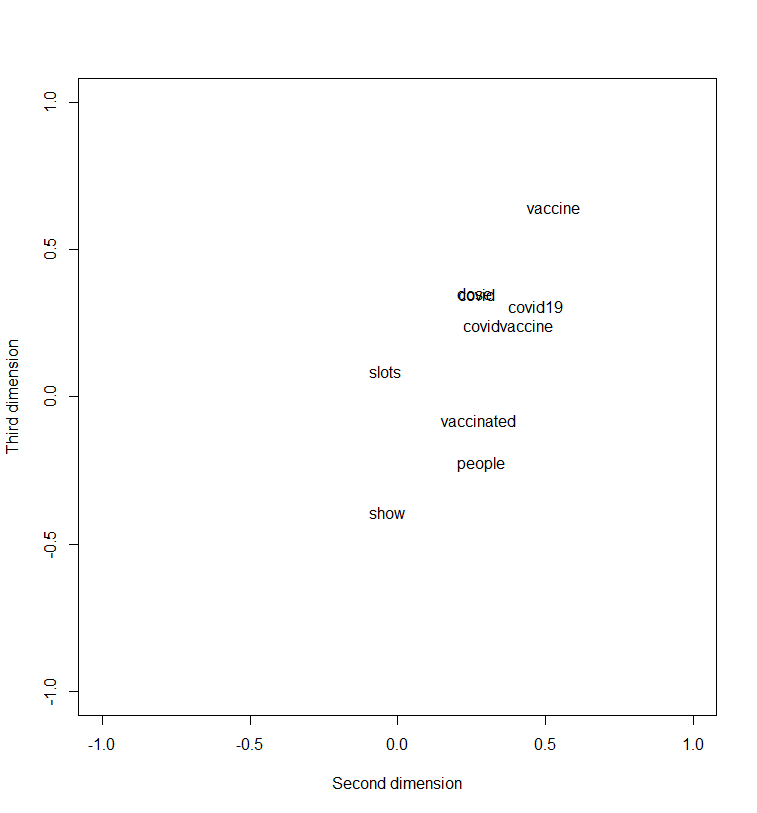
****After we have run a local WE, we plotted the most frequent words from our twitter data. We see some clustering. The words dose, get, covid19 and covidvaccine are close to each other, also vaccinated people, but it is still quite hard to give a clear interpretation. “Vaccinated” and “people“ are close to each other, so a good interpretation might be, that the people should be vaccinated. Another tight relationship is “dose” and “get” with the interpretation of “get your dose of covidvaccine”. In this case we use the cosine distance (the greater the value the less the “intense” of the relationship). There is also another approach to measure - the “relationship” co-occurence - the Euclidian distance who refers to the xy-coordinates of each point. Moreover, we have 100 dimensions and just plot the “relationship” on the second and third one.

Figure 5: Most frequent words and their relationship in terms of cosine.

We also have discovered that the word fear seems to have a the most close association with the word “covid”, “Merachant”, “Westminster” and “thrombosis”. A plausible interpretation might be only given for covid and thrombosis. Fortunately, there is no tight relationship in terms of cosine for “fear” and “vaccine”. Thus, we can assume that in these Twitter statements most of the authors are not afraid of the vaccine, at least do not obviously disclose it.







