Assignments - BIG DATA ANALYTICS

text analysis

Greta Kurpicz & Dominik Walter

FS 2021

University of Lucerne

Professor

Luigi Curini

Table of contents

List of Figures

1 Assignment - Wordfish & Wordscore

```
Call:
textmodel_wordfish.dfm(x = myDfm, dir = c(4, 11))

Estimated Document Positions:

Estimated Feature Scores:
     conserv manifesto    invit    join  govern britain  countri    bond   peopl  strong   sens  nation purpos
beta -0.3029   -0.5594  0.16420 -0.3118 -0.2288 0.09566 -0.07142 -0.0673 -0.2102 -0.3853 0.1910 0.08052 0.4373
psi   3.2212    2.6077 -0.04805  1.5226  4.3925 4.28403  3.93646 -0.1496  4.7599  2.7328 0.9316 4.18098 1.4298


       fray  polit  system  betray problem  overcom    pull      clear   today challeng    face  immens economi overwhelm     debt  social  fabric
beta -0.4992 0.1937 0.01131  0.6837 0.03741 0.009254  0.4356 -0.1018 -0.4047  -0.3507 -0.1573  0.5908  -0.307    0.4508 -0.01656 -0.0196  0.4823
psi  -2.6709 3.1040 3.87730 -1.2976 2.55198 0.170056 -0.1792  2.2831  1.5543   2.6509  2.4263 -1.0267   3.643   -0.2791  2.12272  3.4327 -0.3949
```

Figure 1: Wordfish - Position.

We can see that the parties' positions are in line with the scale. We may assume that it makes sense if we have any prior experience. The Labour Party is on the left (the most pessimistic side of the scale), while UKIP is on the right (the most positive side of the scale). It's also understandable that the Liberals are more left-wing than the Conservatives. What's odd is that although the Conservatives and Liberals had almost identical scores in 2010, there was a significant difference five years later. The Liberal Party shifted to the left, while the Conservative Party shifted somewhat to the left. Since both UKIP and Labour pushed closer to the center in 2015, it seems that they sought to form alliances with the respective middle parties. In 2017, the trend is reversed. With the exception of the Conservatives, all parties return to their initial positions. The Conservative Party continues to move to the right. Throughout the whole time, one thing is clear: the UKIP is a long way from the other parties. The ratings of the words make some sense as well.

For eg, economics is a crucial topic for the middle parties, as evidenced by the score. betray has a fairly high ranking, which is understandable considering how right-wing parties typically talk. However, there are a few that aren't so obvious. Words like social, for example, would seem to be more on the left than they are. Overall, there are a few things that make sense, and it's fascinating to see how the parties presumably moved forward.

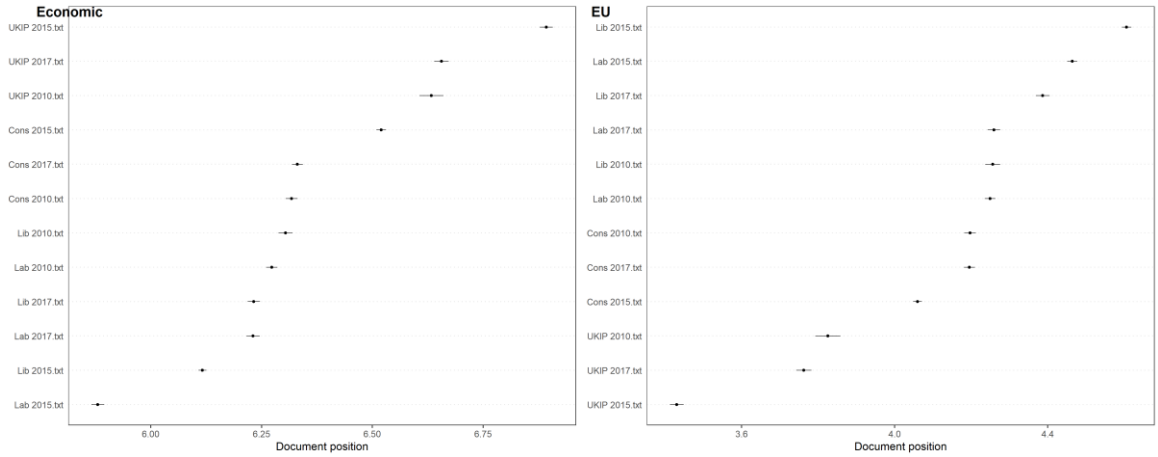

Figure 2: Comparison of the political position in terms of economic and policy of the EU.

The wordscore method estimates the political positions very accurately due to the economic dimension. All the labor parties are on the left side which correspoonds to the political orientation (left. A similar can be seen on the right side of the figure but in the opposite direction. The smaller the score the more likely is a party conservative (right).

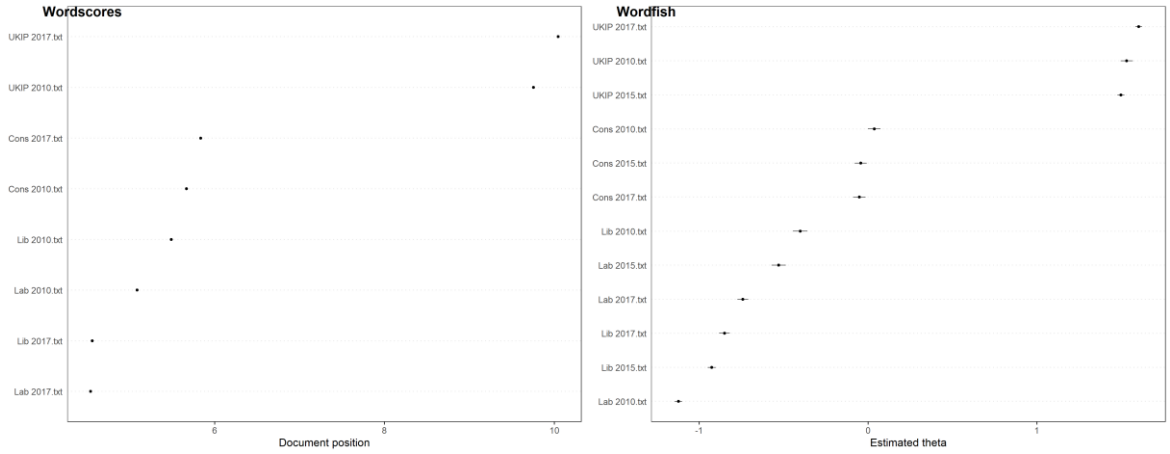

Figure 3: Wordfish - Wordscore

The figure above tells us a similar story as already mentioned, only with the difference, that the right plot is generated by wordfish.

2 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.

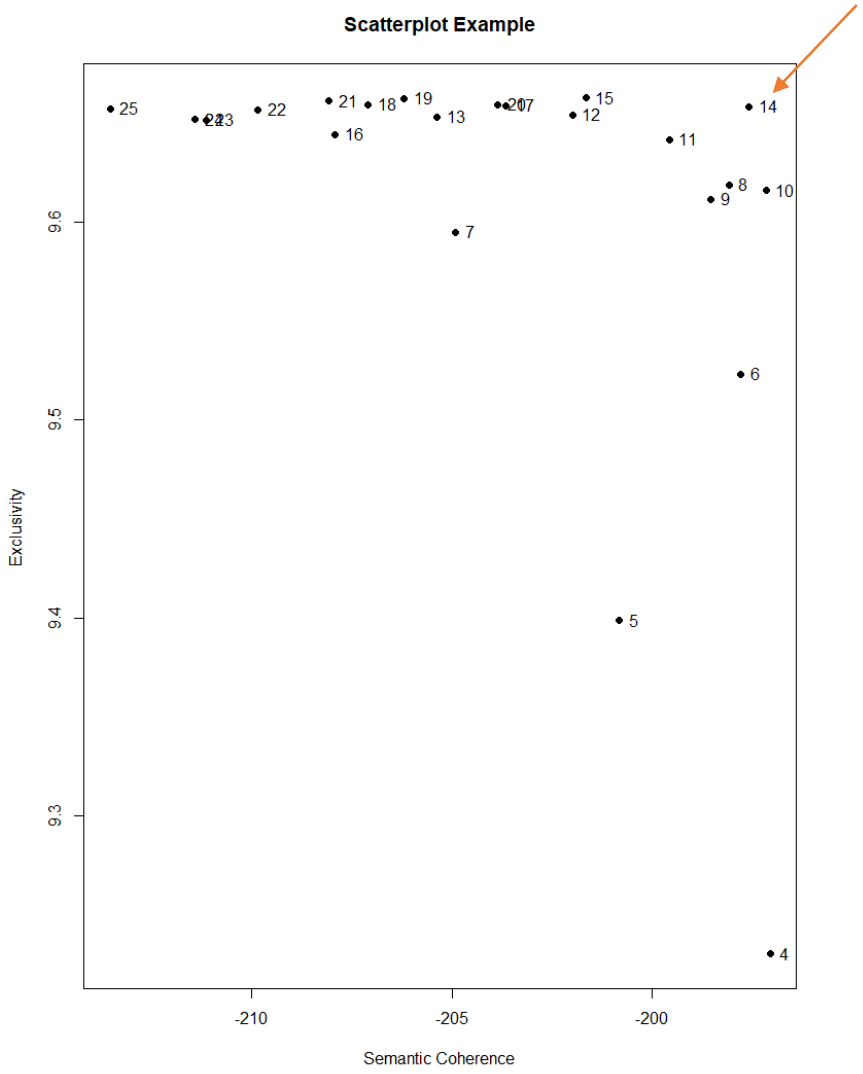

Figure 4: Scatterplot Example, K = 14 is the "best" option.

2.1 Results with and without topic prevalence covariates

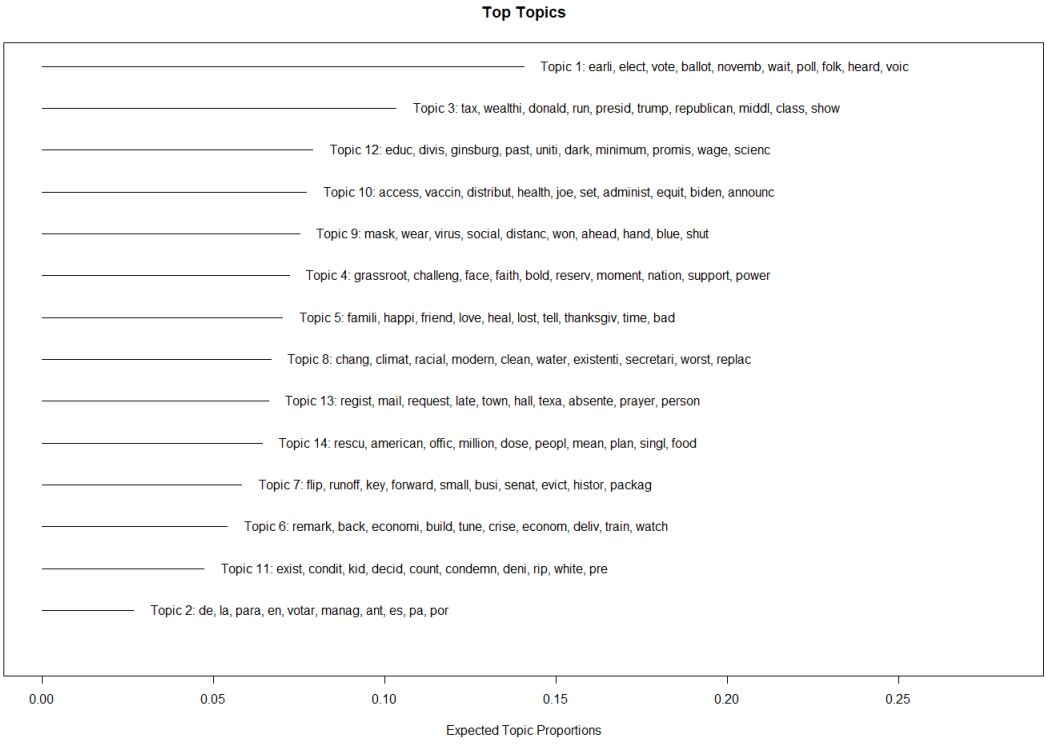

Figure 5: With topic prevalence.

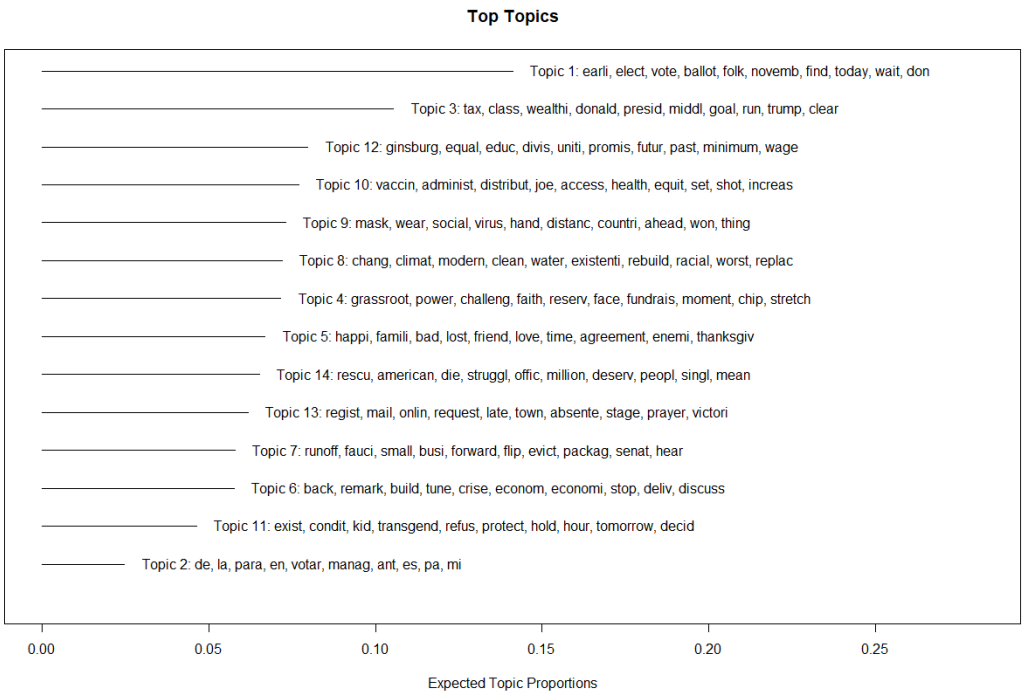

Figure 6: Without topic prevalence.

3 Semi-supervised classification

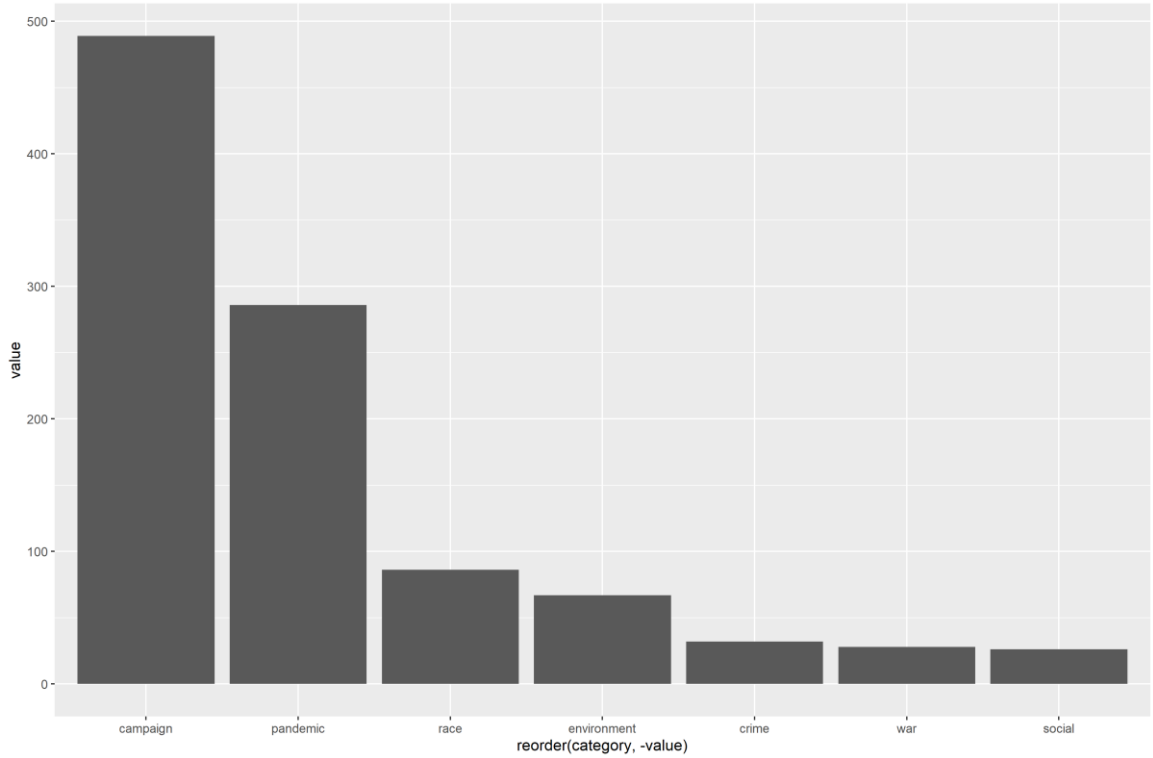

Figure 7: Most "frequent" and "important" words from Joe Bidens last 1'500 word on his twitter account.

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.

4 Naïve Bayes - SVM - Random forest - Gradient Boost

5 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,

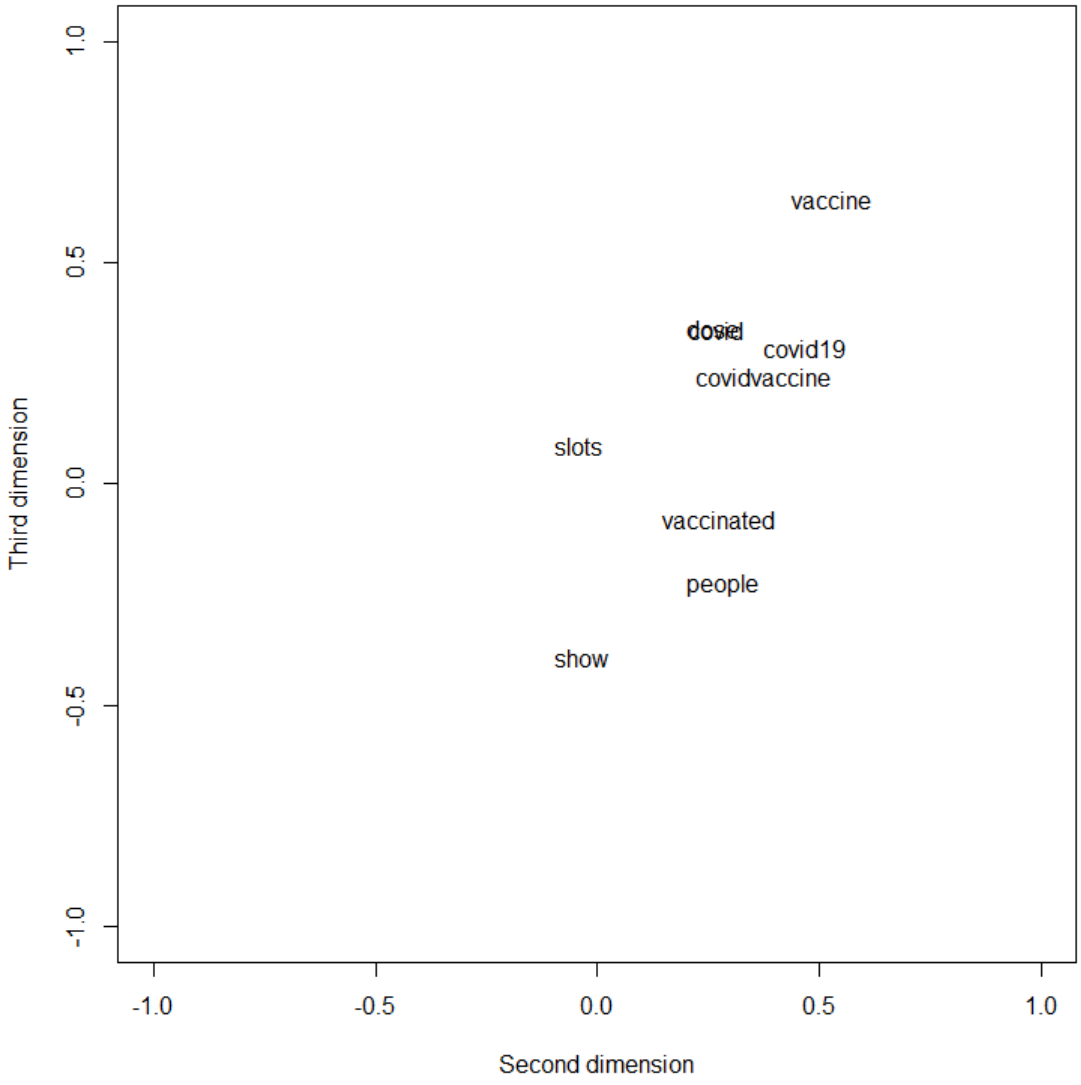

Figure 8: Most frequent words and their relationship in terms of cosine.

we have 100 dimensions and just plot the “relationship” on the second and third one.

We also have discovered that the word fear seems to have a the closest 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.

```
> head(sort(cos_sim[, 1], decreasing = TRUE), 5) #
      fear      covid   Merchant Westminster  thrombosis
 0.6215548  0.4305709  0.3779535  0.3338178  0.3292573
```