Quanteda and Twitter 2

Rodrigo Esteves de Lima-Lopes
State University of Campinas
rll307@unicamp.br

Contents

1 Introduction

Quanteda is a package for managing and analyse text quantitatively. It is quite easy to use and will bring us a number of interesting functions.

1.1 You will need:

1. The package `Quanteda`, which can be installed using RStudio
2. The package `rtweet`, we installed last tutorial.
3. Package `DT` for viewing the KWIC inside R.
4. `Ggplot` for viewing some graphics

1.2 Data

We are going to use the same data we have used in the previous tutorials

2 Making some comparissions

First we need to save our texts as vectors

```r
boulos.v <- boulos$text
covas.v <- covas$text
```

Kwic in the terminal

```r
kwic(boulos.v,"prefeitura")
```

Kwic in the Viewer (it views but does not save)

```r
kwic(covas.v,"prefeitura")%>%View()
```

Now, saving as a variable

```r
prefeitura.b <- kwic(boulos.v,"prefeitura")
```

This command can also be applied to a corpus:

```
kwic(boulos.corpus,"prefeitura")%>%View()
```

Now lets keep on our analysis. We can use the same DFMs we created before, but I will create a couple just to apply some new commands

First step now is to make all lower caps:

```
boulos.lower.v <- char_tolower(boulos.v)
covas.lower.v <- char_tolower(covas.v)
```

Now, we create a lower character vector

```
boulos.word.v <- tokens(boulos.lower.v,
                        remove_punct = TRUE) %>% as.character()
covas.word.v <- tokens(covas.lower.v,
                       remove_punct = TRUE) %>% as.character()
```

So now, we have our second DFM:

```
boulos.dfm.2 <- dfm(boulos.lower.v,
                    remove_punct = TRUE,
                    remove = stopwords("portuguese"))


covas.dfm.2 <- dfm(covas.lower.v,
                   remove_punct = TRUE,
                   remove = stopwords("portuguese"))
```

Now we are creating our word list

```
boulos.wl<-textstat_frequency(boulos.dfm.2)
View(boulos.wl)

covas.wl<-textstat_frequency(covas.dfm.2)
View(covas.wl)
```

Lets us plot a single candidate at a time:

```
library(ggplot2)
theme_set(theme_minimal())
textstat_frequency(covas.dfm.2, n = 50) %>%
  ggplot(aes(x = rank, y = frequency)) +
  geom_point() +
  labs(x = "Frequency rank", y = "Term frequency")
```

```
theme_set(theme_minimal())
textstat_frequency(boulos.dfm.2, n = 50) %>%
  ggplot(aes(x = rank, y = frequency)) +
  geom_point() +
  labs(x = "Frequency rank", y = "Term frequency")
```

The results are:

Quanteda makes key wordlists. Keywords are a comparison between two groups of texts. One is the reference and tells me the statistical baseline for comparison, the other is my research text (or texts). Here, for the sake of the exercise, we will compare Boulos and Covas.

```
all.dfm3 <- dfm(all.corpora, groups = "screen_name", remove = stopwords("portuguese"),
                remove_punct = TRUE)
```
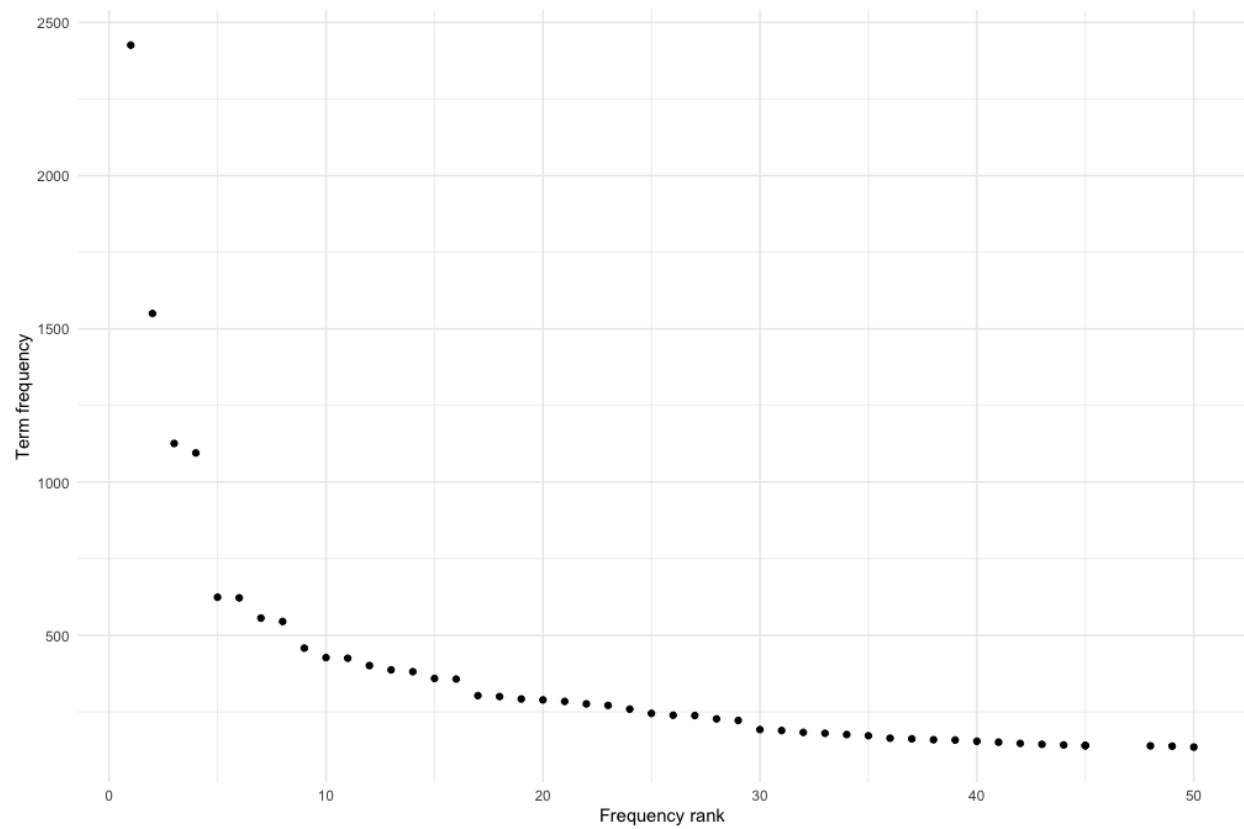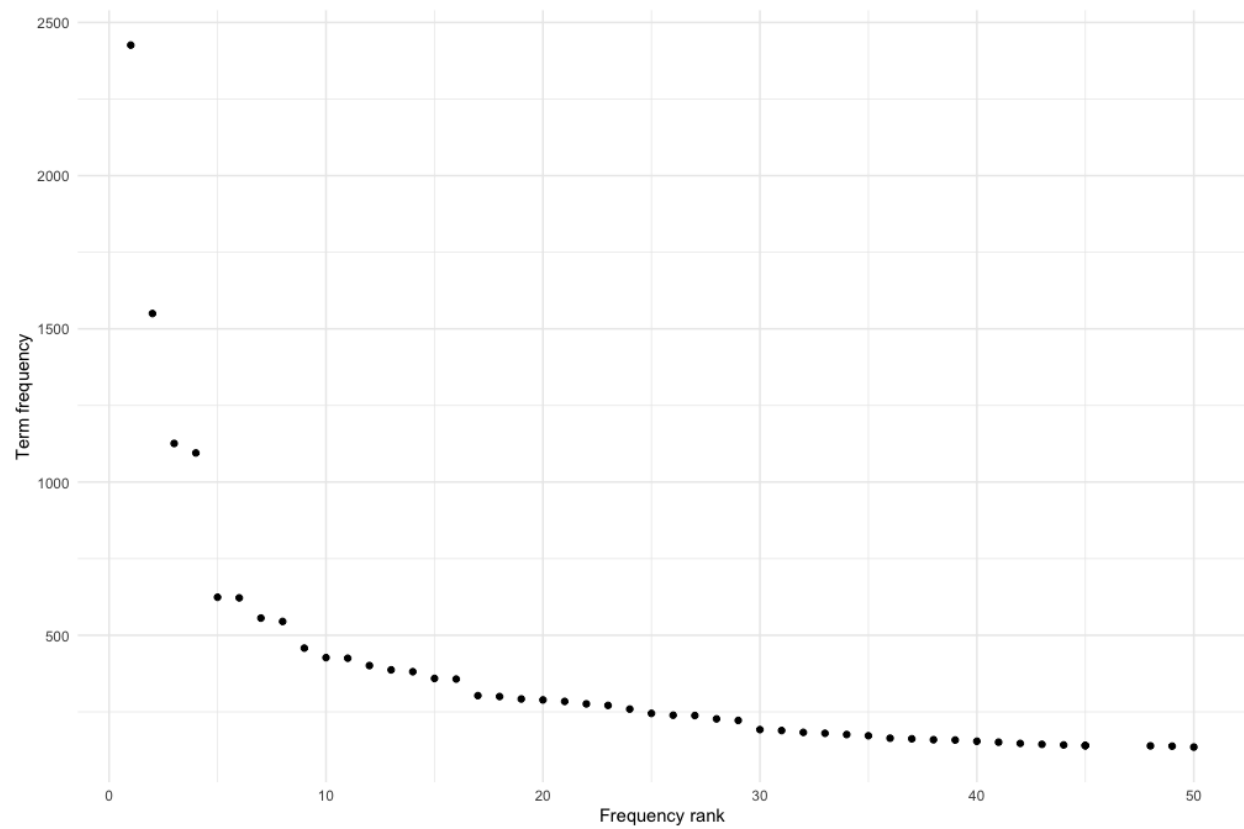
Figure 1: Covas' wordlist

Figure 2: Boulos' wordlist

```
keyness <- textstat_keyness(all.dfm3, target = "brunocovas")
textplot_keyness(keyness)
```
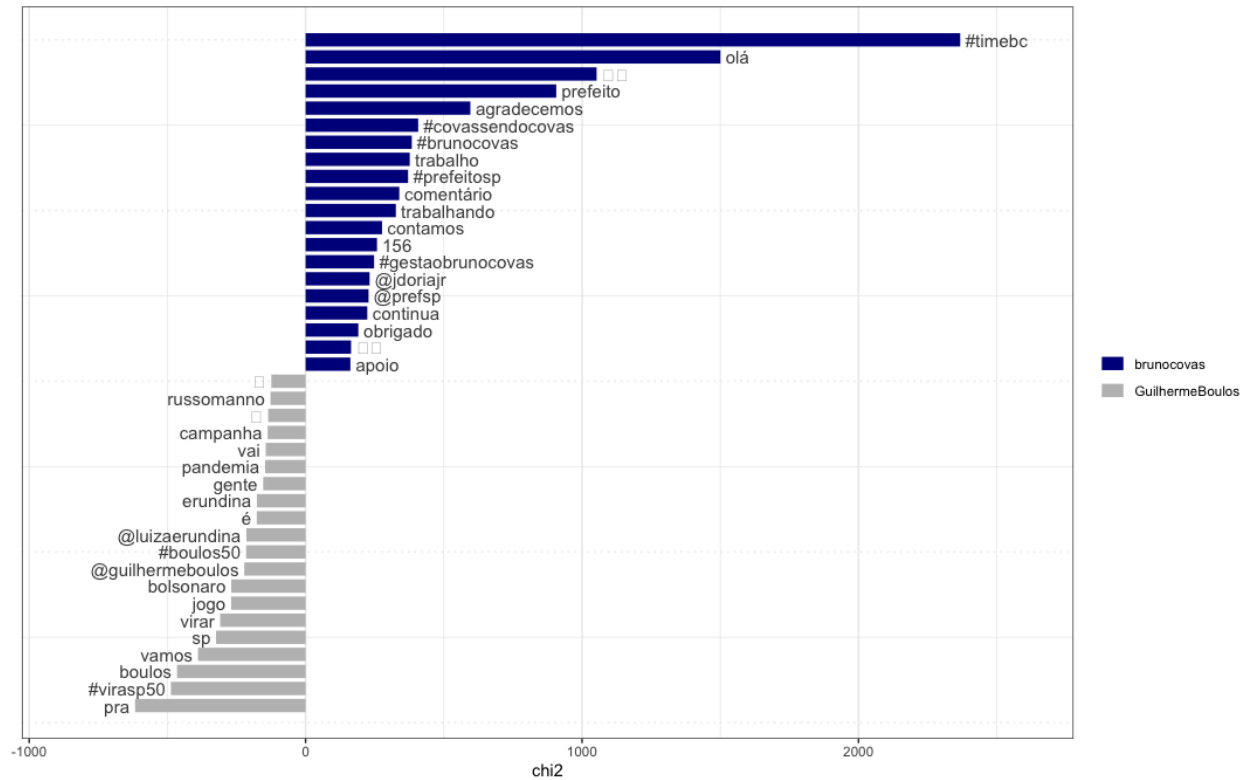
The result is something like

Figure 3: Keywords comparisson