Tagging text data

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

Contents

1 Introduction

In this tutorial, we are going to use a `NLP` package called UDPipe in order to tag a number of texts using POF (par of speech) methodology. We are going to recycle some data from our last tutorial (Lula's interview by Roda Viva).

2 Packages

In this tutorial we will need the following packages:

```
# Packages
library(ggplot2)
library(udpipe)
library(textrank)
library(dplyr)
library(forcats)
```

3 Analysing the texts

After I load the packages above, my first step is to download the models for text analysis. Here we are going to work with Brazilian Portuguese.

```
#Downloading the model
ud_model <- udpipe_download_model(language = "portuguese")
ud_model <- udpipe_load_model(ud_model$file_model)
```

Our next step is to load the data we used last tutorial

```
# Loading the interviews
load('./data/02_scrape_html.RData')
```

Our next step is tagging the corpus. In the command below:

- `ud_model` is the model we just downloaded and loaded.
- `x` is the interviews data frame, please note we are using only the column text.
- `doc_id` is a filed to identify the texts. Here we are using the title of the interviews, but it might be any info you see fit
- `as.data.frame()` is to save the data in a more friendly format

```
Tagged.Interviews <- udpipe_annotate(ud_model,
                          x=base$Text,
                          doc_id = base$Title) %>% as.data.frame()
```

3.1 Which are the most frequent verbs?

```
Verbs <- subset(Tagged.Interviews, upos %in% c("VERB"))
Verbs  <- txt_freq(Verbs$lemma)
Verbs.top <- Verbs[1:25,]
```

Plotting

```
Verbs.top%>%
  mutate(key = fct_reorder(key, freq_pct)) %>%
  ggplot(., aes(x = key, y = freq_pct, fill=key)) +
  geom_segment(aes(x=key, xend=key, y=0, yend=freq_pct),
               color="black",size = 2) +
  geom_point(shape = 21, color="black",
             fill ='#FF0000',
             size=14, stroke = 1) +
  theme_light() +
  labs(caption="Fonte: Dados",
       x = "Verbos",
       y = "Frequência Normalizada (100)") +
  theme(axis.text.x = element_text(angle=40, vjust=0.6),
        legend.position = "none",
        panel.grid.major.x = element_blank(),
        panel.border = element_blank(),
        axis.ticks.x = element_blank(),
        text = element_text(size=30))
```

The result should be like:

3.2 Relationship (verbs and nouns)

```
# Observing verbs+nouns
sub_verb <- keywords_phrases(x = Tagged.Interviews$upos,
                             term = tolower(Tagged.Interviews$token),
                             pattern = "NOUN+VERB",
                             is_regex = TRUE,
                             detailed = FALSE)
```

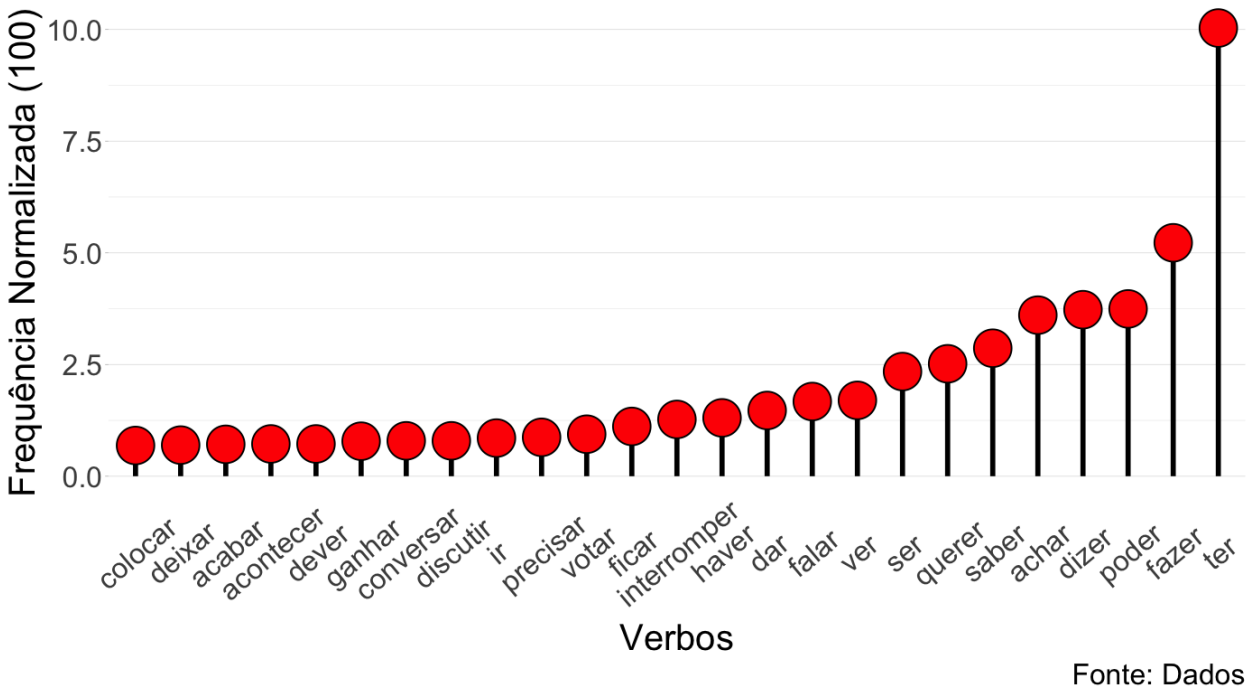

Figure 1: Verbs

3.3 Observing (nouns and adjectives)

```
sub.adj <- keywords_phrases(x = Tagged.Interviews$upos,
                            term = tolower(Tagged.Interviews$token),
                            pattern = "NOUN+ADJ",
                            is_regex = TRUE,
                            detailed = FALSE)
```

3.4 Collocates

```
collocates<-collocation(Tagged.Interviews,
                        term="lemma",
                        group="doc_id",
                        ngram_max = 3,
                        n_min = 2,
                        sep = " ")
```

3.5 Matrix and correlation

```
x <- document_term_frequencies(Tagged.Interviews[, c("doc_id", "lemma")])
dtm <- document_term_matrix(x)
correlation<-dtm_cor(dtm)
```