Wodlist and concordancing: Lexical graduation in the word 'mito' in YouTube Users Comments

Rodrigo Esteves de Lima-Lopes
State University of Campinas | rll307@unicamp.br
Carolina Palma de Sousa Arruda
State University of Campinas | carolpalma203@gmail.com

Contents

1 Packages

We are going to need the following packages for making a simple wordlist:

```
library(tidytext)
library(tidyr)
library(dplyr)
library(quanteda)
```

2 Cleaning and processing the wordlist

First we will extract the comments from the original data frame

```
string.comments <- as.character(comments$textOriginal)
```

Next step is to clean numbers and special characters

```
string.comments <-str_replace_all(string.comments, "[^[:alnum:]]", " ")
```

We are going to use a list of stopwords provided by the package **quanteda**, so no need to import. Probably some users names will pop up, you add them to the list.

```
my.stopwords <- data.frame(stopwords("pt"))
colnames(my.stopwords)<-"words"
```

Now let us create a data frame so we can process the wordlist.

```
geral.list.df <- data.frame(text = string.comments, stringsAsFactors = F)
colnames(geral.list.df) <- "text"
```

Now we make the wordlist *per se*:

```
geral.list.df <- geral.list.df %>%
  unnest_tokens(word, text, to_lower= FALSE) %>%
  count(word, sort = TRUE) %>%
  anti_join(my.stopwords)%>%
  mutate((freq = n / sum(n))*100) %>%
  arrange(desc(n))
colnames(geral.list.df)<-c('word','n','freq')
```

3 Concordancing

I would suggest `quanteda`'s `kwic` command. It simplifies a lot the concondancing process. Please, substitute the term **word** by you research interest.

```
corpus.comments <-corpus(comments,text_field = 'textOriginal')
View(kwic(corpus.comments, pattern = "WORD",case_insensitive=FALSE))
```