Vagas_trans

September 29, 2019

1 Lexical analysis

The purpose of this notebook is to maka available the methodological steps I used in the following article:

Lima-Lopes, RE de. **The Reaction to Social Quotas: A study of Facebook Comments in Brazilian Portuguese**.

The paper was submitted to a major Brazilian journal and the referece will be updated when it is published.

1.1 Dada Colection

Data was collected using the software Netvizz used to scrape data form Facebook's pages and communities. The software could get information such as posts, comments on posts, general statistics of a page and posts in a given period. It only worked with pages that have set their status as public and, by default, anonymises usernames field as it generates a *.tab file. Today the software is discontinued since Facebook has had a more conservative data scrape policy.

1.2 Objective

- To study grammatical patterns in users comments on UFBA's annoucemment that its social quota programme would include imigrants, refugees and transexual people.
- It is believed that the analysis of lexis might revel some interesting characteristics of the discourse of this comments

1.3 Code

[ ]:
```
#Packages
library(tm)
library(plyr)
library(readtext)
library(RColorBrewer)
library(FactoMineR)
library(ggplot2)
library(readr)
library(tidyverse)
library(quanteda)
library(ggplot2)
```

[ ]:
```
#Stopwords
my.stopwords <- read_csv("stop_port2.csv")
```

[ ]:
```
#Reading the comments files (Netvizz would provide \*tab delimited tables
## It generated 4 files, one for each time data was collected. The number of␣
 ↪files represent the number of times I had to collect the comments.
## Restrictions on number of comments were part of Netvizz/Facebook interacion
comentarios_01 <- read_csv("comentarios_01.csv")
comentarios_02 <- read_csv("comentarios_02.csv")
comentarios_03 <- read_csv("comentarios_03.csv")
base1 <- rbind(comentarios_01,comentarios_02,comentarios_03)
```

[ ]:
```
#Creating a corpus using *tm*
text_string <- as.character(base1$comment_message)
corpus.cluster <- Corpus(VectorSource(text_string))
corpus.cluster <- tm_map(corpus.cluster, content_transformer(tolower))
removeURL <- function(x) gsub("http[[:alnum:][:punct:]]*", "", x)
remove.users <-function(x) gsub("@[[:alnum:][:punct:]]*","",x)
corpus.cluster <- tm_map(corpus.cluster, content_transformer(removeURL))
corpus.cluster <- tm_map(corpus.cluster,content_transformer(remove.users))
corpus.cluster  = tm_map(corpus.cluster, stripWhitespace)
corpus.cluster <- tm_map(corpus.cluster, removePunctuation)
corpus.cluster <- tm_map(corpus.cluster,
                         function(x)removeWords(x,c(stopwords("pt"),stopport)))
```

[ ]:
```
#Creating a matrix od terms
cluster.tdm <- TermDocumentMatrix(corpus.cluster)
#Deleting sparse words
cluster.df <- as.data.frame(inspect(cluster.tdm))
#Converting the corpus to a matrix
cluster.m <- as.matrix(cluster.tdm)
```

[ ]:
```
cluster.wf <- rowSums(cluster.m)
#Deleting sparse words (90%)
cluster.m1 <- cluster.m[cluster.wf>quantile(cluster.wf,probs=0.99), ]
#Revmoning 0 columns
cluster.m1 <- cluster.m1[,colSums(cluster.m1)!=0]
#Creating binary relationships
cluster.m1[cluster.m1 > 1] = 1
cluster.m1dist = dist(cluster.m1, method="binary")
```

[ ]:
```
#creating the colour dendogram
dend <- as.dendrogram(clus1)
labelColors <- c("#809acd", "#000000", "#EB6841", "#666666","#80cdb3",
                 "#c5ab8a","#ffa500","#0000ff", "#523415", "#b882ee")
clusMember <- cutree(clus1, 10)
colLab <- function(n) {
```

```
  if (is.leaf(n)) {
    a <- attributes(n)
    labCol <- labelColors[clusMember[which(names(clusMember) == a$label)]]
    attr(n, "nodePar") <- c(a$nodePar, lab.col = labCol)
  }
  n
}
clusDendro = dendrapply(dend, colLab)
plot(clusDendro,cex=0.9)
rect.hclust(clusDendro,k=2)
```

After texts we processed in a dendogram, comments were classified in four types: - In favour of the quotas system - Against the quotas system - Interaction amongst users - Discrimination and racism against Northeast Brazilian citizens

[ ]:
```
#Visualizing the number of word in each kind of discourse

polaridade.raw <- read.csv("polaridade_geral.csv", stringsAsFactors = FALSE,␣
 ↪fileEncoding = "UTF-8")
View(polaridade.raw)

##Sleecting columns
polaridade.raw <- polaridade.raw[, 1:2]
names(polaridade.raw) <- c("Label", "Text")
View(polaridade.raw)

##Converteing classes in values
polaridade.raw$Label <- as.factor(polaridade.raw$Label)

#Observing the value of each theme
prop.table(table(polaridade.raw$Label))
polaridade.raw$TextLength <- nchar(polaridade.raw$Text)
summary(polaridade.raw$TextLength)

#plotting

ggplot(polaridade.raw, aes(x = TextLength, fill = Label)) +
  theme_bw() +
  geom_quantile()
  labs(caption="Source: Data", y = "Text Count", x = "Length of Text")
```

[ ]:
```
#Creating teh corpus for concordancing
trans <-corpus(text_string)
trans.tokens <- tokens(trans, remove_punct = TRUE,
                       remove_numbers = TRUE, remove_url = TRUE)
#general command for concordancing
x.kwic <- kwic(trans, pattern = "x.*", window = 25,
```

```
                    case_insensitive=TRUE, valuetype = "regex")
```