Tutorial 10 - Text data

Ariane Ducellier

2024-11-21

```
Load R packages.
library(corrplot)
## corrplot 0.94 loaded
library(RColorBrewer)
library(tm)
## Loading required package: NLP
library(wordcloud)
Make a word cloud
 # Set number of colors and palette
pal = brewer.pal(6, "RdGy")
# Choose minimum frequency and the range of the size of the words
wordcloud("The objective of this course is to provide students with a comprehensive understanding of da
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation): transformation
## drops documents
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
                                                                                                           fundamentals
                                                                                               visual practical
                                    communicate perception focus - dacht
              using provide decisions cognition
right explore networks exercises better also objective projects
better also objective projects
better also objective projects

and objective
                   ractivity shiny main Oata word user best course through earth of e
interactivity shiny main data
      clear the enabling geographic through types three colors popular static design learning chart creating
                                                                                                                                                                                                            mprehensive
COVEL
                                 make principles maps choosing tices stories packages effective effectively understanding visualize coordinates techniques compelling
```

USE visualizations topics

To use a list of words and their frequencies.



To read a text file and preprocess it, before doing the word cloud.

```
file = readLines("../../data/syllabus.txt")
doc = Corpus(VectorSource(file))
doc = tm_map(doc, tolower)
## Warning in tm_map.SimpleCorpus(doc, tolower): transformation drops documents
doc = tm_map(doc, removePunctuation)
## Warning in tm_map.SimpleCorpus(doc, removePunctuation): transformation drops
## documents
doc = tm_map(doc, removeNumbers)
## Warning in tm_map.SimpleCorpus(doc, removeNumbers): transformation drops
## documents
doc = tm_map(doc, removeWords, stopwords("english"))
## Warning in tm_map.SimpleCorpus(doc, removeWords, stopwords("english")):
## transformation drops documents
wordcloud(as.character(doc), scale=c(2, 0.5))
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation): transformation
## drops documents
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,
## tm::stopwords())): transformation drops documents
         use of inter
               interactive
       visualizations
          principles
```

Make two word clouds

```
files = DirSource("../../data/debate/")
data = Corpus(DirSource("../../data/debate/"))
data = tm_map(data, tolower)
data = tm_map(data, removePunctuation)
data = tm_map(data, removeNumbers)
data = tm_map(data, removeWords, c(stopwords("english"), "biden", "trump"))
data = TermDocumentMatrix(data)
data = as.matrix(data)
colnames(data) = c("biden", "trump")
comparison.cloud(data, max.words=100, title.size=2, colors=c("blue", "red"))
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : obamacare could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : judges could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : november could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : places could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : radical could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : never could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : knows could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : together could not be fit on page. It will not be plotted.
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : thing could not be fit on page. It will not be plotted.
```

```
counted create economyamerica biden president discredited doesn't american totally by guy people sure united totally by an able deal's voteget will jobstrue going make plantax on manable deal's voteget united totally by an able deal's voteget will jobstrue going make plantax on manable deal's voteget united totally by an able deal's voteget will jobstrue going make plantax on the plantax on th
```

c("blue", : seen could not be fit on page. It will not be plotted.

Plot correlations between texts

```
data(crude)
data = tm_map(crude, content_transformer(tolower))
data = tm_map(data, removePunctuation)
data = tm_map(data, removeNumbers)
data = tm_map(data, removeWords, stopwords("english"))
data = TermDocumentMatrix(data)
data = as.matrix(data)
crf = cor(data)
corrplot(crf, method = c("ellipse"))
```

