

STAT 451 - Visualizing Data - Autumn 2025

Ariane Ducellier

11/20/2025

Visualizing text data

Load R packages.

```
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
library(RColorBrewer)
```

library(tm)

```
## Loading required package: NLP
```

```
library(wordcloud)
```

Make a word cloud

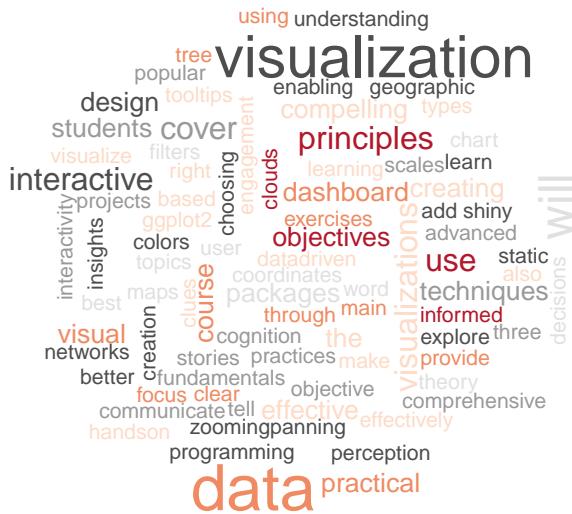
```
# Set number of colors and palette  
pal = brewer.pal(6, "RdGy")
```

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 data analysis and machine learning concepts, and to develop their skills in applying these concepts to real-world problems. The course will cover topics such as statistical methods, data mining, and machine learning algorithms, and will include practical exercises and projects to help students gain hands-on experience with these tools and techniques. By the end of the course, students will be well-prepared to pursue further studies or careers in data science, machine learning, or related fields.")
```

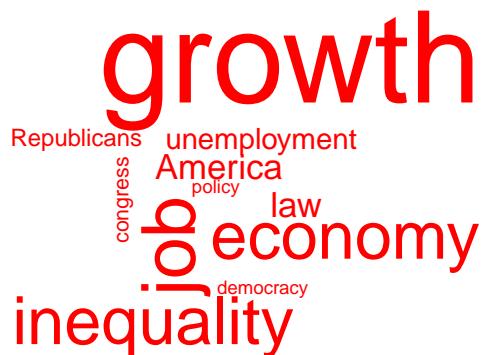
```
## Warning in tm_map.SimpleCorpus(corpus, tm::removePunctuation): transformation
## drops documents
```

```
## Warning in tm_map.SimpleCorpus(corpus, function(x) tm::removeWords(x,  
## tm::stopwords())): transformation drops documents
```



To use a list of words and their frequencies.

```
wordcloud(c("inequality", "law", "policy", "unemployment", "job", "economy", "democracy", "Republicans",
          freq=c(26, 9, 2, 7, 30, 26, 1, 4, 3, 9, 57, 9), min.freq=0, color="red")
```



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

```



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", : happened 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", : 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.

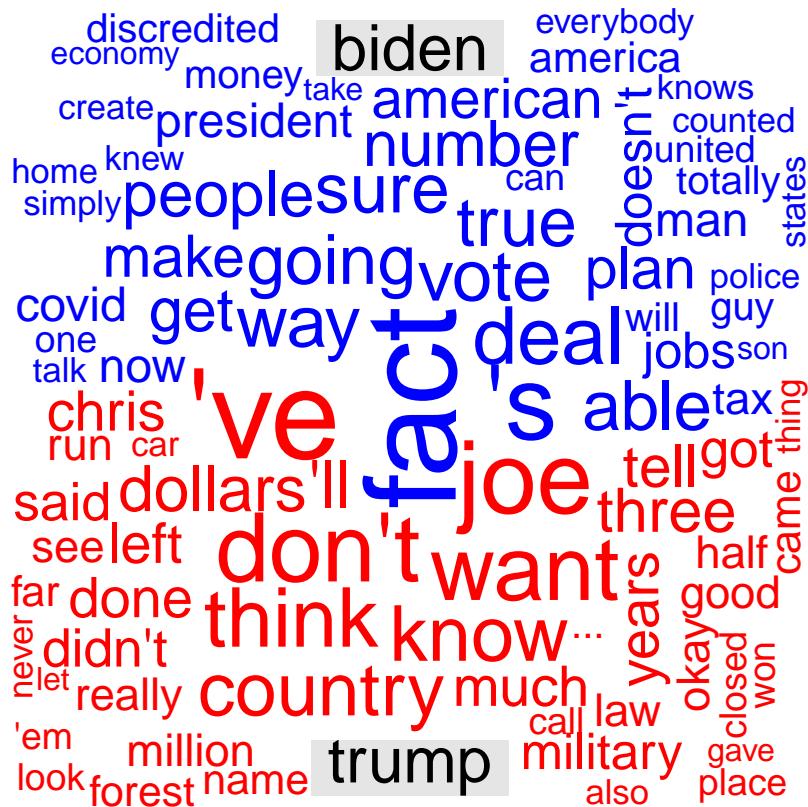
## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : together could not be fit on page. It will not be plotted.

## Warning in comparison.cloud(data, max.words = 100, title.size = 2, colors =
## c("blue", : just 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", : ever 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", : shut 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", : management 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"))
```

