Quantitative Text Analysis

Kostas Gemenis & Bastiaan Bruinsma

2021-07-21

Contents

Chapter 1

Foreword

Welcome to the Quantitative Content Analysis Textbook. This book was originally written as a collection of assignments and slides for the ECPR Winter and Summer Schools. Note that the book is still in active development.

The developments over the past 20 years have made research using quantitative text analysis a particularly exciting undertaking.

First of all, the enormous increase in computing power has made it possible to work with large bodies of text. Secondly, the development of R, a free, open-source, cross-platform statistical software has enabled many researchers and programmers to develop particular packages that implement statistical methods of working with text. In addition, the spread of the Internet has made available in digital format many interesting sources of textual data. To these, we should add the emergence of social media as a massive source of text which is generated daily, by millions of users across the world.

Yet, quantitative text analysis can be a daunting experience for someone who is not familiar with quantitative methods or programming. This book will guide you, with step-by-step explanation of the code, through a series of exercises illustrating a wide range of text analysis methods. Many of these exercises have been given to participants in the ECPR Summer and Winter Schools in Methods and Techniques whom, often, had no prior experience in text analysis, R, or quantitative methods. Therefore, we hope that you will find the exercises easy to understand but also engaging.

1.1 Installing R

R is an open-source programme that allows you to carry out a wide variety of statistical tasks. At its core, it is a modification of the programming languages S and Scheme, making it not only flexible but fast as well. R is available for

Windows, Linux and OS X and receives regular updates. In its basic version, R uses a simple command-line interface. To give it a friendlier look, integrated development environments (IDEs) such as RStudio are available. Apart from looking better, these environments also provide some extra practical features.

1.1.1 R on Windows

To install R on Windows, go to https://cran.r-project.org/bin/windows/base/, download the file, double-click it and run it. Whilst installing, it is best to leave standard options (such as the installation folder) unchanged. This makes it easier for other programmes to know where to find R. Once installed, you will find two shortcuts for R on your desktop. These refer to each of the two versions of R that come with the installation - the 32-bit and the 64-bit version. Which version you need depends on your version of Windows. To see which version of Windows you have, go to This PC (or My Computer, right-click it, and select Properties. Here you should find the version of Windows installed on your PC. If you have the 64-bit version of Windows, you can use both versions. Yet, it is best to use the 64-bit version as this makes better use of the memory of your computer and thus runs smoother. If you have the 32-bit version of Windows, you have to use the 32-bit version of R.

To install RStudio, go to https://www.rstudio.com/products/rstudio/download/, and download the free version of RStudio at the bottom of the page. Make sure to choose **Installers for Supported Platforms** and pick the option for Windows. Once downloaded, install the programme, leaving all settings unchanged. If everything works out fine, RStudio will have found your installation of R and placed a shortcut on the desktop. Whether you have the 32-bit or 64-bit version of Windows or R does not matter for RStudio. What does matter are the slashes. R uses forward slashes (/) instead of the backslashes (\) that Windows uses. Thus, whenever you specify a folder or file within R, make sure to invert the slashes. So, you should refer to a file which in Windows has the address **C:\Users\Desktop\data.csv** as **C:/Users/Desktop/data.csv**.

1.1.2 R on Linux

How to install R on Linux depends on which flavour of Linux you have. In most cases, R is already part of your Linux distribution. You can check this by opening a terminal and typing `R`. If installed, R will launch in the terminal. If R is not part of your system, run the following in the terminal:

1. sudo add-apt-repository 'deb https://cloud.r-project.org/bin/linux/ubuntu focal-cran40/'
2. sudo apt-key adv –keyserver keyserver.ubuntu.com –recv-keys E298A3A825C0D65DFD57CBB6517
3. sudo apt update
4. sudo apt install r-base r-base-core r-recommended r-base-dev

As an alternative, you can use the Synaptic Package Manager and look for the

r-base-dev and **r-base** packages. Select them, and install them.

To install RStudio, go to https://www.rstudio.com/products/rstudio/download/. At the bottom of the page, pick the installer that corresponds to your OS. Then, install the file either through an installation manager or via the terminal. After running the launcher, you can find RStudio in the Dash.

1.1.3 R on macOS

With OS X you must have OS X 10.6 (Snow Leopard) or above. Installing R otherwise is still possible, but you cannot use a certain number of packages (such as some we use here). To check this, click on the Apple icon in the top-left of your screen. Then, click on the "About This Mac" options. A window should then appear that tells you which version of OS X (or macOS) you have.

To install R, go to https://cran.r-project.org/index.html and click **Download R for (Mac) OS X**. Once there, download the .pkg file that corresponds to your version of OS X and install it. Besides, you have to download the **Clang 6.x compiler** and the **GNU Fortran compiler** from https://cran.r-project.org/bin/macosx/tools/. Install both and leave the selected options as they are. After the installation, check if R works by launching the programme.

To install RStudio, go to https://www.rstudio.com/products/rstudio/download/ and download the OSX version at the bottom of the page. After downloading and installing, you can now find RStudio with your other programmes.

1.1.4 R in the Cloud

Aside form installing R on your own system, you can also choose to use its cloud version. This version is hosted by RStudio on https://rstudio.cloud/. To use it, go to the Sign-Up button in the top-right of the screen. Then, select the *Cloud Free* option and once again select Sign-Up. Then, finish the procedure either by filling in your data, or connecting with your Google or GitHub account. Once done, log-in, and you will arrive at your workspace. To get started, you need to make a new project. To do so, click the *New Project* button which takes you to an instance of RStudio. From here on, the programme functions the same as its Desktop version. Note that everything you do - or packages you install - in the project *remain* in the project. Thus, you will have to re-install them if you want to create a new project. Besides, note that RStudio Cloud is quite dependent on both the number of users on the server and your internet connection. Thus, some actions (such as installing packages) might take longer to run.

Chapter 2

Installing Packages

R on its own is a pretty bare-bones experience. What makes it work are the many packages that exist for it. These packages come in two kinds: officially released or in development.

2.1 Installing from CRAN

To be officially released, the package needs to be part of CRAN: the Comprehensive R Archive Network. CRAN is a website that collects and hosts all the material R needs, such as the different distributions, packages, and more. Besides, any package on CRAN has gone through a vetting process. This ensures that the package does not contain any major bugs, has README and NEWS files, and has a clear version number. Many official released packages also have additional documentation and motivating examples published in journals such as *The R Journal* and *The Journal of Statistical Software*. Also, a package being published in CRAN allows us to install the package using the `install.packages` command, or the **Packages** tab in RStudio. Besides, packages on CRAN often receive updates on a regular basis. These updates can add new features to the package, address bugs, or increase performance. To update your packages, go to the **Packages** tab in RStudio and click on the **Update** button.

2.2 Installing from GitHub

Some packages that have not yet had an official release are in development on GitHub (https://github.com/). As a result, these packages change very often and are more unstable as their official counterparts. We can, nevertheless, install packages from Github using the `devtools` package. To install this, type:

```
install.packages("devtools", dependencies=TRUE)
```

Here, `dependencies=TRUE` ensures that if we need other packages to make `devtools` work, R will install these as well. Depending on your operating system, you might have to install some other software for `devtools` to work.

On Windows, `devtools` requires the *RTools* software. To install this, go to https://cran.r-project.org/bin/windows/Rtools/, download the latest *recommended* version (in green), and install it. Then re-open R again and install `devtools` as shown above.

On Linux, how `devtools` installs depends on the flavour of Linux that you have. Most often, installing it as shown above will work fine. If not, the problem is most likely a missing package in your Linux distribution. To address this, close R, open a terminal and type:

1. sudo apt-get update
2. sudo apt-get upgrade
3. sudo apt install build-essential libcurl4-gnutls-dev libxml2-dev libssl-dev
4. Close the terminal, open R, and install `devtools` as shown above.

On OSX (or macOS), `devtools` requires the *XCode* software. To install this, follow these steps:

1. Launch the terminal (which you can find in /Applications/Utilities/), and type:
2. In the terminal, type: **xcode-select –install**
3. A software update window should pop up. Here, click "Install" and agree to the Terms of Service.
4. Go to R and install `devtools` as shown above.

2.3 Quantitative Text Analysis in R

While we will be using several different packages to run quantitative text analysis, we will mostly use `quanteda` (Benoit et al. (2018)). `quanteda` integrates many of the text analysis functions of R that were before spread out over many different packages (see, for example Welbers, Van Atteveldt, and Benoit (2017)). Besides, it is easy to combine with other packages, has simple and logical commands, and a well-maintained website (www.quanteda.io).

The current version of `quanteda` as of writing is `packageVersion("quanteda")`. This version works best with R version 4.0.1 or higher. To check if your system has this, type `R.Version()` in your console. The result will be a list. Look for `$version.string` to see which version number your version of R is. If you do not have the latest version, see the steps above on how to download this.

To install the package, type:

```
install.packages("quanteda", dependencies=TRUE)
```

Note that because we wrote `dependencies=TRUE`, this command also installed three other `quanteda` helper packages that serve to expand upon the basic tools that are already within `quanteda`. In the future, more of these helper packages will be added to expand the package even more. Yet, before these packages get an official release, we can already find them, in development, on GitHub. Here, we install two of them - `quanteda.classifiers` which we will use for supervised learning methods, and `quanteda.dictionaries` which we will use for dictionary analysis:

```
library(devtools)
install_github("quanteda/quanteda.classifiers", dependencies = TRUE)
install_github("kbenoit/quanteda.dictionaries", dependencies = TRUE)
```

Apart from `quanteda` we then need these other packages as well:

```
install_github("mikegruz/kripp.boot", dependencies = TRUE)
install.packages("ca", dependencies = TRUE)
install.packages("combinat", dependencies = TRUE)
install.packages("DescTools", dependencies = TRUE)
install.packages("FactoMineR", dependencies = TRUE)
install.packages("factoextra", dependencies = TRUE)
install.packages("Hmisc", dependencies = TRUE)
install.packages("httr", dependencies = TRUE)
install.packages("jsonlite", dependencies = TRUE)
install.packages("manifestoR", dependencies = TRUE)
install.packages("readr", dependencies = TRUE)
install.packages("readtext", dependencies = TRUE)
install.packages("reshape2", dependencies = TRUE)
install.packages("RTextTools", dependencies = TRUE)
install.packages("R.temis", dependencies = TRUE)
install.packages("rvest", dependencies = TRUE)
install.packages("seededlda", dependencies = TRUE)
install.packages("stm", dependencies = TRUE)
install.packages("tidyverse", dependencies = TRUE)
```

After installation, you will find these packages, as well as the `quanteda` and `devtools` packages, under the **Packages** tab in RStudio.

2.4 Issues, Bugs and Errors

As it is free software, errors are not uncommon in R. Often they arise when you misspell the code or use the wrong code for the job at hand. In these cases, R prints a message (in red) telling you why it cannot do what you ask of it. Sometimes, this message is quite clear, such as telling you to install an extra

package. Other times, it is more complicated and requires some extra work. In these cases, there are four questions you can ask yourself:

1. Did I load all the packages I need?
2. Are all packages up-to-date?
3. Did I spell the commands correct?
4. Is the data in the right shape or format?

If none of these provides a solution, you can always look up online if others have run into the same issue. Often, copy-pasting your error into a search engine can provide you with other instances, and most often a solution. One well-known place for solutions is Stack Overflow (https://stackoverflow.com/). Here, you can share your problem with others and see if someone can offer a solution. Make sure though to read through the problems already posted first, to ensure that you do not post the same problem twice.

Chapter 3

Importing Data

No analysis is possible unless we have some data to work with. In the following exercises, we will look at five different ways to get textual data into R: a) by using .pdf files, b) by using .txt files, c) by using .csv files, d) by using web scraping, and e) by using an API. Before we get to these methods, we will look at how R handles text and how we can work with it.

3.1 Text in R

R sees any form of text as a type of characters vector. In their simplest form, these vectors only have a single character in it. At their most complicated, they can contain many sentences or even whole stories. To see how many characters a vector has, we can use the `nchar` function:

```
vector1 <- "This is the first of our character vectors"
nchar(vector1)
```

```
## [1] 42
```

```
length(vector1)
```

```
## [1] 1
```

This example also shows the logic of R. First, we assign the text we have to a certain object. We do so using the `<-` arrow. This arrow points from the text we have to the object R stores it in, which we here call `vector1`. We then ask R to give us the number of characters inside this object, 40 in this case. The `length` command returns something else, namely 1. This means that we have a single sentence, or word, in our object. If we want to, we can place more sentences inside our object using the `c()` option:

```
vector2 <- c("This is an example", "This is another", "And so we can go on.")
length(vector2)
```

```
## [1] 3
```

```
nchar(vector2)
```

```
## [1] 18 15 20
```

```
sum(nchar(vector2))
```

```
## [1] 53
```

Another thing we can do is extract certain words from a sentence. For this, we use the `substr()` function. With this function, R gives us all the characters that occur between two specific positions. So, when we want the characters between the 4th and 10th characters, we write:

```
vector3 <- "This is yet another sentence"
substr(vector3, 4, 10)
```

```
## [1] "s is ye"
```

We can also split a character vector into smaller parts. We often do this when we want to split a longer text into several sentences. To do so, we use the `strsplit` function:

```
vector3 <- "Here is a sentence - And a second"
parts1 <- strsplit(vector3, "-")
parts1
```

```
## [[1]]
## [1] "Here is a sentence " " And a second"
```

If we now look in the Environment window, we will see that R calls `parts1` a list. This is another type of object that R uses to store information. We will see it more often later on. For now, it is good to remember that lists in R can have many vectors (the layers of the list) and that in each of these vectors we can store many objects. Here, our list has only a single vector. To create a longer list, we have to add more vectors, and then join them together, again using the `c()` command:

```
vector4 <- "Here is another sentence - And one more"
parts2 <- strsplit(vector4, "-")
parts3 <- c(parts1, parts2)
```

We can now look at this new list in the Environment and check that it indeed has two elements. A further thing we can do is to join many vectors together. For this, we can use the `paste` function. Here, the `sep` argument defines how R will combine the elements:

```
fruits <- paste("oranges", "lemons", "pears", sep = "-")
fruits
```

```
## [1] "oranges-lemons-pears"
```

Note that we can also use this command that paste objects that we made earlier together. For example:

```
sentences <- paste(vector3, vector4, sep = ".")
sentences
```

```
## [1] "Here is a sentence - And a second.Here is another sentence - And one more"
```

Finally, we can change the case of the sentence. To do this, we can use `tolower` and `toupper`:

```
tolower(sentences)
```

```
## [1] "here is a sentence - and a second.here is another sentence - and one more"
```

```
toupper(sentences)
```

```
## [1] "HERE IS A SENTENCE - AND A SECOND.HERE IS ANOTHER SENTENCE - AND ONE MORE"
```

Again, we can also run the same command when we have more than a single element in our vector:

```
sentences2 <- c("This is a piece of example text", "This is another piece of example text")
toupper(sentences2)
```

```
## [1] "THIS IS A PIECE OF EXAMPLE TEXT"
## [2] "THIS IS ANOTHER PIECE OF EXAMPLE TEXT"
```

```
tolower(sentences2)
```

```
## [1] "this is a piece of example text"
## [2] "this is another piece of example text"
```

And that is it. As you can see, the options for text analysis in basic R are rather limited. This is why packages such as `quanteda` exist in the first place. Note though, that even `quanteda` uses the same logic of character vectors and combinations that we saw here.

3.2 Import .txt Files

In case that we already have the .txt files somewhere, we can make the above process a bit easier, and begin at the last step:

```
library(readtext)
```

```
txt_directory <- paste0(getwd(), "/Texts")
data_texts <- readtext(paste0(txt_directory, "*"), encoding = "UTF-8")
```

3.3 Import .pdf Files

One of the most popular formats for digital texts is the portable document format (.pdf). To read .pdf files into R, we need two packages. The `pdftools` package to convert the .pdf files into .txt files, and the `readtext` package to read the .txt files into R. Note that this only works if the .pdf files are *readable*. This means that we can select (and copy-paste) the text in them. Thus, `readtext` does not work with .pdf files that the text in them cannot be selected (this is most likely because the pages of the document were scanned as images before turned into a .pdf file). If we have a .pdf file of this type, one solution is to use the `tesseract` package, which can use optical character recognition technology (OCR) to fix this issue.

To import the .pdf files, we start by loading the required libraries into R:

```
library(pdftools)
library(readtext)
```

Then, we go to our working directory (to see where this is, type `getwd()` into the Console). Here, we make two folders: one in which to store the .pdf files - called *PDF* - and another new and empty folder in which to store the .txt files. We call this one *Texts*. Ensure that all the .pdf files are in the *PDF* folder. Then, we tell R about these folders:

```
setwd("Your Working Directory")
pdf_directory <- paste0(getwd(), "/PDF")
txt_directory <- paste0(getwd(), "/Texts")
```

Then, we ask R for a list of all the files in the .pdf directory. This is both to ensure that we are not overlooking anything and to tell R which files are in the folder. Here, setting `recurse=FALSE` means that we only list the files in the main folder and not any files that are in other folders in this main folder.

```
files <- list.files(pdf_directory, pattern = ".pdf", recursive = FALSE,
    full.names = TRUE)

files
```

While we could convert a single document at a time, more often we have to deal with more than one document. To read all documents in at once, we have to write a little function. This function does the following. First, we tell R to make a new function that we label `extract`, and as input give it an element we call `filename`. This filename is at this point an empty element, but to which we will later refer the files we want to extract. Then, we tell it to print the

file name to ensure that we are working with the right files while the function is running. In the next step, we tell it to try to read this filename using the `pdf_text` function and save the result as a file called `text`. Afterwards, we tell it to do so for each of the files that end on .pdf that are in the element `files`. Then, we have it write this text file to a new file. This file is the extracted .pdf in .txt form:

```
extract <- function(filename) {
    print(filename)
    try({
        text <- pdf_text(filename)
    })
    title <- gsub("(.*)/([^/]*).pdf", "\\2", filename)
    write(text, file.path(txt_directory, paste0(title, ".txt")))
}
```

We then use this function to extract all the pdf files in the `pdf_directory` folder. To do so, we use a `for` loop. The logic of this loop is that for each individual `file` in the element `files`, we run the `extract` function we created. This will create an element called `file` for the file R is currently working on, and will create the .txt files in the `txt_directory`:

```
for (file in files) {
    extract(file)
}
```

We can now read the .txt files into R. To do so, we use `paste0(txt_directory, "*")` to tell `readtext` to look into our `txt_directory`, and read any file in there. Besides this, we need to specify the encoding. Most often, this is **UTF-8**, though sometimes you might find **latin1** or **Windows-1252** encodings. While `readtext` will convert all these to **UTF-8**, you have to specify the original encoding. To find out which one you need, you have to look into the properties of the .txt file.

Assuming our texts are in UTF-8 encoding, we run:

```
data_texts <- readtext(paste0(txt_directory, "*"), encoding = "UTF-8")
```

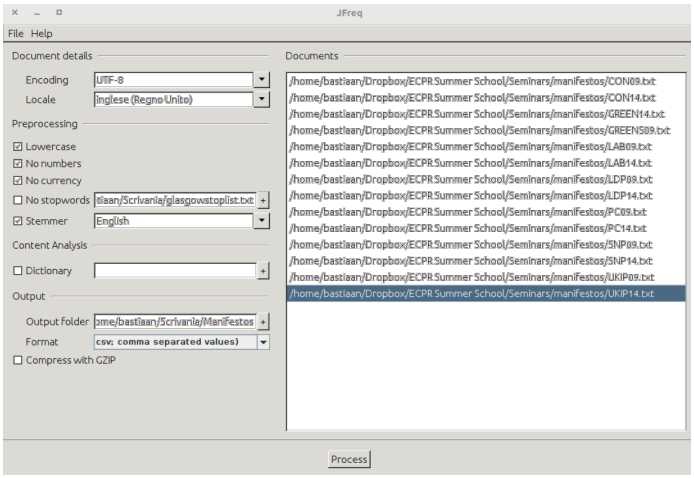
The result of this is a data frame of texts, which we can transform into a corpus for use in `quanteda` or keep as it is for other types of analyses.

3.4 Import .csv Files

We can also choose not to import the texts into R in a direct fashion, but import a .csv file with word counts instead. One way to generate these counts is by using JFreq (Lowe 2011). This is a useful stand-alone programme written in Java that generates a .csv file where the rows represent the documents and the columns represent the individual words contained in the documents. The cells therefore,

contain the wordcounts for each word within each document. JFreq also allows performing some basic pre-processing. JFreq is not actively maintained, but is available at https://conjugateprior.org/software/jfreq/.

To use JFreq, open the programme and drag and drop all the documents you want to process into the window of the programme. Once you do this, the document file names will appear in the document window. Then, you can choose from several pre-processing options. Amongst these are options to make all words lowercase or remove numbers, currency symbols, or stop words. The latter are words that often appear in texts which do not carry an important meaning. These are words such as `and''`,or'' and "but''. As stop words are language-specific and often context-specific as well, we need to tell JFreq what words are stop words. We can do so by putting all the stop words in a separate .txt file and load it in JFreq. You can also find many lists of stopwords for different languages online. For instance, many different lists of stopwords in English are available in this GitHub page: https://github.com/igorbrigadir/stopwords Finally, we can apply a stemmer which reduces words such as Europe and European to a single Europ* stem. JFreq allows us to use pre-defined stemmers by choosing the relevant language from a drop-down menu. In the following screenshot, you can see the JFreq at work importing the .txt files of a number of election manifestos.



Note that here the encoding is UTF-8 while the locale is English (UK). Once we have specified all the options we want, we give a name for the output folder and press *Process*. Now we go to that folder we named and copy-paste the "data.csv'' file into your Working Directory. In R, we then run the following:

```
data_manifestos <- read.csv("data.csv", row.names = 1, header = TRUE)
```

By specifying `row.names=1`, we store the information of the first column in the data frame itself. This column, containing the names of the documents now belongs to the object of the data frame and does not appear as a separate column. The same is true for `header=TRUE` which ensures that the first row

gives names to the columns (in this case containing the words).

3.5 Import from an API

Instead of importing the online data page-by-page, we can also use special programmes to download lots of data at once. We can do so with an Application Programming Interface (API). The main difference between using an API and regular webscraping is that APIs are specifically designed for this purpose. This means that it is easier for R to read the webpages, and that you can download a large amount of data at once. APIs are offered by many popular web sites like Wikipedia, social networking sites like Twitter and Facebook, newspapers such as *The New York Times*, and so on.

While almost all websites can be read by the `rvest` package, for the APIs you often need a specific package. For example, for Twitter there is the `rtweet` package, for Facebook `rFacebook`, and `ggmap` for Google maps. Also, you often, if not always, need to register first before you can use an API. Note, however, that Facebook has recently taken steps in restricting access to their public APIs for research purposes, which means that research on Facebook users' posts is no longer an option (see Freelon (2018) and Perriam, Birkbak, and Freeman (2020)).

Having said this, however, there are many APIs with associated R packages that are made by researchers and for researchers. Let's look at an example using an API for the New York Times. If you look at the website (https://developer.nytimes.com/), you find that we can get information ranging from articles to books and reviews.

Before we start here, we first have to maintain permission to use the API. For this, register on the site and log in. Then, make a new "app" under: https://developer.nytimes.com/my-apps and ensure you select the movie reviews. Then, you can click on the new app to see your key under "API Keys." It is this string of codes and letters you will have to place at the [YOUR_API_KEY_HERE] section.

Now, let us first load the packages:

```
library(tidyverse)
library(httr)
library(jsonlite)
```

We can then build our request. As you can see on the site, the request requires us to give a search term (here we choose "love"). Optionally, we can set a time frame from which we want to sample the reviews:

```
reviews <-  fromJSON("https://api.nytimes.com/svc/movies/v2/reviews/search.json?query=love&openin
```

The result is a JSON object that you can see in the environment. While JSON

(JavaScript Object Notation) is a generic way in which information is easy to share - and is thus often used - it is not in an ideal form. So, we change the JSON information to a data frame using the following:

```
reviews_df <-  fromJSON("https://api.nytimes.com/svc/movies/v2/reviews/search.json?quer
        data.frame()
```

You can now find all the information in the new `reviews_df` object, which contains information about the type of crime, location, month etc. This is one example of an API, but there are many others available, such as those of the EU, OpenStreetMaps, Weather Underground, etc. As we can see though, having a package makes things easier, though more limited.

3.6 Import using Web Scraping

If our text is online (e.g. as part of a website) we can also choose to get it from there without copying it into a .txt file first. To do so, we have to employ web scraping. The logic of web scraping is that we use the structure of the underlying HTML document to find and download the text we want. Note though that not all websites encourage (or even allow) scraping. So, do have a look at their disclaimer before we do so. You can do this by either checking the website's *terms and condition* page, or the robots.txt file that you can usually find appended at the home page (e.g. https://www.facebook.com/robots.txt).

In the following example we will see how one can download information from the Internet Movie Database (IMDb): https://www.imdb.com Note that the IMDb does not allow you to do any web scraping, so the following example is given for illustration purposes only! If you are interested in analyzing data from IMDB you can download the official datasets that are released by IMDB here: https://datasets.imdbws.com/ The documentation for these datasets is available here: https://www.imdb.com/interfaces/ If you would like to learn more about web scraping in the context of quantitative text analysis we suggest the textbook by Munzert et al. (2014).

In the following example we show how to download the user reviews that appear on the IMDB website. The first command, `read_html` downloads this whole page. If you look at this page in your browser, you see that there are many other things on there besides the user review. To tell R which part is the text to download, we use the `html_nodes` command. This command looks for a certain header on the HTML page and starts downloading from there. The `html_text` command then reads that bit of text and puts it into the object. Note that the `%>%` command we use here is what we call a *pipe*. What it does is that it transports the output of one command into another, without saving it to an intermediate object. So here, we first download the HTML, find the right header, and only then save it into an object. Having done this for three reviews, we then bind them together:

```
library(rvest)

review1 <- read_html("http://www.imdb.com/title/tt1979376/") %>%
 html_nodes("#titleUserReviewsTeaser p") %>%
 html_text()

review2 <- read_html("http://www.imdb.com/title/tt6806448/") %>%
 html_nodes("#titleUserReviewsTeaser p") %>%
 html_text()

review3 <- read_html("http://www.imdb.com/title/tt7131622/") %>%
 html_nodes("#titleUserReviewsTeaser p") %>%
 html_text()

reviews_scraping <- c(review1, review2, review3)
```

Chapter 4

Reliability and validity

We could say that the central tenet of quantitative text analysis, which sets it apart from other approaches to analyzing text, is that it strives to be objective and replicable. In measurement theory, we use the terms **reliability** and **validity** to convey this message.

Reliability refers to consistency, that is, the degree to which we get similar results whenever we apply a measuring instrument to measure a given concept. This is similar to the concept of *replicability*. Validity, on the other hand, refers to unbiasedness, that is, the degree to which our measure really measures the concept which intends to measure. In other words, validity looks whether the measuring instrument that we are using is objective.

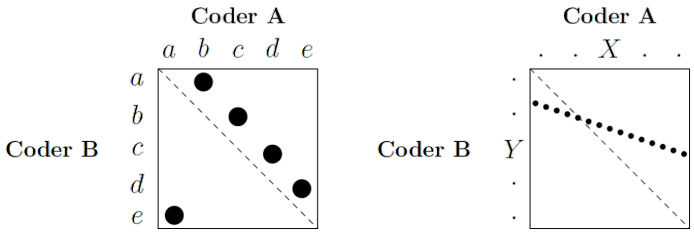
Carmines and Zeller (1979) distinguish among three types of validity. *Content Validity*, which refers to whether our measure represents all facets of the construct of interest; *criterion Validity*, which looks at whether our measure correlates with other measures of the same concept, and *construct Validity*, which looks at whether our measure behaves as expected within a given theoretical context. I should also say here, that the three types of validity are not interchangeable. Ideally, one has to prove that their results pass all three validity tests. In the words of Grimmer and Stewart (2013): "Validate, validate, validate!"

Krippendorff (2004) distinguishes among three types of reliability. *Stability*, which he considers as the weakest form of coding reliability, and which can be measured when the same text is coded by the same coder more than once, *reproducibility*, which is measured by the degree of agreement among independent coders, and *accuracy*, which he considers as the strongest form of coding reliability, and which is measured by the agreement between coders and a given standard. However, in the absence of a benchmark, we are usually interested in measuring reliability as reproducibility, in other words as inter-coder agreement.

4.1 Inter-Coder Agreement

Hayes and Krippendorff (2007, 79) argue that a good measure of the agreement should at least address five criteria. The first is that it should apply to many coders, and not only two. Also, when we use the method for more coders, there should be no difference in how many coders we include. The second is that the method should only take into account the actual number of categories the coders used and not all that were available. This as while the designers designed the coding scheme on what they thought the data would look like, the coders use the scheme based on what the data is. Third, it should be numerical, meaning that we can use it to make a scale between 0 (absence of agreement) and 1 (perfect agreement). Fourth, it should be appropriate for the level of measurement. So, if our data is ordinal or nominal, we should not use a measure that assumes metric data. This ensures that the metric uses all the data and that it does not add or not use other information. Fifth, we should be able to compute (or know), the sampling behaviour of the measure.

With these criteria in mind, we see that popular methods, such as % agreement or Pearson's r, can be misleading. Especially for the latter - as it is a quite popular method - this often leads to problems, as this figure by Krippendorff (2004) shows:

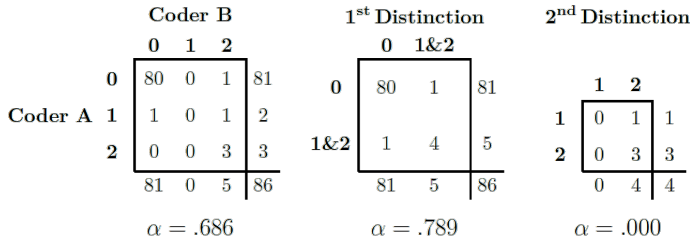


Here, the figure on the left shows two coders: A and B. The dots in the figure show the choices both coders made, while the dotted line shows the line of perfect agreement. If a dot is on this line, it means that both Coder A and Coder B made the same choice. In this case, they disagreed in all cases. When Coder A chose a, Coder B chose e, when Coder A chose b, Coder B chose a, and so on. Yet, when we would calculate Pearson's r for this, we would find a result as shown in the right-hand side of the figure. Seen this way, the agreement between both coders does not seem a problem at all. The reason for this is that Pearson's r works with the distances between the categories *without* taking into account their location. So, for a positive relationship, the only thing Pearson's r requires is that for every increase or decrease for one coder, there is a similar increase or decrease for the other. This happens here with four of the five categories. The result is thus a high Pearson's r, though the actual agreement should be 0.

Pearson's r thus cannot fulfil all our criteria. A measure that can is Krippendorff's α (Krippendorff 2004). This measure can not only give us the agreement

we need, but can also do so for nominal, ordinal, interval, and ratio level data, as well as data with many coders and missing values. Besides, we can compute 95% confidence intervals around α using bootstrapping, which we can use to show the degree of uncertainty around our reliability estimates.

Despite this, Krippendorff's α is not free of problems. One main problem occurs when coders agree on only a few categories and use these categories a considerable number of times. This leads to an inflation of α, making it is higher than it should be (Krippendorff 2004), as in the following example:



Here, in the left-most figure, we see coders A and B who have to code into three categories: 0, 1, or 2. In this example, the categories 1 and 2 carry a certain meaning, while category 0 means that the coders did not know what to assign the case to. Of the 86 cases, both coders code 80 cases in the 0 category. This means that there are only 6 cases on which they can agree or disagree about a code that carries some meaning. Yet, if we calculate α, the result - 0.686 - takes into account all the categories. One solution for this is to add up the categories 1 and 2, as the figure in the middle shows. Here, the coders agree in 84 of the 86 cases (on the diagonal line) and disagree in only 2 of them. Calculating α now shows that it would increase to 0.789. Finally, we can remove the 0 category and again view 1 and 2 as separate categories (as the most right-hand figure shows). Yet, the result of this is quite disastrous. While the coders agree in 3 of the 4 cases, the resulting α equals 0.000, as coder B did not use category 1 at all.

Apart from these issues, Krippendorff's α is a stable and useful measure. A value of $\alpha = 1$ indicates perfect reliability, while a value of $\alpha = 0$ indicates the absence of reliability. This means that if $\alpha = 0$, there is no relationship between the values. It is possible for $\alpha < 0$, which means that the disagreements between the values are larger than they would be by chance and are systematic. As for thresholds, Krippendorff (2004) proposes to use either 0.80 or 0.67 for results to be reliable. Such low reliability often has many causes. One thing might be that the coding scheme is not appropriate for the documents. This means that coders had categories that they had no use for, and lacked categories they needed. Another reason might be that the coders lacked training. Thus, they did not understand how to use the coding scheme or how the coding process works. This often leads to frustration on part of the coders, as in these cases the process often becomes time-consuming and too demanding to carry out.

To calculate Krippendorff's α, we can use the following software:

- KALPHA custom dialogue (SPSS)
- **kalpha** user-written package (Stata)
- KALPHA macro (SAS)
- **kripp.alpha** command in **kripp.boot** package (R) - amongst others

Let us try this in R using an example. Here, we will look at the results of a coding reliability test where 12 coders assigned the sentences of the 1997 European Commission work programme in the 20 categories of a policy areas coding scheme. We can find the results for this on GitHub. To get the data, we tell R where to find it, then to read that file as a .csv file and write is to a new object:

```
library(readr)

urlfile = "https://raw.githubusercontent.com/SCJBruinsma/qta-files/master/reliability_r
reliability_results <- read_csv(url(urlfile))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   coder1 = col_double(),
##   coder2 = col_double(),
##   coder3 = col_double(),
##   coder4 = col_double(),
##   coder5 = col_double(),
##   coder6 = col_double(),
##   coder7 = col_double(),
##   coder8 = col_double(),
##   coder9 = col_double(),
##   coder10 = col_double(),
##   coder11 = col_double(),
##   coder12 = col_double()
## )
```

Notice that in the data frame we just created, the coders are in the columns and the sentences in the rows. As the `kripp.boot` package requires it to be the other way around and in matrix form, we first transpose the data, and then place it in a matrix. Finally, we run the command and specify we want the nominal version:

```
library("kripp.boot")

reliability_results_t <- t(reliability_results)
reliability <- as.matrix(reliability_results_t)
kalpha <- kripp.boot(reliability, iter=1000, method = "nominal")
kalpha$value
```

Note also that `kripp.boot` is a GitHub package. You can still calculate the value (but without the confidence interval) with another package:

```
library("DescTools")

reliability_results_t <- t(reliability_results)
reliability <- as.matrix(reliability_results_t)
kalpha <- KrippAlpha(reliability, method = "nominal")
kalpha$value
```

As we can see, the results point out that the agreement among the coders is 0.634 with an upper limit of 0.650 and a lower limit of 0.618 which is short of Krippendorff's cut-off point of 0.667.

4.2 Visualizing Quality

Lamprianou (2020) notes that existing reliability indices may mask coding problems and that the reliability of coding is not stable across coding units (as illustrated in the example given for Krippendorff's alpha in Section 3.2 above). To investigate the quality of coding he proposes using social network analysis (SNA) and exponential random graph models (ERGM). Here, we illustrate a different approach, based on the idea of sensitivity analysis.

We therefore compare the codings of each coder against all others (and also against a benchmark or a gold standard). For this, we need to bootstrap the coding reliability results to create an uncertainty measure around each coder's results, following the approach proposed by Benoit, Laver, and Mikhaylov (2009). The idea is to use a non-parametric bootstrap for the codings of each coder (using 1000 draws with replacement) at the category level and then calculate the confidence intervals. Their width then depends on both the number of sentences coded by each coder (n) in each category and the number of coding categories that are not empty. Thus, larger documents and fewer empty categories result in narrower confidence intervals, while a small number of categories leads to wider intervals (Lowe and Benoit 2011).

To start, the first thing we do is load two packages we need into R using the `library` command:

```
library(Hmisc)
library(combinat)
```

In the following example we perform the sensitivity analysis on the coded sentences of the 1997 European Commission work programme, as given in Section 3.2. Here, however, the same data is arranged differently. Each row represents a coder, and each column represents a coding category (*c0* to *c19*). In each cell, we see the number of sentences that each coder coded in each category, with the column *n* giving the sum of each row:

```
coderid <- c("coder1", "coder2", "coder3", "coder4", "coder5",
    "coder6", "coder7", "coder8", "coder9", "coder10", "coder11",
    "coder12")
c0 <- c(14, 0, 0, 9, 29, 1, 2, 11, 1, 8, 9, 0)
c01 <- c(4, 1, 1, 2, 2, 3, 2, 1, 1, 1, 6, 0)
c02 <- c(5, 5, 5, 3, 5, 4, 6, 6, 3, 1, 3, 6)
c03 <- c(15, 12, 12, 26, 13, 22, 8, 14, 15, 25, 14, 21)
c04 <- c(5, 6, 6, 5, 4, 6, 6, 5, 6, 6, 6, 6)
c05 <- c(0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)
c06 <- c(9, 10, 22, 12, 9, 11, 11, 7, 9, 11, 6, 20)
c07 <- c(2, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 2)
c08 <- c(3, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2)
c09 <- c(5, 7, 5, 5, 5, 6, 5, 6, 8, 7, 7, 6)
c10 <- c(23, 23, 22, 23, 18, 23, 22, 23, 23, 25, 24, 22)
c11 <- c(31, 31, 33, 40, 25, 23, 25, 30, 40, 16, 40, 31)
c12 <- c(2, 3, 1, 4, 0, 3, 1, 5, 3, 2, 3, 3)
c13 <- c(2, 4, 3, 3, 3, 3, 2, 5, 2, 2, 3, 2)
c14 <- c(13, 12, 11, 13, 9, 14, 18, 14, 2, 22, 12, 14)
c15 <- c(9, 8, 8, 5, 7, 8, 10, 10, 13, 8, 8, 7)
c16 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
c17 <- c(3, 3, 4, 1, 3, 3, 2, 1, 3, 3, 3, 3)
c18 <- c(16, 33, 27, 8, 26, 28, 31, 22, 28, 23, 14, 16)
c19 <- c(3, 3, 2, 1, 3, 3, 3, 1, 4, 2, 3, 3)
c20 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
n <- c(164, 164, 164, 163, 164, 164, 155, 164, 164, 164, 164,
    164)

data_uncertainty <- data.frame(coderid, c0, c01, c02, c03, c04,
    c05, c06, c07, c08, c09, c10, c11, c12, c13, c14, c15, c16,
    c17, c18, c19, c20, n, stringsAsFactors = FALSE)
```

We then tell R how many coders we have. As this number is equal to the number of rows we have, we can get this number using the `nrow` command. We also specify the number of bootstraps we want to carry out (1000) and transform our data frame into an array. We do the latter as R needs the data in this format later on:

```
nman <- nrow(data_uncertainty)
nrepl <- 1000
manifBSn <- manifBSnRand <- array(as.matrix(data_uncertainty[,
    2:21]), c(nman, 20, nrepl + 1), dimnames = list(1:nman, names(data_uncertainty[,
    2:21]), 0:nrepl))
```

We then bootstrap the sentence counts for each coder and compute percentages for each category using a multinomial draw. First, we define `p`, which is the proportion of each category over all the coders. Then, we input this value

together with the total number of codes `n` into the `rmultinomial` command, which gives the random draws. As we want to do this a 1000 times, we place this command into a `for` loop:

```
p <- manifBSn[, , 1]/n

for (i in 1:nrepl) {
    manifBSn[, , i] <- rmultinomial(n, p)
}
```

With this data, we can then ask R to compute the quantities of interest. These are standard errors for each category, as well as the percentage coded for each category:

```
c0SE <- apply(manifBSn[, "c0", ]/n * 100, 1, sd)
c01SE <- apply(manifBSn[, "c01", ]/n * 100, 1, sd)
c02SE <- apply(manifBSn[, "c02", ]/n * 100, 1, sd)
c03SE <- apply(manifBSn[, "c03", ]/n * 100, 1, sd)
c04SE <- apply(manifBSn[, "c04", ]/n * 100, 1, sd)
c05SE <- apply(manifBSn[, "c05", ]/n * 100, 1, sd)
c06SE <- apply(manifBSn[, "c06", ]/n * 100, 1, sd)
c07SE <- apply(manifBSn[, "c07", ]/n * 100, 1, sd)
c08SE <- apply(manifBSn[, "c08", ]/n * 100, 1, sd)
c09SE <- apply(manifBSn[, "c09", ]/n * 100, 1, sd)
c10SE <- apply(manifBSn[, "c10", ]/n * 100, 1, sd)
c11SE <- apply(manifBSn[, "c11", ]/n * 100, 1, sd)
c12SE <- apply(manifBSn[, "c12", ]/n * 100, 1, sd)
c13SE <- apply(manifBSn[, "c13", ]/n * 100, 1, sd)
c14SE <- apply(manifBSn[, "c14", ]/n * 100, 1, sd)
c15SE <- apply(manifBSn[, "c15", ]/n * 100, 1, sd)
c16SE <- apply(manifBSn[, "c16", ]/n * 100, 1, sd)
c17SE <- apply(manifBSn[, "c17", ]/n * 100, 1, sd)
c18SE <- apply(manifBSn[, "c18", ]/n * 100, 1, sd)
c19SE <- apply(manifBSn[, "c19", ]/n * 100, 1, sd)

per0 <- apply(manifBSn[, "c0", ]/n * 100, 1, mean)
per01 <- apply(manifBSn[, "c01", ]/n * 100, 1, mean)
per02 <- apply(manifBSn[, "c02", ]/n * 100, 1, mean)
per03 <- apply(manifBSn[, "c03", ]/n * 100, 1, mean)
per04 <- apply(manifBSn[, "c04", ]/n * 100, 1, mean)
per05 <- apply(manifBSn[, "c05", ]/n * 100, 1, mean)
per06 <- apply(manifBSn[, "c06", ]/n * 100, 1, mean)
per07 <- apply(manifBSn[, "c07", ]/n * 100, 1, mean)
per08 <- apply(manifBSn[, "c08", ]/n * 100, 1, mean)
per09 <- apply(manifBSn[, "c09", ]/n * 100, 1, mean)
per10 <- apply(manifBSn[, "c10", ]/n * 100, 1, mean)
per11 <- apply(manifBSn[, "c11", ]/n * 100, 1, mean)
```

```
per12 <- apply(manifBSn[, "c12", ]/n * 100, 1, mean)
per13 <- apply(manifBSn[, "c13", ]/n * 100, 1, mean)
per14 <- apply(manifBSn[, "c14", ]/n * 100, 1, mean)
per15 <- apply(manifBSn[, "c15", ]/n * 100, 1, mean)
per16 <- apply(manifBSn[, "c16", ]/n * 100, 1, mean)
per17 <- apply(manifBSn[, "c17", ]/n * 100, 1, mean)
per18 <- apply(manifBSn[, "c18", ]/n * 100, 1, mean)
per19 <- apply(manifBSn[, "c19", ]/n * 100, 1, mean)
```

We then bind all these quantities together in a single data frame:

```
dataBS <- data.frame(cbind(data_uncertainty[, 1:22], c0SE, c01SE,
    c02SE, c03SE, c04SE, c05SE, c06SE, c07SE, c08SE, c09SE, c10SE,
    c11SE, c12SE, c13SE, c14SE, c15SE, c16SE, c17SE, c18SE, c19SE,
    per0, per01, per02, per03, per04, per05, per06, per07, per08,
    per09, per10, per11, per12, per13, per14, per15, per16, per17,
    per18, per19))
```

While we can now inspect the results by looking at the data, it becomes more clear when we visualise this. While R has some inbuilt tools for visualisation (in the `graphics` package), these tools are rather crude. Thus, here we will use the `ggplot2` package, which extends our options, and which has an intuitive structure:

```
library(ggplot2)
```

First, we make sure that the variable `coderid` is a factor and make sure that it is in the right order:

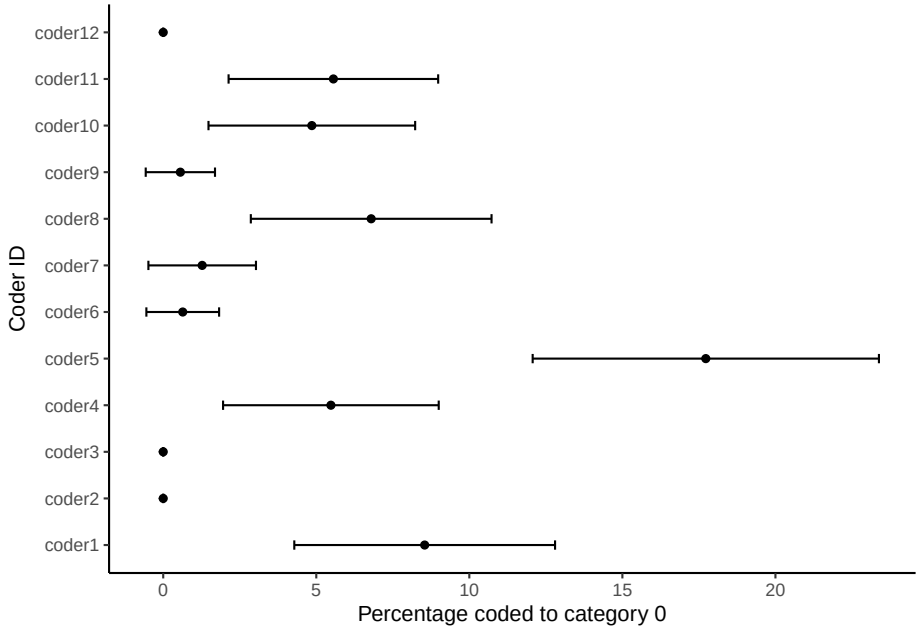
```
dataBS$coderid <- as.factor(dataBS$coderid)
dataBS$coderid <- factor(dataBS$coderid, levels(dataBS$coderid)[c(1,
    5:12, 2:4)])
```

Then, we calculate the 95% confidence intervals for each category. We do so using the percent of each category and the respective standard error, and add these values to our data-set:

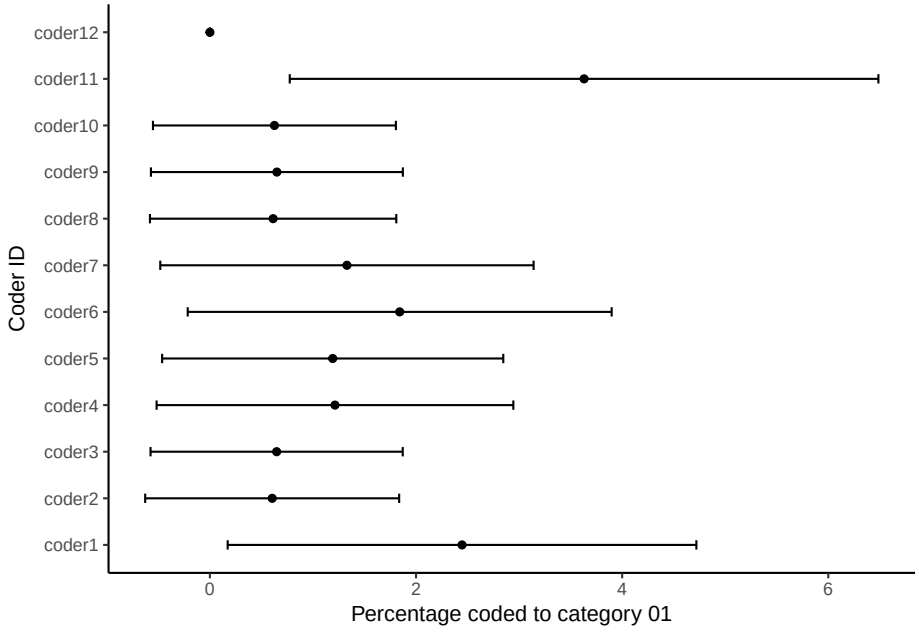
```
c0_lo <- per0 - (1.96 * c0SE)
c0_hi <- per0 + (1.96 * c0SE)
c01_lo <- per01 - (1.96 * c01SE)
c01_hi <- per01 + (1.96 * c01SE)
c02_lo <- per02 - (1.96 * c02SE)
c02_hi <- per02 + (1.96 * c02SE)

dataBS <- cbind(dataBS, c0_lo, c0_hi, c01_lo, c01_hi, c02_lo,
    c02_hi)
```

Finally, we generate the graphs for each individual category:

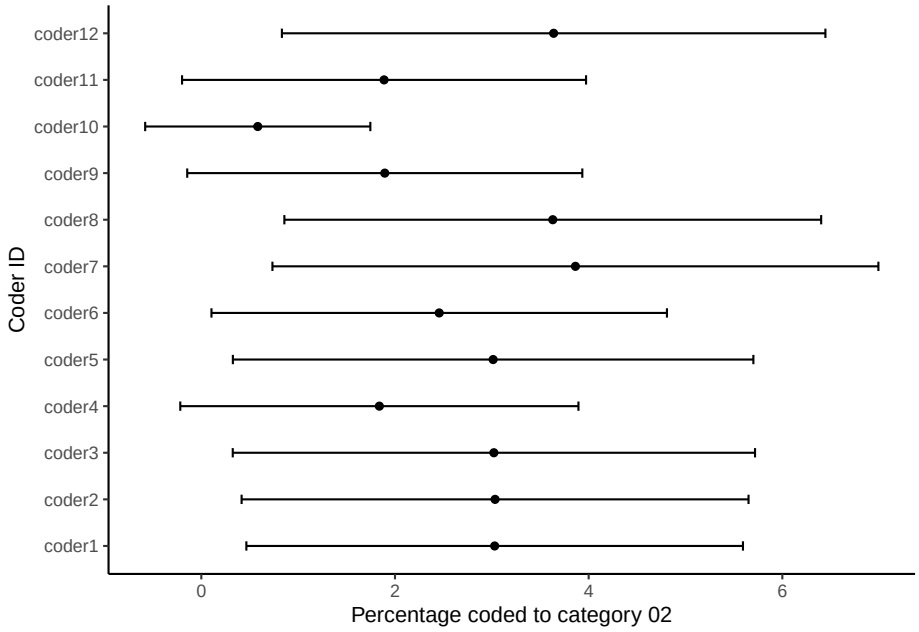
```
ggplot(dataBS, aes(per0, coderid)) + geom_point() + geom_errorbarh(aes(xmax = c0_hi,
    xmin = c0_lo), height = 0.2) + xlab("Percentage coded to category 0") +
    ylab("Coder ID") + theme_classic()
```




```
ggplot(dataBS, aes(per01, coderid)) + geom_point() + geom_errorbarh(aes(xmax = c01_hi,
    xmin = c01_lo), height = 0.2) + xlab("Percentage coded to category 01") +
    ylab("Coder ID") + theme_classic()
```




```
ggplot(dataBS, aes(per02, coderid)) + geom_point() + geom_errorbarh(aes(xmax = c02_hi,
    xmin = c02_lo), height = 0.2) + xlab("Percentage coded to category 02") +
    ylab("Coder ID") + theme_classic()
```



Each figure shows the percentage that each of the coders coded in the respective

category of the coding scheme. We thus use the confidence intervals around the estimates to look at the degree of uncertainty around each estimate. We can read the plots by looking if the dashed line is within the confidence intervals for each coder. The larger the coders deviate from the benchmark or standard, the less likely that they understood the coding scheme in the same way. It also means that it is more likely that a coder would have coded the work programme much different from the benchmark coder. Thus, such a sensitivity analysis is like having a single reliability coefficient for each coding category.

Chapter 5

Preliminaries

Before we start with any kind of analyses, it pays to have a brief look at the preliminaries first. The goal of these preliminaries is to give us a better understanding of "what" are texts are about, who there authors are, and what kind of words and information we can expect to find in them. This is necessary for us to know our data. Not only is it standard academic practise to know your data well, it also helps us to decide whether results we will encounter later on really do make sense. Here, we look at three different preliminaries: the idea of keywords-in-context, several visualisations, and a range of textual statistics. Before that though, we will have a brief look at the idea of the "corpus," as it is central to the idea of how `quanteda` works.

5.1 The Corpus

Within `quanteda`, the main way to store documents is in the form of a `corpus` object. This object contains all the information that comes with the texts and does not change during our analysis. Instead, we make copies of the main corpus, change them into the type we need, and run our analyses on them. The advantage of this is that we always can go back to our original data.

Apart from importing texts ourselves, `quanteda` contains several corpora as well. Here, we use one of these, which contains the inaugural speeches of all the US Presidents. For this, we first have to load the main package and then load the data into R:

```
library(quanteda)

data(data_corpus_inaugural)
head(data_corpus_inaugural)
```

```
## Corpus consisting of 6 documents and 4 docvars.
```

```
## 1789-Washington :
## "Fellow-Citizens of the Senate and of the House of Representa..."
##
## 1793-Washington :
## "Fellow citizens, I am again called upon by the voice of my c..."
##
## 1797-Adams :
## "When it was first perceived, in early times, that no middle ..."
##
## 1801-Jefferson :
## "Friends and Fellow Citizens: Called upon to undertake the du..."
##
## 1805-Jefferson :
## "Proceeding, fellow citizens, to that qualification which the..."
##
## 1809-Madison :
## "Unwilling to depart from examples of the most revered author..."
```

You should now see the corpus appear in the Environment tab. If you click on it, you can see, amongst others, that the corpus comes with information on the Year of the release of the speech and the president it belongs to. As the corpus is quite large, we make it a bit more manageable by only selecting the speeches from 1900 onwards. We can do this by using the `corpus_subset` command for both:

```
corpus_inaugural <- corpus_subset(data_corpus_inaugural, Year > 1900)
```

Now we have our corpus, we can start with the analysis. As noted, we try not to carry out any analysis on the corpus itself. Instead, we keep it as it is and work on its copies. Often, this means transforming the data into another shape. One of the more popular shapes is the data frequency matrix (dfm). This is a matrix which contains the documents in the rows and the word counts for each word in the columns.

Before we can do so however, we have to split up our texts into unique words. To do this, we first have to construct a `tokens` object. In the command that we use to do this, we can specify how we want our texts to be split (here we use the standard option), and in addition clean our data a bit. For example, we can specify that we want to convert all the texts into lowercase and remove any numbers and special characters.

```
data_inaugural_tokens <- tokens(corpus_inaugural, what = "word",
    remove_punct = TRUE, remove_symbols = TRUE, remove_numbers = TRUE,
    remove_url = TRUE, remove_separators = TRUE, split_hyphens = FALSE,
    include_docvars = TRUE, padding = FALSE, verbose = TRUE)
```

```
## Creating a tokens object from a corpus input...
```

```
##  ...starting tokenization
##  ...1901-McKinley to 2021-Biden.txt
##  ...preserving hyphens
##  ...preserving social media tags (#, @)
##  ...segmenting into words
##  ...7,013 unique types
##  ...removing separators, punctuation, symbols, numbers, URLs
##  ...complete, elapsed time: 0.15 seconds.
## Finished constructing tokens from 31 documents.
```

We can also remove certain stopwords so that words like "and" or "the" do not influence our analysis too much. We can either specify these words ourselves or we can use a list that is already present in R. To see this list, type `stopwords("english")` in the console. Stopwords for other languages are also available (such as German, French and Spanish). Even more stopwords can be found in the `stopword` package, that can easily be integrated with `quanteda`. For now, we will use the English ones. First, however, as all the stopwords are lower-case, we will have to lower case our words as well:

```
data_inaugural_tokens <- tokens_tolower(data_inaugural_tokens, keep_acronyms = FALSE)
data_inaugural_tokens <- tokens_select(data_inaugural_tokens, stopwords("english"), selection = '
```

Then, we can construct our dfm:

```
data_inaugural_dfm <- dfm(data_inaugural_tokens)
```

5.2 Keywords in Context

Besides all the analytical techniques we have available, we can also use `quanteda` to look at various rather simpler ones. One popular option is known as keywords in context (kwic), in which we are interested with which other words a certain word appears in our texts. This is also known as looking at the "concordance" of our text. Here, we can easily find this with our tokens dataframe. Let's say we are interested in all those words that start with "secure" and we want to know which three words occur before and after this word. We can then run:

```
tokens <- tokens(corpus_inaugural)
kwic_output <- kwic(tokens, pattern = "secure*", valuetype = "glob", window = 3)
```

In the outputted object, we find a column labelled "pre" and another labelled "post." These refer to the words that came either before or after the word "secure*". We can easily take these out and combine them:

```
text_pre <- kwic_output$pre
text_post <- kwic_output$post
text_word <- kwic_output$keyword
text <- as.data.frame(paste(text_pre, text_word, text_post))
```

We then combine this information with the name of the document it came from so that we know which text the word is from:

```
extracted <- cbind(kwic_output$docname, text)
names(extracted) <- c("docname", "text")
head(extracted)
```

```
##         docname                                        text
## 1 1901-McKinley be adapted to secure a government capable
## 2     1909-Taft               , and to secure at the same
## 3     1909-Taft          is needed to secure a more rapid
## 4     1909-Taft  . This should secure an adequate revenue
## 5     1909-Taft     business . To secure the needed speed
## 6     1909-Taft    duties as to secure an adequate income
```

5.3 Visualisations and Descriptives

Another thing we can do with this dfm is to generate a frequency graph using the `topfeatures` function. For this, we first have to save the 50 most frequently occurring words in our texts (note that there is also the `textstat_frequency` function in the `quanteda.textstats` helper package that can do this):

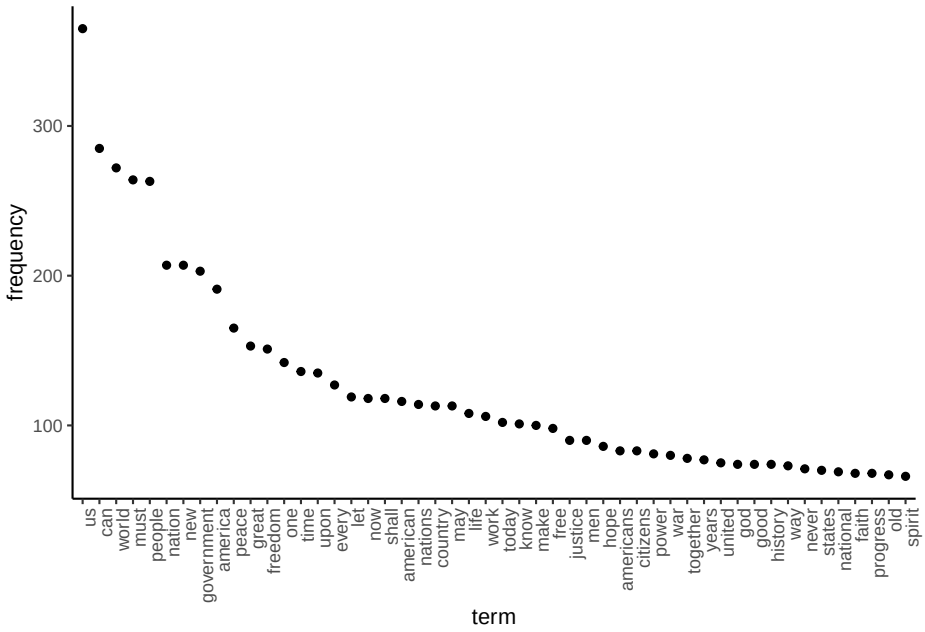
```
features <- topfeatures(data_inaugural_dfm, 50)
```

We then have to transform this object into a data frame, and sort it by decreasing frequency:

```
features_plot <- data.frame(list(term = names(features),frequency = unname(features)))
features_plot$term <- with(features_plot, reorder(term, -frequency))
```

Then we can plot the results:

```
library(ggplot2)
ggplot(features_plot) +
 geom_point(aes(x=term, y=frequency)) +
 theme_classic()+
 theme(axis.text.x=element_text(angle=90, hjust=1))
```



We can also generate word clouds. As these show all the words we have, we will trim our dfm first to remove all those words that occurred less than 40 times. We can do this with the `dfm_trim` function. Then, we can use this newly trimmed dfm to generate the word cloud:

```
library(quanteda.textplots)

wordcloud_dfm_trim <- dfm_trim(data_inaugural_dfm, min_termfreq = 40)
textplot_wordcloud(wordcloud_dfm_trim)
```

If we would want to, we can also split up this word cloud based on which words belong to which party. For this, we have to generate a new dfm and within it, specify the groups that well which words belong to which party. Given that we have only Democratic and Republican presidents, we end up with two groups:

```
library(quanteda.textplots)

wordcloud_dfm_comp <- dfm_group(data_inaugural_dfm, groups = Party)
wordcloud_dfm_comp <- dfm_trim(wordcloud_dfm_comp, min_termfreq = 20,
    max_words = 40)
textplot_wordcloud(wordcloud_dfm_comp, comparison = TRUE)
```

5.4 Text Statistics

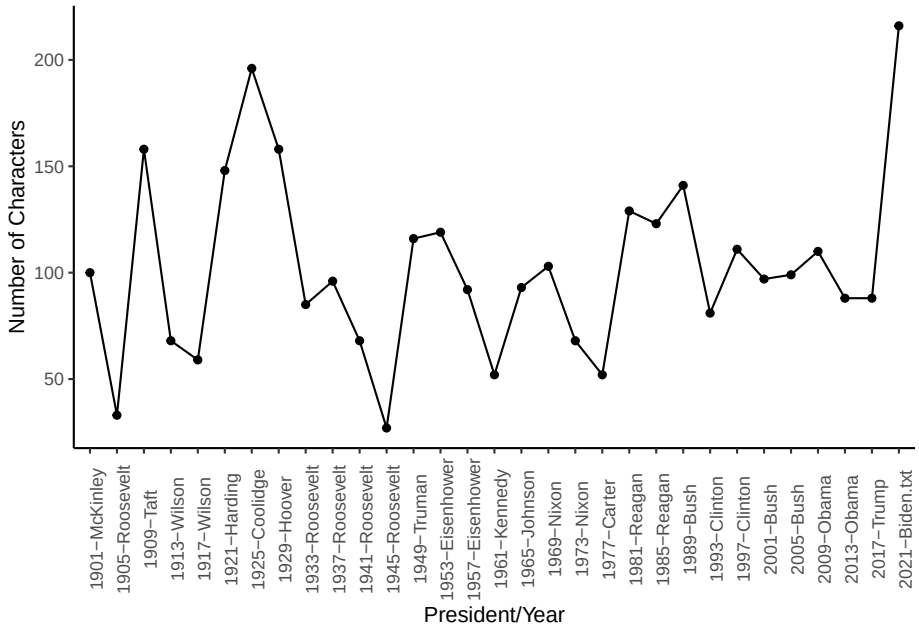
Finally, *quanteda* also allows us to calculate quite a number of textual statistics. These are all collected in the *quanteda.textstats* helper package. Here, we will look at several of them, starting with a simple overview of our corpus in the terms of a summary that tells us the number of characters, sentences, tokens, etc. for each of the texts:

```
library(quanteda.textstats)
corpus_summary <- textstat_summary(corpus_inaugural)
```

If we want, we can then use this data to make some simple graphs telling us various things about the texts in our corpus. As an example, let's look at the number of sentences the various presidents put in their speeches:

```
ggplot(data=corpus_summary, aes(x=document, y=sents, group=1)) +
  geom_line()+
  geom_point()+
  ylab("Number of Characters")+
```

```
  xlab("President/Year")+
  theme_classic()+
  theme(axis.text.x = element_text(angle = 90))
```

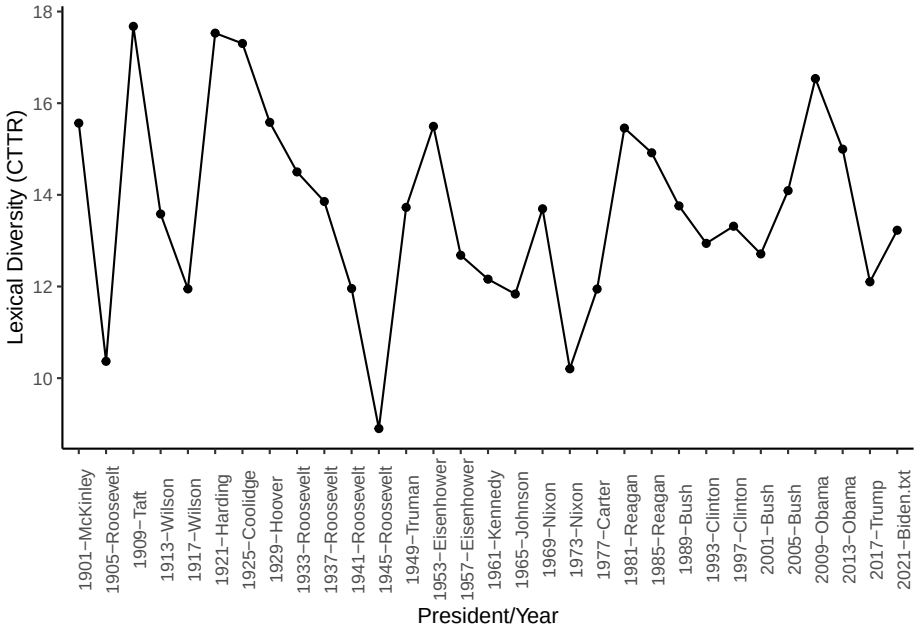


Another thing we can look at are the readability and lexical diversity of the texts. The former one of these refers to how readable a text is (i.e. how easy or difficult it is to read), while the latter tells us how many different types of words are used in the texts and thus how "diverse" the text is in word choice and use. Given that there are many ways to calculate both metrics, please have a look at the help file to see which one works best for you. Here, we will use the most popular:

```
corpus_readability <- textstat_readability(corpus_inaugural, measure = c("Flesch.Kinca
corpus_lexdiv <- textstat_lexdiv(data_inaugural_tokens, c("CTTR", "TTR", "MATTR"), MAT
```

As before, we can easily plot this data in a graph to see how lexical diversity developed over time:

```
ggplot(data=corpus_lexdiv, aes(x=document, y=CTTR, group=1)) +
  geom_line()+
  geom_point()+
  ylab("Lexical Diversity (CTTR)")+
  xlab("President/Year")+
  theme_classic()+
  theme(axis.text.x = element_text(angle = 90))
```



Another thing we can do is look at the similarities and distances between documents. With this, we can answer questions such as: how "different" are these documents from each other? And if different (or similar), how different (or similar)? The idea is that the larger the similarity is, the smaller the distance is as well. A good way to understand the idea of similarity is to consider how many operations you need to perform to change one text into the other. The more "replace" options you have to carry out, the more different the text. As for the distances, it is best to consider the texts as having positions on a Cartesian plane (with positions based on their word counts). The distance between these two points (either Euclidean, Manhattan or other) is then the distance between the texts.

Let's start with a look at these similarities (note again that there are many different methods to calculate this):

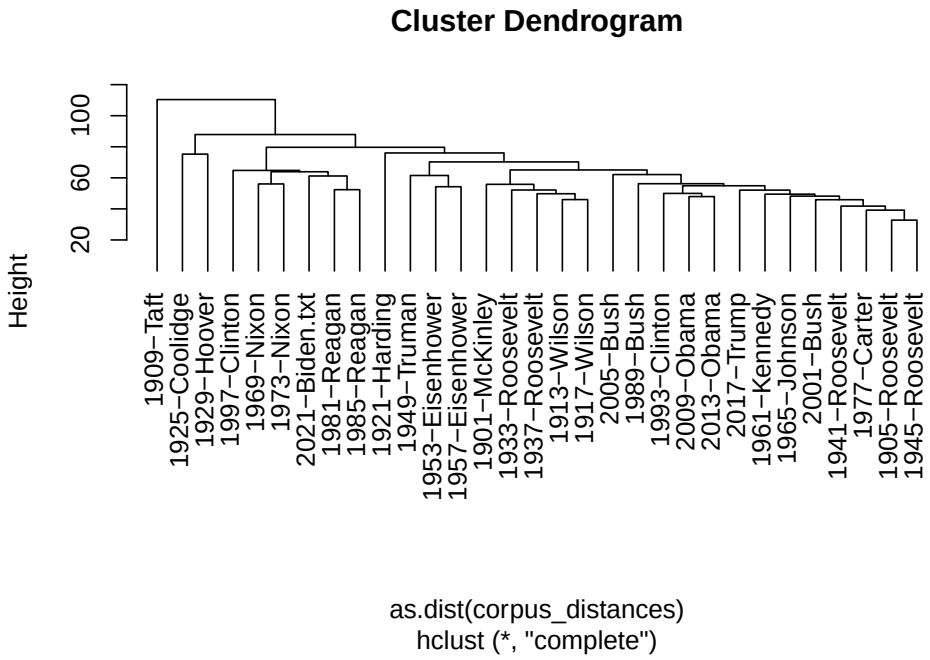
```
corpus_similarties <- textstat_simil(data_inaugural_dfm, method = "correlation", margin = "docume
corpus_similarties <- as.data.frame(corpus_similarties)
```

As brief look at these results tells us that the 1981 and 1985 Reagan speeches show the highest degree of similarity, while the
1945 Roosevelt and 2017 Trump speeches are the most different. Note that while we look here at the documents, we could also look at individual words (set `margin="features"`). For now, let us look at the distances between the documents, choosing the Euclidean distance between the documents as our metric:

```
corpus_distances <- textstat_dist(data_inaugural_dfm, margin = "documents", method = "euclidean")
corpus_distances_df <- as.data.frame(corpus_distances)
```

Here, we find the 1905 and 1945 Roosevelt speeches (the two different Roosevelts) to be the closest, and the 1909 Taft and 1997 Clinton speeches to be furthest apart. If we want to, we can even convert this data into a dendrogram. We do this by taking the information on the distances out of the `corpus_distances` object, make them into a triangular matrix, and plot them:

```
plot(hclust(as.dist(corpus_distances)), hang = -1)
```



Here, we can easily see that - amongst others - the 1909 Taft document is the "farthest" away from all the others, that the 1981 and 1985 Reagan speeches were very close, and that the 1997 Clinton speech was closer to Nixon's speeches, than his 1993 speech (which was close to the 2009 and 2013 Obama speeches).

Finally, let us look at the entropy of our texts. The entropy of a document measures the "amount" of information the is produced by each letter of the text. To get an idea of what this means, consider the "e" is a frequently occurring letter in an English text, while "z" is not. Thus, a word with a "z" in it, it more unique and thus likely to carry unique and interesting information. The "higher" the entropy of a text, the less "information" is in it:

```
corpus_entropy_docs <- textstat_entropy(data_inaugural_dfm, "documents")
corpus_entropy_docs <- as.data.frame(corpus_entropy_docs)
```

As we can see, the Roosevelt speeches had the lowest entropies, while the 1909 Taft and 1925 Coolidge speeches were the highest (in relative terms). While not as common as the other distances metrics, entropy is sometimes used to

measure the similarity between texts and can be useful if we want to know the importance of certain words. This because if a certain word is not "important," we might consider it to become a stop word:

```
corpus_entropy_feats <- textstat_entropy(data_inaugural_dfm, "features")
corpus_entropy_feats <- as.data.frame(corpus_entropy_feats)
corpus_entropy_feats <- corpus_entropy_feats[order(-corpus_entropy_feats$entropy),]
head(corpus_entropy_feats, 10)
```

```
##     feature  entropy
## 164  people 4.766391
## 488    life 4.747627
## 385  nation 4.737440
## 5     great 4.654392
## 114     can 4.651396
## 317  future 4.639222
## 197   world 4.616910
## 212    time 4.616614
## 402    must 4.610073
## 231     god 4.601430
```

Looking at the data, we find that "people," "life" and "nation" have pretty high entropies, indicating that the words added little in terms of information to the documents, and it would be a candidate to remove from our corpus.

Chapter 6

Dictionary Analysis

One of the simplest forms of quantitative text analysis is dictionary analysis. We can define dictionary methods as those which simply use the rate at which key words appear in a text to classify documents into categories or to measure the extent to which documents belong to particular categories, without making further assumptions. In many respects, dictionary methods present a non-statistical, categorical analysis approach.

One of the most well-known examples of using dictionary methods is the measuring the tone in newspaper articles, speeches, children's writings, and so on, by using the so-called sentiment analysis dictionaries. Another well-known example is the measuring of policy content in different documents as illustrated by the Policy Agendas Project dictionary (Albaugh et al. (2013)). Here, we will carry out three such analyses, the first a standard analysis and the other two focusing on sentiment. For the former, we will use political party manifestos, while for the latter we will use movie reviews and Twitter data.

6.1 Classical Dictionary Analysis

As for our dictionaries, we can either make the dictionary ourselves or use an off-the-shelf version. For the latter, we can either import the files we already have into R or use some of the versions that come with the `quanteda.dictionaries` package. For this, we first load the package:

```
library(quanteda.dictionaries)
```

We then apply one of these dictionaries to the document feature matrix we made earlier. As a dictionary, we will use the one made by Laver and Garry (2000), meant for estimating policy positions from political texts. We first load this dictionary into R and then run it on the dfm using the `dfm_lookup` command:

```
data_dictionary_LaverGarry
dictionary_results <- dfm_lookup(data_inaugural_dfm, data_dictionary_LaverGarry)
dictionary_results
```

Apart from off-the-shelf dictionaries, it is also possible to create our own which could suit our research question better. One approach in dictionary construction is to use prior theory deductively to come up with different categories and their associated words. Another approach is to use reference texts in order to come up with categories and words inductively. We can also combine different dictionaries as illustrated by Young and Soroka (2012), or different dictionaries and keywords from categories in manual coding scheme (Lind et al. (2019)). Finally, one can use expert or crowdcoding assessments to determine the words that best match different categories in a dictionary (Haselmayer and Jenny (2017)).

If we want to create our own dictionary in `quanteda` we use the same commands as above, but we first have to create the dictionary. To do so, we specify the words in a named list. This list contains keys (the words we want to look for) and the categories to which they belong. We then transform this list into a dictionary. Here, we choose some words which we believe will allow us to easily identify the different parties:

```
dic_list <- list(economy = c("tax*", "vat", "trade"), social = c("Medicare",
    "GP", "health"), devolution = c("states", "senate", "independence"),
    government = c("Washington", "Congress", "White House"))
dic_created <- dictionary(dic_list, tolower = FALSE)
dic_created
```

```
## Dictionary object with 4 key entries.
## - [economy]:
##   - tax*, vat, trade
## - [social]:
##   - Medicare, GP, health
## - [devolution]:
##   - states, senate, independence
## - [government]:
##   - Washington, Congress, White House
```

If you compare the `dic_list` file with the `data_dictionary_LaverGarry` file, you will find that it has the same structure. To see the result, we can use the same command:

```
dictionary_created <- dfm_lookup(data_inaugural_dfm, dic_created)
dictionary_created
```

```
## Document-feature matrix of: 31 documents, 4 features (42.74% sparse) and 4 docvars.
##                  features
## docs              economy social devolution government
```

```
##   1901-McKinley        2      0          9          9
##   1905-Roosevelt       0      0          0          1
##   1909-Taft           17      0         12         14
##   1913-Wilson          1      2          1          0
##   1917-Wilson          1      0          2          0
##   1921-Harding         7      0          2          1
## [ reached max_ndoc ... 25 more documents ]
```

6.2 Sentiment Analysis

The logic of dictionaries is that we can use them to see which kind of topics are present in our documents. Yet, we can also use them to provide us with measurements that are most often related to scaling. One way to do so is with *sentiment* analysis. Here, we look at whether a certain piece of text is happy, angry, positive, negative, and so on. One case in which this can help us is with movie reviews. These reviews give us a description of a movie and then tell us their opinion. Another is when we look at Twitter data, to capture the "mood of the moment." Here, we will look at both, starting with the movie reviews.

6.2.1 Movie Reviews

First, we load some reviews into R. The corpus we use here contains 50,000 movie reviews, each with a 1-10 rating (amongst others). As 50,000 reviews make the analysis quite slow, we will first select 30 reviews at random from this corpus. We do so via `corpus_sample`, after which we transform it via a tokens object into a dfm:

```
library(quanteda.classifiers)
reviews <- corpus_sample(data_corpus_LMRD, 30)
reviews_tokens <- tokens(reviews)
reviews_dfm <- dfm(reviews_tokens)
```

The next step is to load in a sentiment analysis dictionary. Here, we will use the Lexicoder Sentiment Dictionary, included in `quanteda` and run it on the dfm:

```
data_dictionary_LSD2015
results_dfm <- dfm_lookup(reviews_dfm, data_dictionary_LSD2015)
results_dfm
```

The next step is to convert the results to a data frame and view them:

```
sentiment <- convert(results_dfm, to="data.frame")
sentiment
```

```
##                   doc_id negative positive neg_positive neg_negative
## 1    test/pos/8273_9.txt        2        5            0            0
## 2    test/neg/6071_3.txt        5        8            0            0
## 3   test/pos/3535_10.txt        4       13            0            0
```

```
## 4   test/neg/10038_4.txt       15       10            0            0
## 5   train/pos/8503_8.txt       26       27            0            0
## 6    test/neg/9519_1.txt        5        4            0            0
## 7   train/neg/4256_4.txt        3       12            0            0
## 8   test/pos/1536_10.txt        5        5            0            0
## 9   train/neg/1414_3.txt        5        4            0            0
## 10  train/neg/4664_3.txt        5        5            0            0
## 11  train/neg/6683_1.txt       11        7            0            0
## 12   test/pos/4714_8.txt        7        2            0            0
## 13  train/neg/2414_3.txt       16       12            0            0
## 14  test/pos/9208_10.txt        6       10            0            0
## 15  train/pos/1102_8.txt       10       15            0            0
## 16  train/neg/9169_1.txt        7        2            0            0
## 17  train/pos/5187_7.txt        6       13            0            0
## 18   test/neg/3567_4.txt        7       10            0            0
## 19  train/neg/7802_1.txt       11        2            0            0
## 20  test/neg/10108_1.txt       17       12            0            0
## 21  train/neg/2454_4.txt       10       15            0            0
## 22 train/pos/9596_10.txt        1       10            0            0
## 23 train/pos/10286_9.txt        7       11            0            0
## 24 train/pos/6448_10.txt        3       12            0            0
## 25 train/pos/2569_10.txt        1        5            0            0
## 26  test/pos/7651_10.txt        4       16            0            0
## 27  test/neg/10588_1.txt        5        2            0            0
## 28   test/neg/8988_1.txt       15        8            0            0
## 29 train/neg/11305_1.txt       17        3            0            0
## 30  train/pos/108_10.txt        1        4            0            0
```

Since movie reviews usually come with some sort of rating (often in the form of stars), we can see if this relates to the sentiment of the review. To do so, we have to take the rating out of the dfm and place it in a new data-frame with the positive and negative sentiments:

```
star_data <- reviews_dfm@docvars$rating
stargraph <- as.data.frame(cbind(star_data, sentiment$negative, sentiment$positive))
names(stargraph) <- c("stars","negative","positive")
```

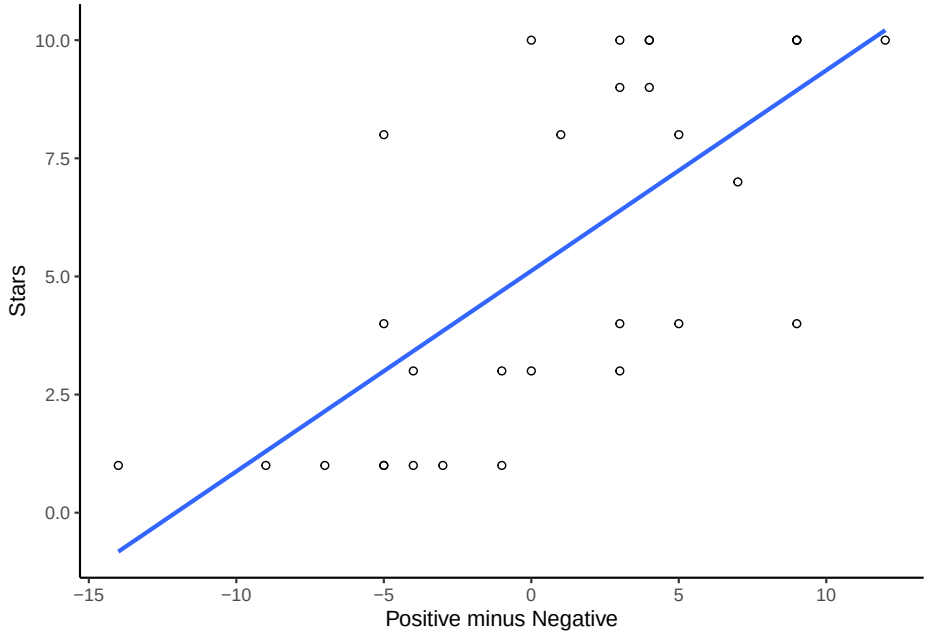
To compare the sentiment with the stars, we first have to combine the senitments into a scale. Of the many ways to do so, the simplest is to take the difference between the positive and negative words (positive – negative). Another option is to take the ratio of positive words against both positive and negative (positive/positive+negative). Here, we do both:

```
sentiment_difference <- stargraph$positive - stargraph$negative
sentiment_ratio <- (stargraph$positive/(stargraph$positive +
    stargraph$negative))
stargraph <- cbind(stargraph, sentiment_difference, sentiment_ratio)
```

Then, we can plot the ratings and the scaled sentiment measures together with a linear regression line:

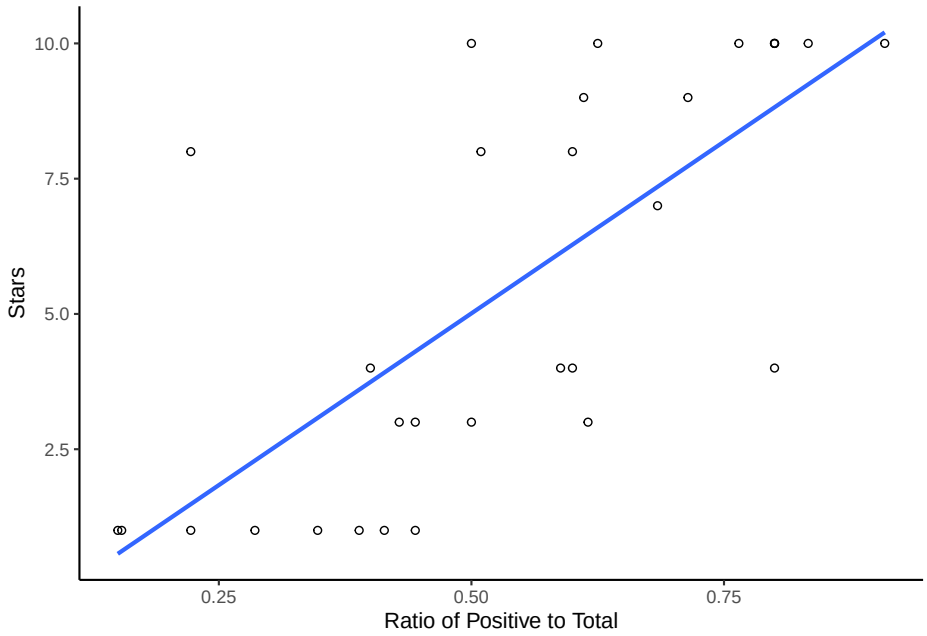
```
library(ggplot2)

ggplot(stargraph, aes(x = sentiment_difference, y = stars)) +
    geom_point(shape = 1) + geom_smooth(method = lm, se = FALSE) +
    xlab("Positive minus Negative") + ylab("Stars") + theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(stargraph, aes(x = sentiment_ratio, y = stars)) + geom_point(shape = 1) +
    geom_smooth(method = lm, se = FALSE) + xlab("Ratio of Positive to Total") +
    ylab("Stars") + theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Finally, we would like to illustrate how one can make inferences by using the output of a dictionary analysis, by estimating confidence intervals around the point estimates. To do so, again the first step is to add a column which will be the total of positive and negative words scored by the dictionary. We do so by copying the data frame to a new data frame and adding a new column filled with NA values:

```
reviews_bootstrap   <- sentiment
reviews_bootstrap$n <- NA
```

We then again specify the number of reviews, the replications that we want and change the data frame into an array:

```
library(combinat)

nman <- nrow(reviews_bootstrap)
nrepl <- 1000
manifBSn <- manifBSnRand <- array(as.matrix(reviews_bootstrap[,
    2:3]), c(nman, 2, nrepl + 1), dimnames = list(1:nman, names(reviews_bootstrap[,
    2:3]), 0:nrepl))
```

Then, we bootstrap the word counts for each movie review and compute percentages for each category using a multinomial draw:

```
n <- apply(manifBSn[1:nrow(manifBSn), , 1], 1, sum)
p <- manifBSn[, , 1]/n
```

```
for (i in 1:nrepl) {
    manifBSn[, , i] <- rmultinomial(n, p)
}
```

We can then ask R to compute the quantities of interest. These are standard errors for each category, as well as the percentage coded for each category.

```
NegativeSE <- apply(manifBSn[, "negative", ]/n * 100, 1, sd)
PositiveSE <- apply(manifBSn[, "positive", ]/n * 100, 1, sd)
perNegative <- apply(manifBSn[, "negative", ]/n * 100, 1, mean)
perPositive <- apply(manifBSn[, "positive", ]/n * 100, 1, mean)
```

We then save these quantities of interest in a new data frame:

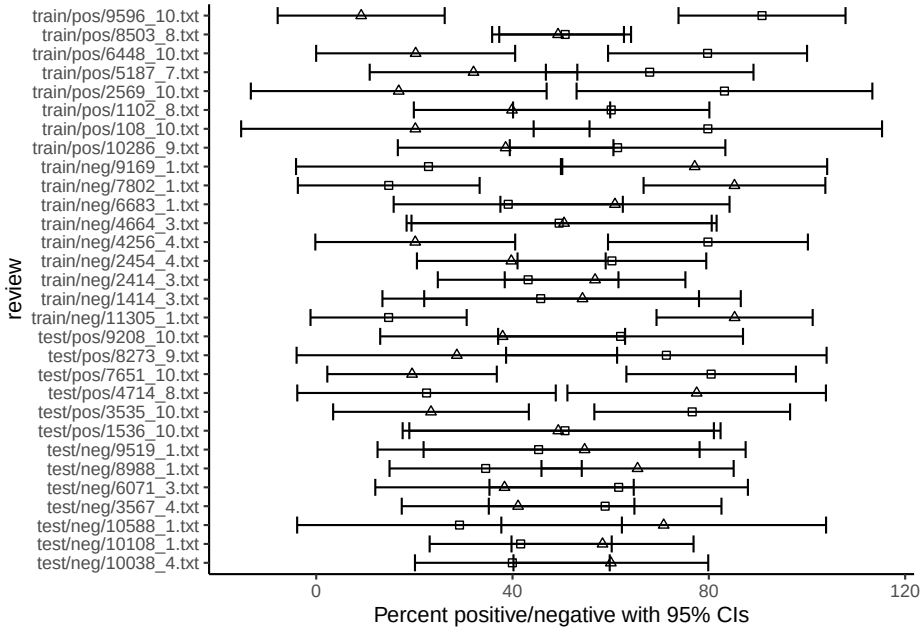
```
dataBS <- data.frame(cbind(reviews_bootstrap[, 1:3], NegativeSE,
    PositiveSE, perNegative, perPositive))
```

Then, we first calculate the confidence intervals and add these:

```
pos_hi <- dataBS$perPositive + (1.96 * dataBS$PositiveSE)
pos_lo <- dataBS$perPositive - (1.96 * dataBS$PositiveSE)
neg_lo <- dataBS$perNegative - (1.96 * dataBS$NegativeSE)
neg_hi <- dataBS$perNegative + (1.96 * dataBS$NegativeSE)
dataBS <- cbind(dataBS, pos_hi, pos_lo, neg_lo, neg_hi)
```

Finally, we can then make the graph. Here, we plot each of the positive and negative points and then overlay them with their error bars:

```
library(ggplot2)

ggplot() +
 geom_point(data = dataBS,aes(x = perPositive, y = doc_id), shape = 0) +
 geom_point(data = dataBS,aes(x = perNegative, y = doc_id), shape = 2) +
 geom_errorbarh(data = dataBS,aes(x = perPositive, xmax = pos_hi,xmin = pos_lo, y = doc_id)) +
 geom_errorbarh(data = dataBS,aes(x = perNegative, xmax = neg_hi,xmin = neg_lo, y = doc_id)) +
 xlab("Percent positive/negative with 95% CIs") +
 ylab("review")+
 theme_classic()
```



As can be seen in this particular example, the fact that some documents are much less lengthier than others introduces a lot of uncertainty in the estimates. As evident from the overlapping confidence intervals in the figure, for most reviews, the percentage of negative words is not much different from the percentage of positive words. In other words: the sentiment for these reviews is rather mixed.

6.2.2 Twitter

Now, let us turn to an example using Twitter data. Here, we will look at the major problems that have occurred to several of the major US airlines. For this, data was scraped from Twitter between 16 and 24 February of 2015. Then, using using the Crowdflower platform, contributors were asked to classify each tweet (their sentiment) as either negative, positive, or neutral, and, if negative, what their reason was for classifying it as such. In addition, the data-set also contains information on how "confident" coders were on their classification and reason, as well as information on the Airline, and some info on the Tweet. Finally, we get some information on the "gold" tweets, which are used by Crowdflower to figure out how well there coders are doing (that is, these tweets have been expert coded).

We download the data from its website (https://www.kaggle.com/crowdflower/twitter-airline-sentiment), but for ease-of-use, we also placed it on GitHub, so we can directly import it into R:

```
urlfile = "https://raw.githubusercontent.com/SCJBruinsma/qta-files/master/Tweets.csv"
tweets <- read.csv(url(urlfile))
```

Given that this is Twitter data, we have to do quite some cleaning in order to filter out everything we do not want. While we earlier on saw that we can perform cleaning on a corpus, we can also clean text in a dataframe directly (basically in any string), with R's in-house `gsub` command, which basically replaces parts of the string. To understand how this works, say that we want to remove all the mentions of websites from our tweets. We then do as such:

```
tweets$text <- gsub("http.*","",  tweets$text)
```

Thus, we substitute those strings that start with "http.*" (the asterisk denotes a wildcard, which means that anything can follow) and replace it with "" (that is, nothing). We do this for any string that is in `tweets$text`. Using this technique, we also remove slashes, punctuation, various symbols, "RT" (retweets), and references ("href"):

```
tweets$text <- gsub("https.*","", tweets$text)
tweets$text <- gsub("\\$", "", tweets$text)
tweets$text <- gsub("@\\w+", "", tweets$text)
tweets$text <- gsub("[[:punct:]]", "", tweets$text)
tweets$text <- gsub("[ |\t]{2,}", "", tweets$text)
tweets$text <- gsub("^ ", "", tweets$text)
tweets$text <- gsub(" $", "", tweets$text)
tweets$text <- gsub("RT", "", tweets$text)
tweets$text <- gsub("href", "", tweets$text)
```

We then transform our dataframe into a corpus (specifying that our text is in the `tweets$text` field), and then transform this into a tokens object, lower all the words, remove the stop words, and finally make it into a dfm:

```
corpus_tweets <- corpus(tweets, text_field = "text")
data_tweets_tokens <- tokens(corpus_tweets)
data_tweets_tokens <- tokens_tolower(data_tweets_tokens, keep_acronyms = TRUE)
data_tweets_tokens <- tokens_select(data_tweets_tokens, stopwords("english"), selection = "remove
data_tweets_dfm <- dfm(data_tweets_tokens)
```

Now we can apply our dictionary. We can do this in two ways: applying it to the dfm, and applying it to the tokens object. Both should give roughly similar results. Yet, given that `dfm_lookup()` cannot detect multi-word expressions (as the dfm gets rid of all word order), we can use the `tokens_lookup()` and then convert this into a dfm, to compensate for this. One reason we might want to do this here, is because the LSD2015 dictionary contains some multi word expressions that `dfm_lookup()` might miss. As a comparison, let's have a look at both:

```
results_tokens <- tokens_lookup(data_tweets_tokens, data_dictionary_LSD2015)
results_tokens <- dfm(results_tokens)
results_tokens <- convert(results_tokens, to="data.frame")

results_dfm <- dfm_lookup(data_tweets_dfm, data_dictionary_LSD2015)
results_dfm <- convert(results_dfm, to="data.frame")
```

Now, let us see how well our dictionary has done. To see this, we compare the sentiment of the tweet according to the dictionary with the sentiment assigned by the coder. We take this information out of our original data, and recode it (so it has got numerical values):

```
library(car)
```

```
## Caricamento del pacchetto richiesto: carData
```

```
##
## Caricamento pacchetto: 'car'
```

```
## Il seguente oggetto è mascherato da 'package:dplyr':
##
##     recode
```

```
## Il seguente oggetto è mascherato da 'package:purrr':
##
##     some
```

```
labels <- tweets$airline_sentiment
labels <- car::recode(labels, "'positive'=1;'negative'=-1;'neutral'=0")
table(labels)
```

```
## labels
##   -1    0    1
## 9178 3099 2363
```

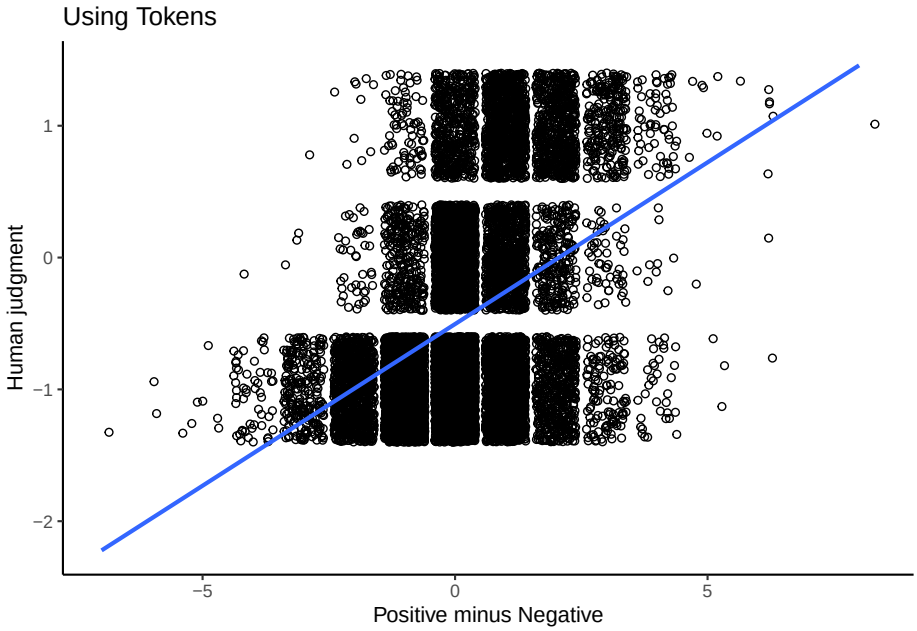
A quick look at the data (with `table()`) reveals that the majority of the tweets is negative (which is to be expected), a fair share neutral, and finally some positive ones. Now, let us bind this data to the output of our dictionary analysis, and calculate an overall score for each tweet by subtracting the positive score from the negative score (that is, the higher the score, the more positive the tweet):

```
comparison_tokens <- as.data.frame(cbind(results_tokens$positive, results_tokens$negati
difference_tokens <- results_tokens$positive - results_tokens$negative
comparison_tokens <- cbind(comparison_tokens, difference_tokens)

comparison_dfm <- as.data.frame(cbind(results_dfm$positive, results_dfm$negative, label
difference_dfm <- results_dfm$positive - results_dfm$negative
comparison_dfm <- cbind(comparison_dfm, difference_dfm)
```

Finally, we can place this all in a graph, in which we plot both the human judgement scores and the scores calculated by subtracting the positive and negative codes. In addition, we plot a simple linear equation to better understand the relation:

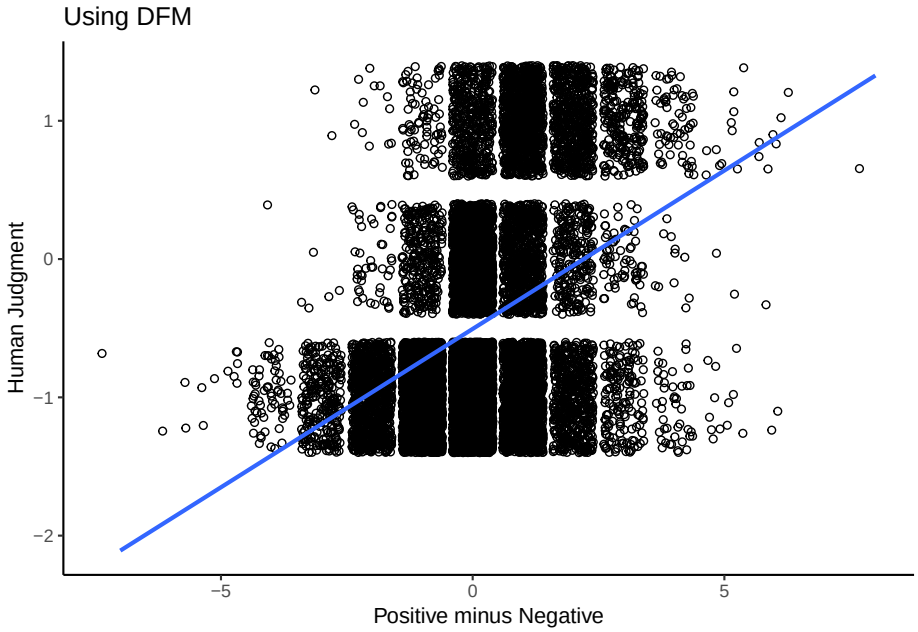
```
library(ggplot2)

ggplot(comparison_tokens ,aes(x = difference_tokens, y = labels)) +
  geom_jitter(shape = 1) +
  geom_smooth(method = lm, se = FALSE) +
  xlab("Positive minus Negative") +
  ylab("Human judgment") +
  ggtitle("Using Tokens")+
  theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(comparison_dfm, aes(x = difference_dfm, y = labels)) +
  geom_jitter(shape = 1) +
  geom_smooth(method = lm, se = FALSE) +
  xlab("Positive minus Negative") +
  ylab("Human Judgment") +
  ggtitle("Using DFM")+
  theme_classic()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



As we can see, there is a slightly positive relation (0.6947 for the tokens and 0.6914 for the dfm), which is a relatively good thing considering our approach does not involve any human coders at all. Though, how good it is of course depends on the benchmark that we set ourselves.

Chapter 7

Scaling

With a dictionary, we aimed to classify our texts into different categories based on the words they contain. While practical, there is no real way to compare these categories: one category is no better or worse than the other. If we do want to compare texts, we have to place them on some sort of scale. Here, we will look at three ways in which we can do so: *Wordscores* (Laver, Benoit, and Garry 2003), *Wordfish* (Slapin and Proksch 2008), and Correspondence Analysis. The first two methods used to be part of the main `quanteda` package, but have now moved to the `quanteda.textmodels` package, while we find CA in the `FactoMineR` package.

7.1 Wordscores

The idea of Wordscores is to use reference texts (from which we know the position) to position our virgin texts (from which we do not know the position). Here, we will use the data from the 2001 and 2005 party manifestos of the five largest parties in the United Kingdom, and will use the 2001 documents as reference texts and the 2005 documents as virgin texts. Also the scale we want to position our documents on is the general left-right scale. Thus, we need to know the positions for the 2001 documents on this. Here, we will use the left-right scale from the 2002 Chapel Hill Expert Survey (Bakker et al. 2012) to do so. So, we load our data, make the subset, transform it into a dfm, and clean it:

```
library(quanteda)
library(quanteda.corpora)

data(data_corpus_ukmanifestos)
corpus_manifestos <- corpus_subset(data_corpus_ukmanifestos,
    Year == 2001 | Year == 2005)
corpus_manifestos <- corpus_subset(corpus_manifestos, Party ==
```

```
    "Lab" | Party == "LD" | Party == "Con" | Party == "SNP" |
    Party == "PCy")

data_manifestos_tokens <- tokens(corpus_manifestos, what = "word",
    remove_punct = TRUE, remove_symbols = TRUE, remove_numbers = TRUE,
    remove_url = TRUE, remove_separators = TRUE, split_hyphens = FALSE,
    include_docvars = TRUE, padding = FALSE, verbose = TRUE)

data_manifestos_tokens <- tokens_tolower(data_manifestos_tokens,
    keep_acronyms = FALSE)
data_manifestos_tokens <- tokens_select(data_manifestos_tokens,
    stopwords("english"), selection = "remove")

data_manifestos_dfm <- dfm(data_manifestos_tokens)
```

Then , we check the order of the documents inside our dfm:

```
data_manifestos_dfm@Dimnames$docs
```

```
##  [1] "UK_natl_2001_en_Con" "UK_natl_2001_en_Lab" "UK_natl_2001_en_LD"
##  [4] "UK_natl_2001_en_PCy" "UK_natl_2001_en_SNP" "UK_natl_2005_en_Con"
##  [7] "UK_natl_2005_en_Lab" "UK_natl_2005_en_LD"  "UK_natl_2005_en_PCy"
## [10] "UK_natl_2005_en_SNP"
```

We can then set the scores for the reference texts. For the virgin texts, we set `NA` instead. Then, we run the wordscores model - providing the dfm and the reference scores - and save it into an object:

```
library(quanteda.textmodels)

scores <- c(7.72,5.18,3.82,3.2,3,NA,NA,NA,NA,NA)
ws <- textmodel_wordscores(data_manifestos_dfm, scores)
summary(ws)
```

```
##
## Call:
## textmodel_wordscores.dfm(x = data_manifestos_dfm, y = scores)
##
## Reference Document Statistics:
##                       score total min max   mean median
## UK_natl_2001_en_Con  7.72  7179   0  92 0.8606      0
## UK_natl_2001_en_Lab  5.18 16395   0 166 1.9654      0
## UK_natl_2001_en_LD   3.82 12337   0 101 1.4789      0
## UK_natl_2001_en_PCy  3.20  3508   0  72 0.4205      0
## UK_natl_2001_en_SNP  3.00  5693   0 108 0.6825      0
## UK_natl_2005_en_Con    NA  4350   0  46 0.5215      0
## UK_natl_2005_en_Lab    NA 13370   0 147 1.6027      0
```

```
## UK_natl_2005_en_LD     NA  9265   0 109 1.1106      0
## UK_natl_2005_en_PCy    NA  4204   0 148 0.5040      0
## UK_natl_2005_en_SNP    NA  1509   0  49 0.1809      0
##
## Wordscores:
## (showing first 30 elements)
##         time       common        sense conservative    manifesto introduction
##        5.838        6.540        7.376        7.161        4.478        3.982
##        lives      raising       family       living       safely      earning
##        6.047        4.427        5.519        4.719        5.743        6.046
##      staying      healthy      growing        older      knowing        world
##        6.946        4.294        4.745        6.280        7.720        4.366
##       leader     stronger      society         town      country    civilised
##        4.524        4.910        4.342        7.515        4.401        4.278
##        proud    democracy   conclusion      present    ambitious    programme
##        6.069        5.267        6.946        3.594        4.466        4.233
```

When we run the `summary` command, we can see the word scores for each word. This is the position of that word on our scale of interest. We then only need to figure out how often these words occur in each of the texts, add up their scores, and divide this by the total number of words of the texts. This gives us the *raw score* of the text. Yet, this raw score has some problems. Most important of which is that as some words occur in almost all texts, all the scores will be very clustered in the middle of our scale. To prevent this, we can spread out the scores again, so they look more like the scores of our reference texts. This rescaling has two versions. The first was the original as proposed by Laver, Benoit, and Garry (2003), and focuses on the variance of the scores. The idea here is that the distribution of the scores of the virgin texts has the correct mean, but an incorrect variance which needs rescaling. The second, proposed by Martin and Vanberg (2008), focuses on the extremes of the scores. What it does is to take the scores of the virgin texts and stretch them out to match the extremes of the scores of the reference texts. Here, we run both so we can compare them. For the MV transformation, we will calculate the standard errors for the scores as well:

```
pred_lbg <- predict(ws, rescaling = "lbg")
```

```
## Warning: 2203 features in newdata not used in prediction.
```

```
pred_mv <- predict(ws, rescaling = "mv", se.fit = TRUE, interval = "confidence")
```

```
## Warning: 2203 features in newdata not used in prediction.
```

```
## Warning in predict.textmodel_wordscores(ws, rescaling = "mv", se.fit = TRUE, :
## More than two reference scores found with MV rescaling; using only min, max
## values.
```

```
pred_lbg
```

```
## UK_natl_2001_en_Con UK_natl_2001_en_Lab  UK_natl_2001_en_LD UK_natl_2001_en_PCy
##            8.794566            5.440327            3.971305            1.921840
## UK_natl_2001_en_SNP UK_natl_2005_en_Con UK_natl_2005_en_Lab  UK_natl_2005_en_LD
##            2.166928            5.656940            5.128174            5.047475
## UK_natl_2005_en_PCy UK_natl_2005_en_SNP
##            3.752962            4.289754
```

```
pred_mv
```

```
## $fit
##                          fit      lwr      upr
## UK_natl_2001_en_Con 7.720000 7.633952 7.806048
## UK_natl_2001_en_Lab 5.331214 5.295467 5.366960
## UK_natl_2001_en_LD  4.285022 4.243678 4.326365
## UK_natl_2001_en_PCy 2.825456 2.750505 2.900406
## UK_natl_2001_en_SNP 3.000000 2.932910 3.067090
## UK_natl_2005_en_Con 5.485479 5.387391 5.583567
## UK_natl_2005_en_Lab 5.108908 5.060253 5.157564
## UK_natl_2005_en_LD  5.051437 4.989127 5.113747
## UK_natl_2005_en_PCy 4.129525 4.039903 4.219146
## UK_natl_2005_en_SNP 4.511812 4.323620 4.700003
##
## $se.fit
## UK_natl_2001_en_Con UK_natl_2001_en_Lab  UK_natl_2001_en_LD UK_natl_2001_en_PCy
##          0.04390309          0.01823830          0.02109396          0.03824078
## UK_natl_2001_en_SNP UK_natl_2005_en_Con UK_natl_2005_en_Lab  UK_natl_2005_en_LD
##          0.03423029          0.05004577          0.02482464          0.03179121
## UK_natl_2005_en_PCy UK_natl_2005_en_SNP
##          0.04572606          0.09601792
```

Note that this does not only predict the 2005 texts, but also the 2001 texts. As such, we can use these scores to see how well this procedure can recover the original scores. One reason why this might be a problem is because of a warning you most likely received. This says that "n features in newdata not used in prediction." This is as the method does not use all the words from the reference texts to score the virgin texts. Instead, it only uses the words that occur in them both. Thus, when we compare the reference scores with the scores the method gives to the reference documents, can see how well the method does.

To compare the scores, we will use the Concordance Correlation Coefficient as developed by Lin (1989). This coefficient estimates how far two sets of data deviate from a line of 45 degrees (which indicates perfect agreement). To calculate this, we take the scores (here we take the LBG version) from the object we created and combine them with the original scores. From this, we only select the first five texts (those from 2001) and calculate the CCC:

```
library(DescTools)
```

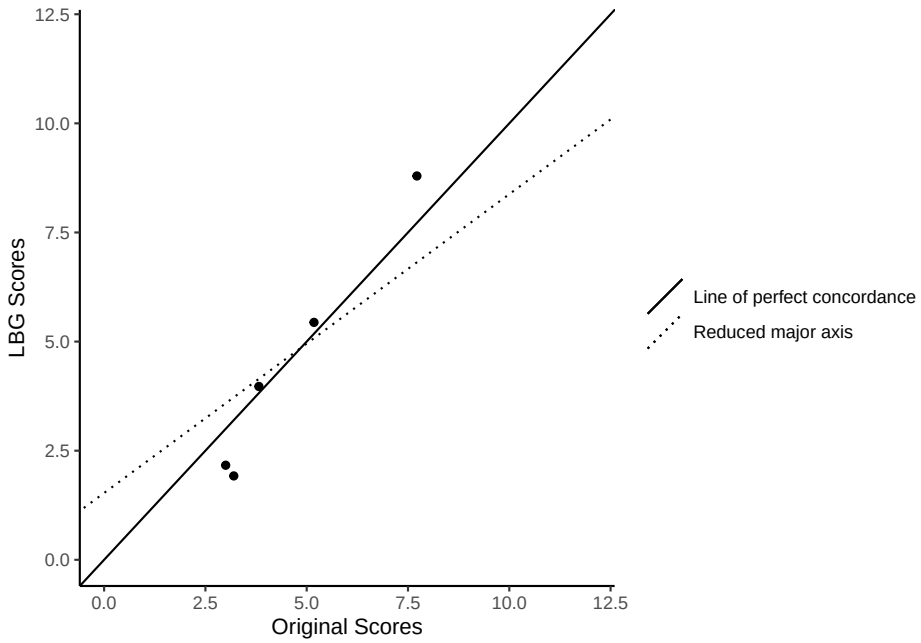
```
comparison <- as.data.frame(cbind(pred_lbg, scores))
comparison <- comparison[1:5, ]

CCC(comparison$scores, comparison$pred_lbg, ci = "z-transform",
    conf.level = 0.95, na.rm = TRUE)
```

```
## $rho.c
##         est    lwr.ci   upr.ci
## 1 0.9239205 0.8242101 0.968064
##
## $s.shift
## [1] 1.443978
##
## $l.shift
## [1] -0.05966866
##
## $C.b
## [1] 0.9345491
##
## $blalt
##       mean      delta
## 1 8.257283 -1.0745660
## 2 5.310163 -0.2603267
## 3 3.895653 -0.1513052
## 4 2.560920  1.2781600
## 5 2.583464  0.8330719
```

The result here is not bad, though the confidence intervals are rather large. We can have a further look at why this is the case by plotting the data. In this plot, we will show the position of the texts, as well as a 45-degree line. Also, we plot the reduced major axis, which shows the symmetrical relationship between the two variables. This line is a linear regression, which we compute first using the `lm` command:

```
library(ggplot2)

lm_line <- lm(comparison$scores ~ comparison$pred_lbg)

ggplot(comparison, aes(x=scores, y=pred_lbg)) +
 geom_point()+
 xlab("Original Scores")+
 ylab("LBG Scores")+
 ylim(0, 12)+
 xlim(0, 12)+
```

```
geom_abline(aes(intercept = 0,
                slope =1,
                linetype = "dashed"))+
geom_abline(aes(intercept = lm_line$coefficients[1],
                slope = lm_line$coefficients[2],
                linetype = "solid" ))+
scale_shape_manual(name = "",
                   values=c(1,3),
                   breaks=c(0,1),
                   labels=c("Line of perfect concordance" , "Reduced major axis"))+
scale_linetype_manual(name = "",
                      values=c(1,3),
                      labels=c("Line of perfect concordance" , "Reduced major axis"))+
theme_classic()
```

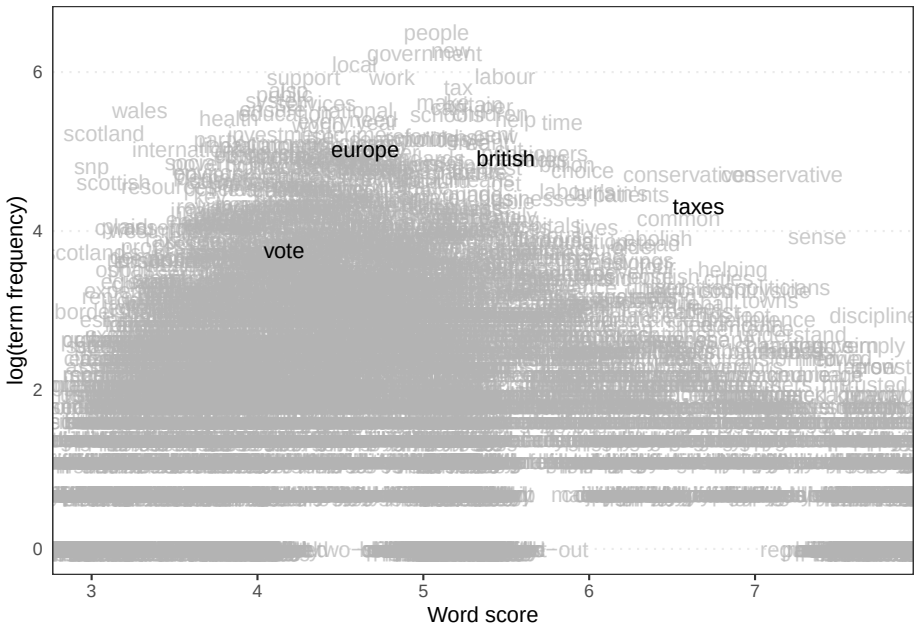


This graph allows us to spot the problem. That is that while we gave the manifesto for Plaid Cymru (PCy) a reference score of 3.20, Wordscores gave it 1.91. Removing this manifesto from our data-set would thus improve our estimates.

Apart from positioning the texts, we can also have a look at the words themselves. We can do this with the `textplot_scale1d` command, for which we also specify some words to highlight:

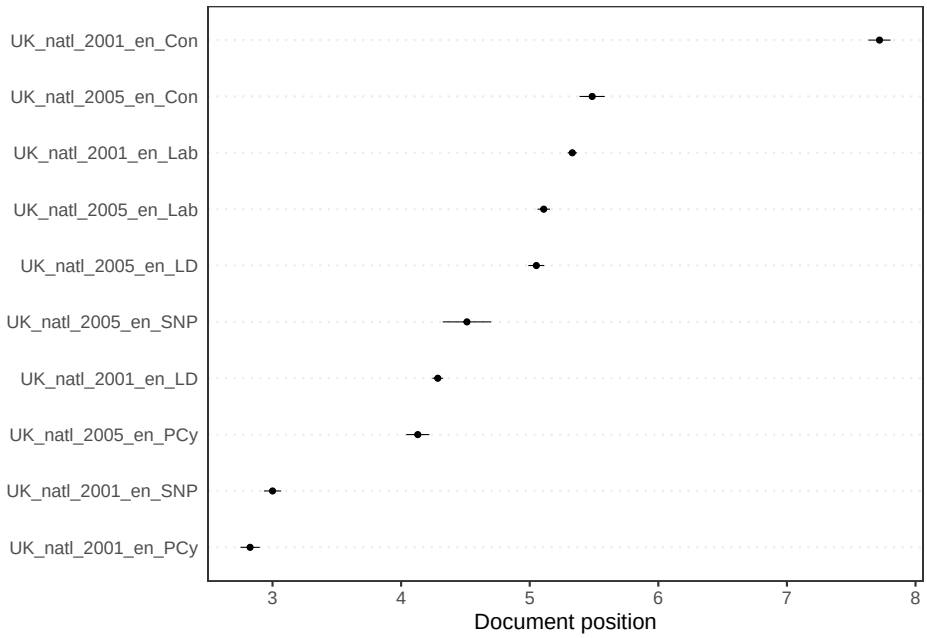
```
library(quanteda.textplots)
```

```
textplot_scale1d(ws, margin = "features", highlighted = c("british",
    "vote", "europe", "taxes"))
```



Finally, we can have a look at the confidence intervals around the scores we created. For this, we use the same command as above, though instead of specifying `features` (referring to the words), we specify `texts`. Note that we can only do this for the MV scores, as only here we also calculated the standard errors:

```
textplot_scale1d(pred_mv, margin = "documents")
```



Note that we can also make this graph ourselves. This requires some data-wrangling using the `dplyr` package. This package allows us to use pipes, which are denoted by the `%>%` command. This pipe transports an output of a command to another one before saving it. This saves us from constructing too many intermediate data-sets. Thus, here we first bind together the row names of the fit (which denote the documents), the fit itself, and the standard error of the fit (which also includes the lower and upper bound). We then transform this into a tibble (which is similar to a dataframe), rename the first and fifth columns, and finally ensure that all the values (which are still characters), are numeric (and year a factor):

```
library(dplyr)

data_textplot <- cbind(rownames(as.data.frame(pred_mv$se.fit)), pred_mv$fit, pred_mv$se
  as_tibble() %>%
  rename(id = 1,
         se = 5) %>%
  mutate(fit = as.numeric(fit),
         lwr = as.numeric(lwr),
         upr = as.numeric(upr),
         se = as.numeric(se),
         year = as.factor(stringr::str_sub(id, start = 9, end = 12)))
```

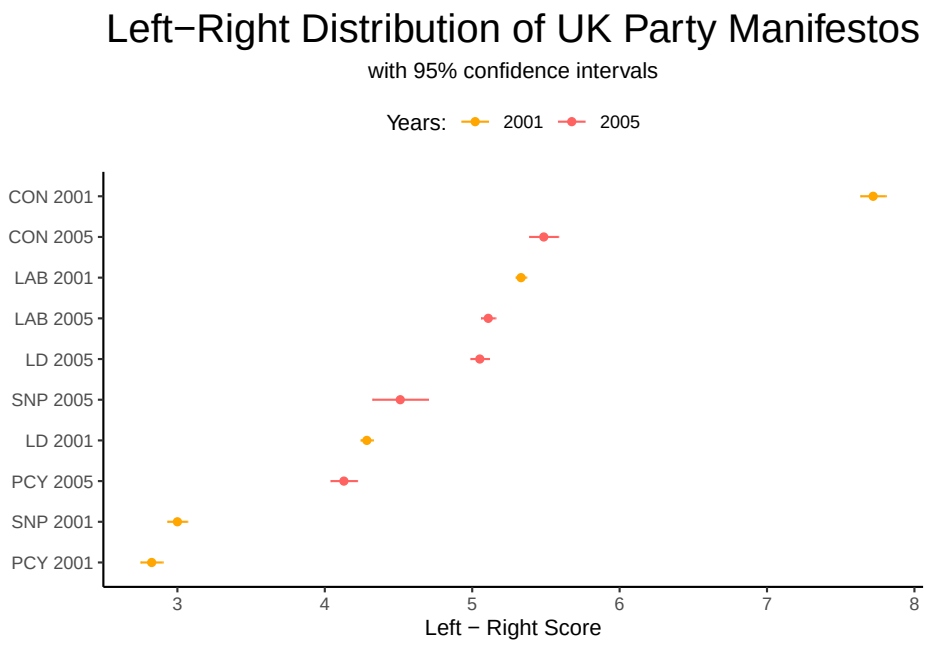
```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if
## Using compatibility `.name_repair`.
```

If we now look at our `data_textplot` object, we see that we have all the data we need: the fit (the average value), the lower and upper bounds, the year and the id that tells us with which party and year we are dealing. The only thing that we perhaps can do is to give the parties slightly better names. To see the current ones, type `data_textplot$id` in the console. We can then give them different names (just ensure that the order remains the same). We then sort them in decreasing order based on their fit:

```
data_textplot$id <- as.character(c("CON 2001", "LAB 2001", "LD 2001", "PCY 2001", "SNP 2001", "C
data_textplot$id <- with(data_textplot,  reorder(id, fit))
```

Then, we can plot this data using `ggplot`:

```
ggplot() +
  geom_point(data = data_textplot, aes(x = fit, y = id, colour = year)) +
  geom_errorbarh(data = data_textplot, aes(xmax = upr, xmin = lwr,  y = id, colour = year), heigh
  theme_classic() +
  scale_colour_manual(values = c("#ffa600", "#ff6361"),
                      name = "Years:",
                      breaks = c("2001", "2005"),
                      labels = c("2001", "2005")) +
  labs(title = "Left-Right Distribution of UK Party Manifestos",
       subtitle = "with 95% confidence intervals",
       x = "Left - Right Score",
       y = NULL) +
  theme_classic()+
  theme(plot.title = element_text(size = 20, hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5),
        legend.position = "top")
```



7.2 Wordfish

Different from Wordscores, for Wordfish we do not need any reference text. Instead of this, the method using a model (based on a Poisson distribution) to calculate the scores for the texts. The only thing we have to tell Wordfish is which texts define the extremes of our scale. While this might seems very practical, it also leaves us with a problem: which scale do we want? For example, let us say that we consider our corpus of inaugural speeches of American presidents we saw earlier. What scale should we be interested in? Let us for now say that we care about a general left-right position. As benchmarks, we then set the 1965 Johnson speech as the most "left" and the 1985 Reagan speech as the most "right." Also, we set a seed as the model draws random numbers and we want our work to be replicable:

```
set.seed(42)

data_inaugural_dfm@Dimnames$docs
```

```
##  [1] "1901-McKinley"   "1905-Roosevelt"  "1909-Taft"       "1913-Wilson"
##  [5] "1917-Wilson"     "1921-Harding"    "1925-Coolidge"   "1929-Hoover"
##  [9] "1933-Roosevelt"  "1937-Roosevelt"  "1941-Roosevelt"  "1945-Roosevelt"
## [13] "1949-Truman"     "1953-Eisenhower" "1957-Eisenhower" "1961-Kennedy"
## [17] "1965-Johnson"    "1969-Nixon"      "1973-Nixon"      "1977-Carter"
## [21] "1981-Reagan"     "1985-Reagan"     "1989-Bush"       "1993-Clinton"
## [25] "1997-Clinton"    "2001-Bush"       "2005-Bush"       "2009-Obama"
```

```
## [29] "2013-Obama"      "2017-Trump"      "2021-Biden.txt"
```

```
wordfish <- textmodel_wordfish(data_inaugural_dfm, dir = c(17,
    22))
summary(wordfish)
```

```
##
## Call:
## textmodel_wordfish.dfm(x = data_inaugural_dfm, dir = c(17, 22))
##
## Estimated Document Positions:
##                    theta      se
## 1901-McKinley     1.5265 0.03498
## 1905-Roosevelt    0.5503 0.07556
## 1909-Taft         2.1184 0.01580
## 1913-Wilson       0.9487 0.05162
## 1917-Wilson       0.8646 0.05786
## 1921-Harding      1.3072 0.03094
## 1925-Coolidge     1.4402 0.02757
## 1929-Hoover       1.5401 0.02740
## 1933-Roosevelt    1.1203 0.04539
## 1937-Roosevelt    0.6489 0.05203
## 1941-Roosevelt    0.1151 0.06610
## 1945-Roosevelt   -0.1673 0.10037
## 1949-Truman       0.8432 0.04453
## 1953-Eisenhower   0.2793 0.04693
## 1957-Eisenhower   0.1598 0.05757
## 1961-Kennedy     -0.5388 0.05631
## 1965-Johnson     -0.7804 0.05190
## 1969-Nixon       -0.9598 0.03916
## 1973-Nixon       -0.4417 0.05235
## 1977-Carter      -0.3437 0.06414
## 1981-Reagan      -0.6960 0.04169
## 1985-Reagan      -0.6409 0.04007
## 1989-Bush        -0.8890 0.03935
## 1993-Clinton     -1.1441 0.04009
## 1997-Clinton     -0.8663 0.03910
## 2001-Bush        -0.7422 0.04953
## 2005-Bush        -0.4094 0.04772
## 2009-Obama       -1.0796 0.03436
## 2013-Obama       -1.0532 0.03719
## 2017-Trump       -1.3810 0.03639
## 2021-Biden.txt   -1.3289 0.02999
##
## Estimated Feature Scores:
##      fellow-citizens assembled    4th   march   great anxiety  regard currency
```

```
## beta           2.084   -0.2363  1.308 -0.0627 -0.1155  0.5387  0.8081    2.057
## psi           -4.874   -2.7959 -4.776 -1.7305  1.5216 -2.7801 -2.2013   -3.848
##       credit    none  exists     now treasury receipts inadequate    meet
## beta  0.8941  0.4225 -0.7698 -0.5027    1.468    2.291      0.180 -0.3738
## psi  -2.1652 -2.1725 -2.4684  1.2740   -4.312   -5.629     -2.235  0.1264
##      current obligations government sufficient   public   needs surplus instead
## beta   1.455      0.5204     0.1712     0.5477  0.42723 -0.2554   1.308 -1.0352
## psi   -3.600     -1.0900     1.6961    -1.4018 -0.07204 -0.7792  -4.776 -0.9148
##      deficit    felt constrained convene congress extraordinary
## beta -0.1073  0.3725       1.308   1.308   0.7189        0.7233
## psi  -1.4313 -2.3578      -4.776  -4.776  -0.2453       -2.9579
```

Here, *theta* gives us the position of the text. As with Wordscores, we can also calculate the confidence intervals (note that *theta* is now called *fit*):

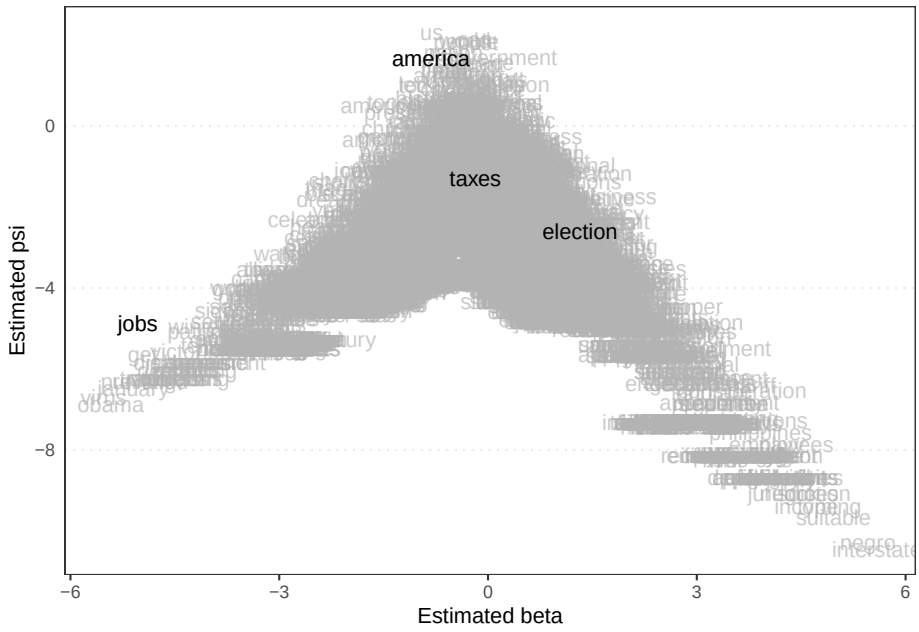
```
pred_wordfish <- predict(wordfish, interval = "confidence")
pred_wordfish
```

```
## $fit
##                         fit         lwr        upr
## 1901-McKinley     1.5264549  1.45790117  1.5950087
## 1905-Roosevelt    0.5503249  0.40223178  0.6984181
## 1909-Taft         2.1183666  2.08740335  2.1493298
## 1913-Wilson       0.9487067  0.84753556  1.0498779
## 1917-Wilson       0.8645950  0.75120033  0.9779896
## 1921-Harding      1.3071772  1.24653710  1.3678173
## 1925-Coolidge     1.4402120  1.38616969  1.4942542
## 1929-Hoover       1.5400842  1.48638959  1.5937789
## 1933-Roosevelt    1.1202695  1.03130137  1.2092377
## 1937-Roosevelt    0.6488960  0.54692619  0.7508658
## 1941-Roosevelt    0.1151337 -0.01441162  0.2446789
## 1945-Roosevelt   -0.1673134 -0.36404388  0.0294171
## 1949-Truman       0.8431987  0.75592790  0.9304694
## 1953-Eisenhower   0.2792604  0.18727234  0.3712484
## 1957-Eisenhower   0.1598369  0.04700292  0.2726709
## 1961-Kennedy     -0.5387860 -0.64916020 -0.4284118
## 1965-Johnson     -0.7804433 -0.88215746 -0.6787291
## 1969-Nixon       -0.9598015 -1.03655447 -0.8830485
## 1973-Nixon       -0.4416846 -0.54428665 -0.3390825
## 1977-Carter      -0.3436878 -0.46940309 -0.2179725
## 1981-Reagan      -0.6960482 -0.77775955 -0.6143369
## 1985-Reagan      -0.6409237 -0.71946264 -0.5623847
## 1989-Bush        -0.8890203 -0.96614097 -0.8118997
## 1993-Clinton     -1.1441087 -1.22269262 -1.0655248
## 1997-Clinton     -0.8663188 -0.94294521 -0.7896924
## 2001-Bush        -0.7422147 -0.83928343 -0.6451460
## 2005-Bush        -0.4094409 -0.50296128 -0.3159205
```

```
## 2009-Obama      -1.0796208 -1.14695763 -1.0122840
## 2013-Obama      -1.0531958 -1.12609628 -0.9802953
## 2017-Trump      -1.3810308 -1.45235286 -1.3097088
## 2021-Biden.txt  -1.3288773 -1.38765502 -1.2700996
```

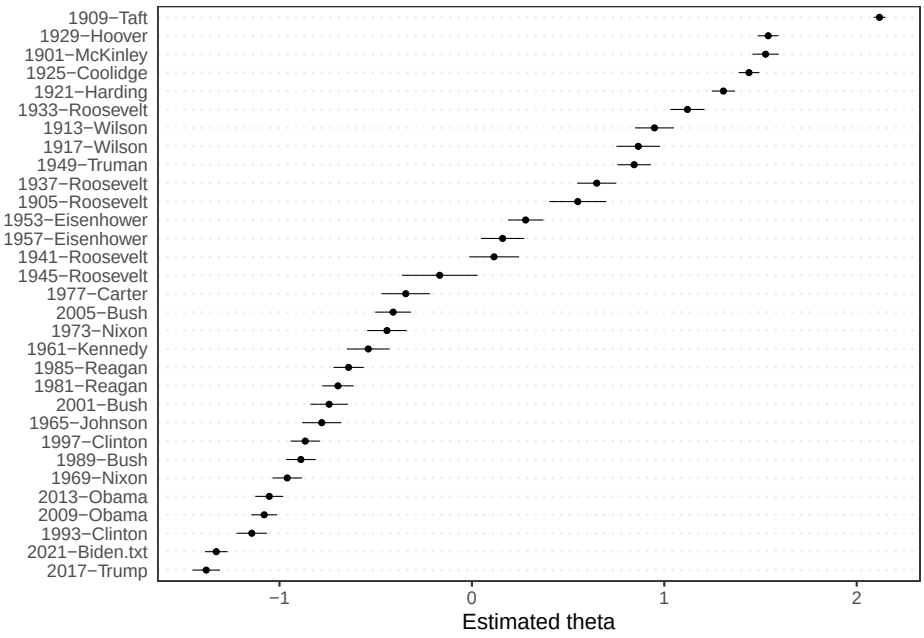
As with Wordscores, we can also plot graphs for Wordfish, using the same commands. The first graph we will again be looking at is the distribution of the words, which here forms an "Eifel Tower" like graph:

```
textplot_scale1d(wordfish, margin = "features", highlighted = c("america",
    "jobs", "taxes", "election"))
```



And then we can do the same for the documents as well. Note that we can also make a similar graph to the one we made ourselves above (just replace `pred_mv` with `pred_wordfish`):

```
textplot_scale1d(wordfish, margin = "documents")
```



Looking at the results here gives us an interesting picture. Remember that we chose our benchmark texts to look at the left-right position of our texts? Here, we see that both these texts (the 1965 Johnson and 1985 Reagan) are actually quite close to each other. Sticking with our interpretation that Reagan is more right-wing than Johnson, this would mean that the 1909 Taft address was the most right-wing and the 2017 Trump text the most left wing. Whether this is true is of course up to our own interpretation.

7.3 Correspondence Analysis

Correspondence Analysis has a similar logic as Principal Component Analysis. Yet, while PCA requires metric data, CA only requires nominal data (such as text). The idea behind both is to reduce the complexity of the data by looking for new dimensions. These dimensions should then explain as much of the original variance that is present in the data as possible. Within R many packages can run CA (such as the `ca` and `FactoMineR` packages and even `quanteda.textmodels`). One interesting package is the `R.temis` package. The `R.temis` package is interesting as it aims to bring the techniques of qualitative text analysis into R. Thus, the package focus on the import of corpus from programs such as Alceste (https://www.image-zafar.com/Logicieluk.html) and sites such as LexisNexis (https://www.lexisnexis.com) - programs that are often used in qualitative text analysis. The package itself is build on the popular `tm` package and has a largely similar logic.

To carry out the Correspondence Analysis, `R.temis` uses the `FactoMineR` and

`factoextra` packages. Here, we will look at an example with these packages using data from the an article on the stylistic variations in the Twitter data of Donald Trump between 2009 and 2018 (Clarke and Grieve 2019). Here, the authors aimed to figure out whether the way Trump's tweets were written fluctuated over time. To do so, they downloaded 21,739 tweets and grouped them into 63 categories over 4 dimensions based on their content. Given that all the data used in the article is available for inspection, we can attempt to replicate part of the analysis here.

First, we load the packages we need for the Correspondence Analysis:

```
library(FactoMineR)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(readr)
```

Then, we import the data. You can do so either by downloading the replication data yourselves, or use the file we already put up on GitHub:

```
urlfile = "https://raw.githubusercontent.com/SCJBruinsma/qta-files/master/TRUMP_DATA.txt"
tweets <- read_csv(url(urlfile))
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   TWEETID = col_double(),
##   WORDCOUNT = col_double(),
##   DATE = col_date(format = ""),
##   TIME = col_time(format = ""),
##   RETWEET = col_double(),
##   FAV = col_double()
## )
## i Use `spec()` for the full column specifications.
```

This data-set contains quite some information we do not need. To begin with, we remove all those variables that do not contain information about the 63 categories and the length of the tweet in words. Also, for clarity's sake, we sample 200 of the tweets:

```
tweets <- tweets[sample(nrow(tweets), 200), ]
tweets_mat <- tweets[,2:65]
```

We can then run the MCA with the `FactoMineR` package. For this, we have to give the data-set and the number of dimensions we think are in the data. We can set the latter either by establishing the dimensions as in a regular PCA (for example through a scree plot) or based on theory. Here we combine both and use the 5 dimensions established in the article. In addition, we set a supplementary

quantitative variable as `quanti.sup=1`. As this is a quantitative variable, it is not taken into consideration by the MCA, but does allow us to assess later on how it correlates with each of the five dimensions:

```
mca_tweets <- MCA(tweets_mat, ncp=5, quanti.sup=1, graph = FALSE)
```

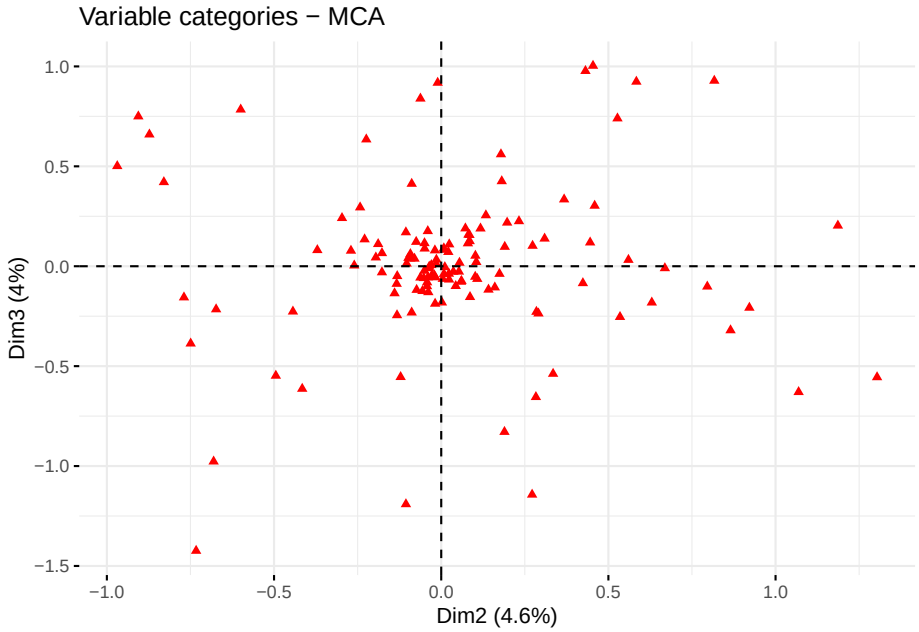
First, let's start by looking at the association of the wordlength with the five dimensions:

```
mca_tweets$quanti.sup
```

```
## $coord
##                Dim 1      Dim 2      Dim 3      Dim 4       Dim 5
## WORDCOUNT 0.8668104 -0.2107786 0.02090331 0.03345767 -0.05383935
```

As we can see, the word length has a strong correlation with Dimension 1. This basically means that this dimension captures the length of the words and not a seperate dimension we are interested in. Thus, when we want to look at the correspondence between the categories and the dimensions, we can ignore this dimension. Thus, for the MCA, we will look at dimensions 2 and 3:

```
fviz_mca_var(mca_tweets,
             repel = TRUE,
             geom = c("point"),
             axes = c(2, 3),
             ggtheme = theme_minimal())
```



Here, we only plot the points as adding the labels as well will make the picture

quite cluttered. In the article, Dimension 2 is identified as "Conversational Style" and Dimension 3 as "Campaigning Style." The plot thus nicely shows us that some categories belong to one of these dimensions and not to the other. To see for which cases this is mostly the case (the ones that have the most extreme positions), we can have a look at their coordinates:

```
var <- get_mca_var(mca_tweets)
coordinates <- as.data.frame(var$coord)
coordinates <- coordinates[order(coordinates$`Dim 2`),]
head(coordinates)
```

```
##                    Dim 1      Dim 2      Dim 3        Dim 4       Dim 5
## POSESPRPN_P    0.03383985 -0.9689511  0.5011988  0.036488377  0.36083073
## PROGRESSIVE_P  0.79201463 -0.9059134  0.7499092 -0.674316668  0.46134193
## SUPERLATIVE_P  0.37755724 -0.8728057  0.6594311 -0.147514294 -0.41714831
## GERUND_P       0.20213101 -0.8295159  0.4209137  0.089642937  0.82377111
## MULTIWVB_P     0.42162442 -0.7694130 -0.1558600 -0.003865414  0.37457619
## COLON_P       -0.24915096 -0.7493897 -0.3866968  1.023830211 -0.07073923
```

Here, remember to look only at the results from the second column onward. Here, we see that one extreme category for the second dimension (Conversational Style) was the use of a colon (:) or possessive proper nouns (such as Hillary's). This seems to fit well with the idea of conversational style. We can also see that the latter one also corresponds quite well with Dimension 3 (Campaigning Style), while the first one does not.

Chapter 8

Supervised Methods

While with scaling we try to place our texts on a scale, with supervised methods we go back to what we did with dictionary analysis: classification. Within `quanteda` there are many different models for supervised methods, of which we will cover two. These are Support Vector Machines (SVM) and Naive Bayes (NB). The first classifies texts by looking at their position on a hyperplane, the second by their (Bayesian) probabilities.

8.1 Support Vector Machines

To show how SVM works, we will look at an example of SVM in `quanteda` and one in `RTextTools`, and an example of NB in `quanteda`.

8.1.1 SVM with RTextTools

For the SVM, we will start with an example using our Twitter data and the `RTextTools` package. First, we load the Twitter data:

```
library("RTextTools")
```

```
## Caricamento del pacchetto richiesto: SparseM
```

```
##
## Caricamento pacchetto: 'SparseM'
```

```
## Il seguente oggetto è mascherato da 'package:base':
##
##     backsolve
```

```
## Registered S3 method overwritten by 'tree':
##   method     from
##   print.tree cli
```

```
library("car")

urlfile = "https://raw.githubusercontent.com/SCJBruinsma/qta-files/master/Tweets.csv"
tweets <- read.csv(url(urlfile))
tweets$text <- gsub("http.*","",  tweets$text)
tweets$text <- gsub("https.*","", tweets$text)
tweets$text <- gsub("\\$", "", tweets$text)
tweets$text <- gsub("@\\w+", "", tweets$text)
tweets$text <- gsub("[[:punct:]]", "", tweets$text)
tweets$text <- gsub("[ |\t]{2,}", "", tweets$text)
tweets$text <- gsub("^ ", "", tweets$text)
tweets$text <- gsub(" $", "", tweets$text)
tweets$text <- gsub("RT", "", tweets$text)
tweets$text <- gsub("href", "", tweets$text)

labels <- tweets$airline_sentiment
labels <- car::recode(labels, "'positive'=1;'negative'=-1;'neutral'=0")
```

The goal of the supervised learning task is to use part of this dataset to train a certain algorithm, and then use the trained algorithm to assign categories to the remaining sentences. Since we know the coded categories for the remaining sentences, we will be able to evaluate how well this training was in guessing/estimating what the codes for these sentences were. We start by creating a document term matrix;

```
doc_matrix <- create_matrix(tweets$text, language = "english", removeNumbers = TRUE, st
```

```
## Warning in TermDocumentMatrix.SimpleCorpus(x, control): custom functions are
## ignored
```

```
## Warning in TermDocumentMatrix.SimpleCorpus(x, control): custom tokenizer is
## ignored
```

```
doc_matrix
```

```
## <<DocumentTermMatrix (documents: 14640, terms: 694)>>
## Non-/sparse entries: 84547/10075613
## Sparsity           : 99%
## Maximal term length: 18
## Weighting          : term frequency (tf)
```

Note that `RTextTools` gives you plenty of options in preprocessing. Apart from the options used above, you can also strip whitespace, remove punctuation, and remove stopwords from lists that are already defined in the package. Stemming and stopword removal is language specific, so when you select the language in the option as above (`language=''english''`), the stemming and stopword removal will be done according to the language of your choice. At the moment, the stopwords included are those for Danish, Dutch, English, Finnish, French,

German, Italian, Norwegian, Portuguese, Russian, Spanish, and Swedish.

We then create a container parsing the document matrix into a training set, and a test set. The training set will be used to train the algorithm and the test set to test how well this algorithm was trained. The following command instructs R to use the first 4000 sentences for the training set the remaining 449 sentences for the test set. Moreover, we specify to append to the document matrix the variable that contains the assigned coders:

```
container <- create_container(doc_matrix, labels, trainSize = 1:10000, testSize = 10001:14640, vi
```

We can then train a model using one of the available algorithms. For instance, we can use the Support Vector Machines algorithm (SVM) as follows:

```
SVM <- train_model(container, "SVM")
```

Other algorithms available are glmnet (GLMNET), maximum entropy (MAXENT), scaled linear discriminant analysis (SLDA), bagging (BAGGING), boosting (BOOSTING), random forest (RF), neural networks (NNET), classification tree (TREE).

We then use the model we just trained to classify the texts in the test set. The following command instructs R to classify the documents in the test set of the container using the SVM model that we previously trained.

```
SVM_CLASSIFY <- classify_model(container, SVM)
```

We can also view the classification that was performed by the SVM model as follows. The first columns corresponds to the label that was assigned to each of the tweets in the training set, while the second column gives the probability that the sentence was assigned to that particular category by the SVM algorithm. As you can see, while the probability for some sentences is quite highfor others is quite low even though the classification always chooses the category with the highest probability.

```
head(SVM_CLASSIFY)
```

The next step is to check the performance of the model we just tested in terms of classification. To do this, we first request a function which returns a container with different summaries. For instance, we can request summaries on the basis of the labels that were attached to the sentences, the documents (or in this case, the sentences) by label, or on the basis of the algorithm.

```
analytics <- create_analytics(container, SVM_CLASSIFY)
summary(analytics)
```

```
## ENSEMBLE SUMMARY
##
##        n-ENSEMBLE COVERAGE n-ENSEMBLE RECALL
## n >= 1                   1               0.8
```

```
##
##
## ALGORITHM PERFORMANCE
##
## SVM_PRECISION    SVM_RECALL    SVM_FSCORE
##     0.6800000     0.6733333     0.6733333
```

Precision gives the proportion of bills that were classified as belonging to a category and actually belong to this category (true positives) to all the bills that were classified in that category (irrespective of where they belong). Recall is the proportion of bills that were classified as belonging to a category and actually belong to this category (true positives) to all the bills that belong to this category (true positives plus false negatives). The F score is a weighted average between precision and recall ranging fro 0 to 1.

Finally, we can compare the scores between the labels given by the coders and those based on our SVM:

```
compare <- as.data.frame(cbind(labels[10001:14640], SVM_CLASSIFY$SVM_LABEL))
table(compare)
```

```
##     V2
## V1     -1    0    1
##   -1 3013  296  110
##   0   289  347   55
##   1   131   59  340
```

8.1.2 SVM with Quanteda

Instead of using a separate package, we can also use `quanteda` to carry out an SVM. For this, we load some movie reviews, select 1000 of them at random, and place them into our corpus:

```
set.seed(42)

library(quanteda)
library(quanteda.classifiers)
corpus_reviews <- corpus_sample(data_corpus_LMRD, 1000)
```

Our aim here will be to see how well the SVM algorithm can predict the rating of the reviews. To do this, we first have to create a new variable `prediction`. This variable contains the same scores as the original rating. Then, we remove 30% of the scores and replace them with NA. We do so by creating a `missing` variable what contains 30% 0s and 70% 1s. We then place the 0s with NAs. These NA scores are then the ones we want the algorithm to predict. Finally, we add the new variable to the corpus:

```
prediction <- corpus_reviews$rating
```

```
set.seed(42)

missing <- rbinom(1000, 1, 0.7)
prediction[missing == 0] <- NA

docvars(corpus_reviews, "prediction") <- prediction
```

We then transform the corpus into a data frame, and also remove stopwords, numbers and punctuation:

```
dfm_reviews <- dfm(corpus_reviews, remove = stopwords("english"), remove_punct = TRUE, remove_num
```

```
## Warning: 'dfm.corpus()' is deprecated. Use 'tokens()' first.

## Warning: '...' should not be used for tokens() arguments; use 'tokens()' first.

## Warning: 'remove' is deprecated; use dfm_remove() instead
```

Now we can run the SVM algorithm. To do so, we tell the model on which dfm we want to run our model, and which variable contains the scores to train the algorithm. Here, this is our `prediction` variable with the missing data:

```
library(quanteda.textmodels)
svm_reviews <- textmodel_svm(dfm_reviews, y = docvars(dfm_reviews, "prediction"))
svm_reviews
```

```
##
## Call:
## textmodel_svm.dfm(x = dfm_reviews, y = docvars(dfm_reviews, "prediction"))
##
## 707 training documents; 129,216 fitted features.
## Method: L2-regularized L2-loss support vector classification dual (L2R_L2LOSS_SVC_DUAL)
```

Here we see that the algorithm used 720 texts to train the model (the one with a score) and fitted 133,728 features. The latter refers to the total number of words in the training texts and not only the unique ones. Now we can use this model to predict the ratings we removed earlier:

```
svm_predict <- predict(svm_reviews)
```

While we can of course look at the resulting numbers, we can also place them in a two-way table with the actual rating, to see how well the algorithm did:

```
rating     <- corpus_reviews$rating
table_data <- as.data.frame(cbind(svm_predict, rating))
table(table_data$svm_predict,table_data$rating)
```

```
##
##        1   2   3   4   7   8   9  10
##   1  175  14  10  11   4   3   1   6
```

```
##   2   13  65   5   3   0   0   0   3
##   3    5   2  82   4   1   4   0   3
##   4    4   5   5  90   1   5   2   6
##   7    0   1   2   2  75   6   0   1
##   8    2   1   1   0   3  83   5   7
##   9    3   0   1   6   4  11  74   7
##   10   1   3   3   2   3  10  14 137
```

Here, the table shows the prediction of the algorithm from top to bottom and the original rating from left to right. What we want is that all cases are on the diagonal: in that case, the prediction is the same as the original rating. Here, this happens in the majority of cases. Also, only in a few cases is the algorithm far off.

8.2 Naive Bayes

For the NB example, we will use data from the Manifesto Project (Volkens et al. 2019), also known as the Comparative Manifesto Project (CMP), Manifesto Research Group (MRG), and MARPOR (Manifesto Research on Political Representation)). After you have signed up and downloaded the API key, load the package and set the key:

```
library(manifestoR)
mp_setapikey("manifesto_apikey.txt")
```

While we can download the whole dataset, as it is rather large, it makes more sense to only download download a part of it. Here, we take the manifestos for the United Kingdom in 2015. To tell R we want only these documents, we make a small dataframe listing the party and the year we want, and then place this into the `mp_corpus` command. Note that instead of the names of the parties, the Manifesto Project assigns unique codes to each party. To see which code belongs to which party, see: https://manifesto-project.wzb.eu/down/data/2019a/codebooks/parties_MPDataset_MPDS2019a.pdf. Also note that the date includes both the year and month of the election:

```
manifestos <- data.frame(party=c(51320, 51620, 51110, 51421, 51901, 51902, 51951), dat
manifesto_corpus <- mp_corpus(manifestos)
```

```
## Connecting to Manifesto Project DB API... corpus version: 2020-2
## Connecting to Manifesto Project DB API... corpus version: 2020-2
```

For now, we are only interested in the (quasi)-sentences the of the manifestos, the codes the coders gave them, and names of the parties. To make everything more clear, we will take these elements from the corpus, combine them into a new data-frame, and remove all the NA values. We do this because otherwise the data would also include the headers and titles of the document, which do not have any codes assigned to them:

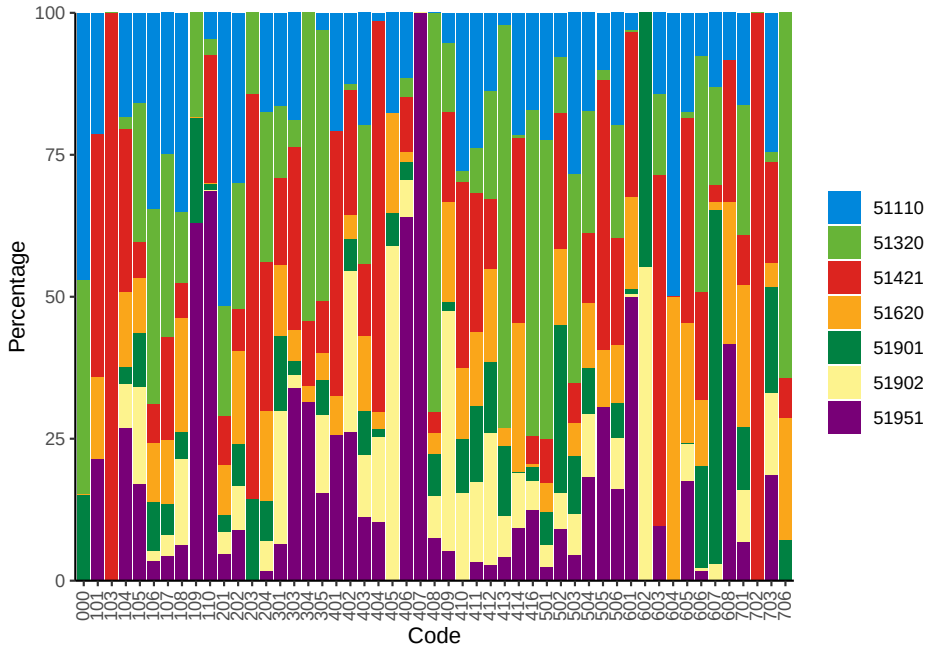
```
text_51320 <- manifesto_corpus$content[[1]]$content$text
text_51620 <- manifesto_corpus$content[[2]]$content$text
text_51110 <- manifesto_corpus$content[[3]]$content$text
text_51421 <- manifesto_corpus$content[[4]]$content$text
text_51901 <- manifesto_corpus$content[[5]]$content$text
text_51902 <- manifesto_corpus$content[[6]]$content$text
text_51951 <- manifesto_corpus$content[[7]]$content$text

texts <- c(text_51320,text_51620,text_51110,text_51421,text_51901,text_51902,text_51951)

party_51320 <- rep(51320,length.out=length(text_51320))
party_51620 <- rep(51620,length.out=length(text_51620))
party_51110 <- rep(51110,length.out=length(text_51110))
party_51421 <- rep(51421,length.out=length(text_51421))
party_51901 <- rep(51901,length.out=length(text_51901))
party_51902 <- rep(51902,length.out=length(text_51902))
party_51951 <- rep(51951,length.out=length(text_51951))

party <- c(party_51320,party_51620,party_51110,party_51421,party_51901,party_51902,party_51951)

cmp_code <- codes(manifesto_corpus)

manifesto_data <- data.frame(texts,cmp_code,party)
```

To get an idea of how much a party "owns" a code, we can calculate the row percentages. These inform us how much of the appearance of a certain code is due to a single party. To calculate these, we use the `prop.table` command. Here, the `,1` at the end tells R to look at the rows (no value would give the cell proportions, and 2 would give the column proportions). We then multiply the proportions by 100 to get the percentages. Then, we place the output in a data-frame, and provide some names to the columns using the `names` command:

```
prop_row <- as.data.frame((prop.table(table(manifesto_data$cmp_code,
    manifesto_data$party), 1) * 100))
names(prop_row) <- c("Code", "Party", "Percentage")
```

While we can look at the results by looking at the `prop_row` object, it is clearer to do this in a graph. To build this graph, in the command we first specify the data, the x variable (the codes), the y variable (the percentages), and the filling of the bar (which should be the party colours). These party colours we provide in the next line (in hexadecimal notation). Then we tell `ggplot` to draw the bar chart and *stack* the bars on top of each other (the alternative is to *dodge*, in which R places the bars next to each other). Then, we specify our theme, turn the text for the codes 90 degrees, and move the codes a little bit so they are under their respective bars:

```
library(ggplot2)

ggplot(data = prop_row, aes(x = Code, y = Percentage, fill = Party)) +
    scale_fill_manual("", values = c("#0087DC", "#67B437", "#DC241F",
        "#FAA61A", "#008142", "#FDF38E", "#780077")) + geom_bar(stat = "identity",
    position = "stack") + scale_y_continuous(expand = c(0, 0)) +
    theme_classic() + theme(axis.text.x = element_text(angle = 90)) +
    theme(axis.text.x = element_text(vjust = 0.4))
```



Now, we can see that some parties dominate some categories, while for others the spread is more even. For example, UKIP dominates the categories 406 and 407 - dealing with positive and negative mentions of protectionism, while the Conservatives do the same with category 103 (*Anti-Imperialism*). Note though, that these are percentages. This means that the reason the Conservatives dominate category 103 is as they have two (quasi)-sentences with that category. The others do not have the category at all (702 on *Negative Mentioning of Labour Groups* has the same issue). Other categories, such as 403 (*Market Regulation*) and 502 (*Positive Mentions of Culture*) are way better spread out over all the parties.

Another thing we can look at is what part of a party's manifesto belongs to any of the codes. This can help us answer the question: "what are the parties talking about?" To see this, we have to calculate the column percentages:

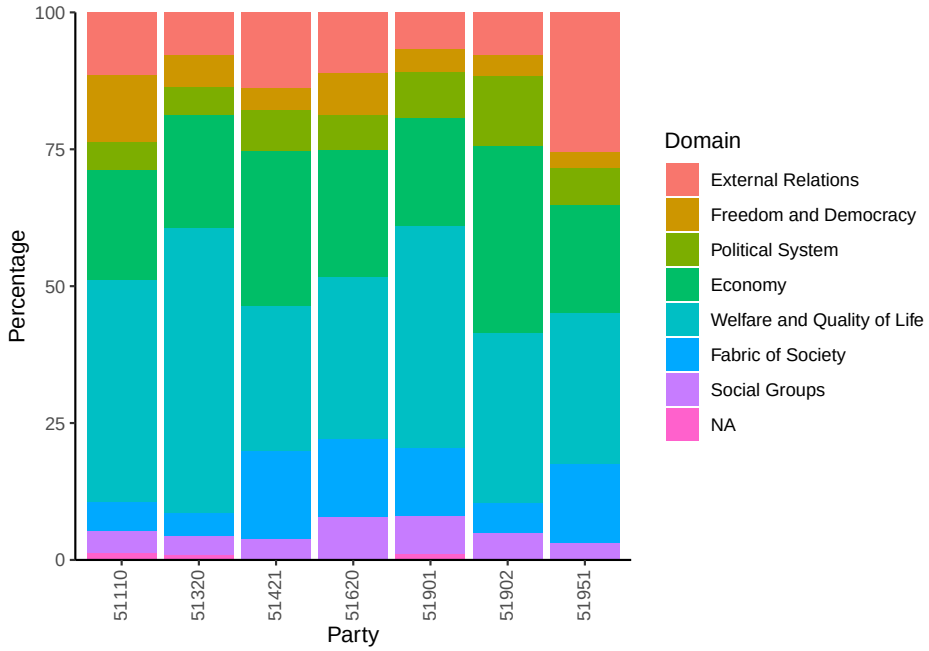
```
prop_col <- as.data.frame((prop.table(table(manifesto_data$cmp_code,
    manifesto_data$party), 2) * 100))
names(prop_col) <- c("Code", "Party", "Percentage")
```

If we now type `prop_col`, we can see what percentage of a party manifesto was about a certain code. Yet, given that there are 57 possible codes, it is more practical to cluster these in some way. Here, we do this using the Domains to which they belonged in the codebook. In total there are 7 domains (https://manifesto-project.wzb.eu/down/papers/handbook_2014_version_5.pdf), and a category which houses the 0 code. To cluster the codes, we make a new variable called `Domain`. To do so, we first transform the codes into numeric format, create an empty variable called `Domain`, and then replace the NA values in this empty category with the name of the domain based on the values in the Code variable. This we do using various operators R uses: `>=` means greater than and equal to, while `<=` means smaller than and equal to. Then, we make this new variable into a factor, and sort this factor in the way the codes occur:

```
prop_col$Code <- as.numeric(as.character(prop_col$Code))
prop_col$Domain <- NA

prop_col$Domain[prop_col$Code >= 101 & prop_col$Code <= 110] <- "External Relations"
prop_col$Domain[prop_col$Code >= 201 & prop_col$Code <= 204] <- "Freedom and Democracy"
prop_col$Domain[prop_col$Code >= 301 & prop_col$Code <= 305] <- "Political System"
prop_col$Domain[prop_col$Code >= 401 & prop_col$Code <= 416] <- "Economy"
prop_col$Domain[prop_col$Code >= 501 & prop_col$Code <= 507] <- "Welfare and Quality of Life"
prop_col$Domain[prop_col$Code >= 601 & prop_col$Code <= 608] <- "Fabric of Society"
prop_col$Domain[prop_col$Code >= 701 & prop_col$Code <= 706] <- "Social Groups"
prop_col$Domain[prop_col$Code == 0] <- "NA"

prop_col$Domain <- as.factor(prop_col$Domain)
prop_col$Domain <- factor(prop_col$Domain, levels(prop_col$Domain)[c(2,
    4, 6, 1, 8, 3, 7, 5)])
```

We then construct a plot as we did above:

```
ggplot(data = prop_col, aes(x = Party, y = Percentage, fill = Domain)) +
    geom_bar(stat = "identity", position = "stack") + scale_y_continuous(expand = c(0,
    0)) + theme_classic() + theme(axis.text.x = element_text(angle = 90)) +
    theme(axis.text.x = element_text(vjust = 0.4))
```



Here, we see that the Domain of *Welfare and Quality of Life* was the most dominant in all the manifestos, with *Economy* coming second. Also, especially UKIP paid a lot of attention to *External Relations*, while the Green party paid little attention to the *Fabric of Society*. In all, this gives us a good idea of what type of data we are actually dealing with.

Now let's get back to the classification. For this, we need to transform the corpus from the `manifestoR` package into a corpus for the `quanteda` package. To do so, we first have to transform the former into a data frame, and then turn it into a corpus. We then look at the first 10 entries:

```
corpus_data <- mp_corpus(manifestos) %>%
  as.data.frame(with.meta=TRUE)

manifesto_corpus <- corpus(corpus_data)

summary(manifesto_corpus, 10)
```

Here, we see that the corpus treats each sentence as a separate document (which is confusing). We can still identify to which party they belong due to the *party* variable, which shows the party code. The *cmp_code* variable shows the code assigned to the sentence (here it is all NA as the first sentences have the 0 category). To run the NB, instead of providing our training documents using a vector with NA values, we have to split our data-set into a training and a test set. For this, we first generate a string of 8000 random numbers between 0 and 10780 (the total number of sentences). We do so to prevent our training or test

set to exist only of sentences from a single party document:

```
set.seed(42)
id_train <- sample(1:10780, 8000, replace = FALSE)
head(id_train, 10)
```

```
##  [1]  2369  5273  9290  1252  8826 10289   356  7700  3954 10095
```

Then we generate a unique number for each of the 10780 sentences in our corpus. This so we can later match them to the sentences we would like to place in our training set or our test set:

```
docvars(manifesto_corpus, "id_numeric") <- 1:ndoc(manifesto_corpus)
```

We should now see this new variable *id_numeric* appear in our corpus. We can now construct our training and test set using these id's. For the training set, the logic is to create a sub set of the main corpus, and to take only those sentences whose *id_numeric* is also in *id_train*. For the test set, we do the same, only now taking only those sentences whose *id_numeric* is not in *id_train* (note that the ! mark signifies this). Then, we use the `%>%` pipe to transform the resulting object via a tokens object into a dfm:

```
manifesto_train <- corpus_subset(manifesto_corpus, id_numeric %in% id_train) %>%
  tokens() %>%
    dfm()

manifesto_test <- corpus_subset(manifesto_corpus, !id_numeric %in% id_train) %>%
    tokens() %>%
    dfm()
```

We then run the model using the `textmodel_nb` command, and ask it to use as classifiers the codes in the *cmp_code* variable:

```
manifesto_nb <- textmodel_nb(manifesto_train, docvars(manifesto_train, "cmp_code"))
summary(manifesto_nb)
```

Notice that the textmodel gives us a prediction of how likely it is that an individual word belongs to a certain code (the estimated feature scores). While this can be interesting, what we want to know here is how good the algorithm was. This is when we move from the training of the model using the training set to the prediction of the test set.

A problem is that Naive Bayes can only use featuers that were both in the training and the test set. To ensure this happens, we use the `dfm_match` option, which matches all the features in our dfm to a specified vector of features:

```
manifesto_matched <- dfm_match(manifesto_test, features = featnames(manifesto_train))
```

If we look at this new corpus we see that little has changed (there are still 2780 features). This means that all features that were in the test set were also

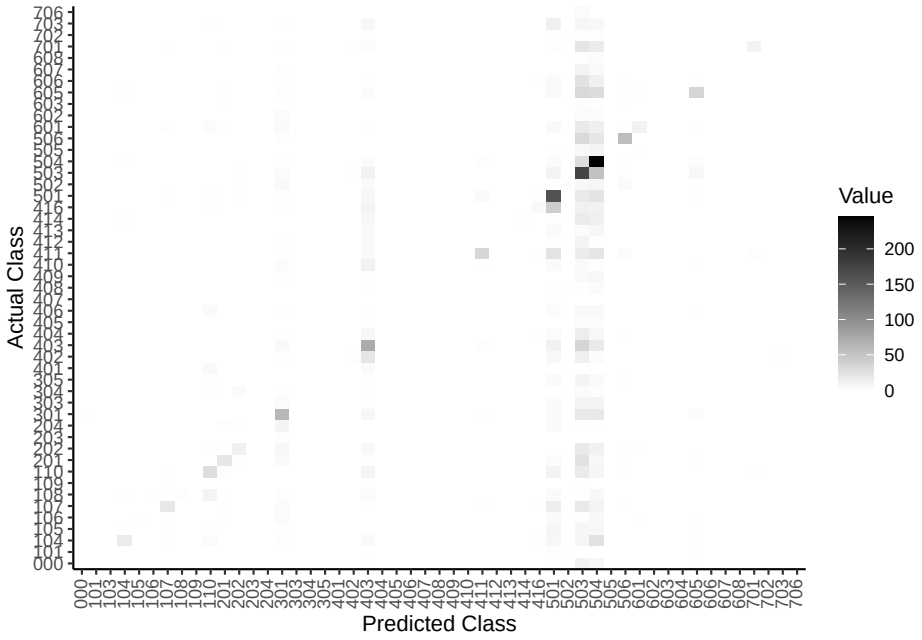
there in the training set. This is good news as this means the algorithm has all the information needed for a good prediction. Yet, the lower the number of sentences, the less likely this is to occur, so matching is always a good idea.

Now we can predict the missing codes in the test set (now the manifesto_matched dfm) using the model we trained earlier. The resulting classes are what the model predicts (we already set this when we trained the model). If we would then open the `predicted_class` object we can see to which code R assigned each sentence. Yet, as before, this is a little too much information. Moreover, we do not want to know what the model assigned the sentence to, but how this corresponds to the original code. To see this, we take the actual classes from the `manifesto_matched` dfm and place them with the predicted classes into a cross table:

```
predicted_class <- predict(manifesto_nb, newdata = manifesto_matched)
actual_class <- docvars(manifesto_matched, "cmp_code")
table_class <- table(actual_class, predicted_class)
table_class
```

While this is already better (we have to pay attention to the diagonal), the large number of codes still makes this hard to read. So, as before, we can better visualise these results - here with the help of a heatmap. To do this, we first tranform our table into a dataframe which gives us all the possible combinations of codes and their occurrence. We put this into the command and also use a scaling gradient that gets darker when the value in a cell is higher:

```
table_class <- as.data.frame(table_class)

ggplot(data = table_class, aes(x = predicted_class, y = actual_class)) +
    geom_tile(aes(fill = Freq)) + scale_fill_gradient(high = "black",
    low = "white", name = "Value") + xlab("Predicted Class") +
    ylab("Actual Class") + scale_y_discrete(expand = c(0, 0)) +
    theme_classic() + theme(axis.text.x = element_text(angle = 90)) +
    theme(axis.text.x = element_text(vjust = 0.4))
```



Here, we can see that a high number of cases are on the diagonal, which indicates that the algorithm did a good job. Yet, it also classified a large number of sentences into the 503 and 504 categories, while they belonged to any of the other categories.

Besides this, we can also summarize how good the algorithm is by means of Krippendorff's α. To do so, we take the predicted codes, transform them from factors to numeric values, and store them in an object. Then, we bind them together with the actual codes and place them into a data frame. Finally, we transpose the data frame (so that rows are now columns) and make it into a matrix:

```
predict <- as.numeric(as.character(predicted_class))
reliability <- as.data.frame(cbind(actual_class, predict))
reliability_t <- t(reliability)
reliability <- as.matrix(reliability_t)
```

Then, we load the `kripp.boot` package, and calculate the nominal version of Krippendorff's α, as we are working with nominal codes:

```
library(kripp.boot)
kripp.boot(reliability, iter = 500, method = "nominal")
```

Alternatively, we can use the `DescTools` package:

```
library(DescTools)
KrippAlpha(reliability, method = "nominal")
```

Here we see that the number of subjects was 2780 (the number of sentences in the test set), the number of coders 2 (the actual and the predicted codes), and the value of α 0.318 with an interval between 0.297 and 0.337. While this might not look particularly encouraging, when we realise that Mikhaylov, Laver, and Benoit (2012) estimate the agreement among trained coders by the Manifesto Project to be between 0.350 and 0.400, then 0.305 is quite a good score for a simple model!

Chapter 9

Unsupervised Methods

While supervised models often work fine for text classification, one disadvantage is that we need to set specifics for the model. As an alternative, we can not specify anything and have R find out which classifications work. There are various algorithms to do so, of which we here will focus on Latent Dirichlet Allocation (LDA), a "seeded" version of LDA that uses information from other sources to improve the results of the LDA, and finally we will look at a Structural Topic Model.

9.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation, or LDA, relies on the idea is that each text is in fact a mix of topics, and each word belongs to one these. To run LDA, we will use the `topicmodels` package, and use the inaugural speeches as an example. First, we will use the `convert` function to convert the data frequency matrix to a data term matrix as this is what `topicmodels` uses:

```
library(topicmodels)
inaugural_dtm <- convert(data_inaugural_dfm, to = "topicmodels")
```

Then, we fit an LDA model with 10 topics. First, we have to define some a priori parameters for the model. Here, we will use the Gibbs sampling method to fit the LDA model (Griffiths and Steyvers 2004) over the alternative VEM approach (Blei, Ng, and Jordan 2003). Gibbs sampling performs a random walk over the distribution so we need to set a seed to ensure reproducible results. In this particular example, we set five seeds for five independent runs. We also set a burn-in period of 2000 as the first iterations will not reflect the distribution well, and take the 200th iteration of the following 1000:

```
burnin <- 2000
iter <- 1000
```

```
thin <- 200
seed <- list(42, 5, 24, 158, 2500)
nstart <- 5
best <- TRUE
```

The LDA algorithm estimates topic-word probabilities as well as topic-document probabilities that we can extract and visualize. Here, we will start with the topic-word probabilities called *beta*. To do this, we will use the `tidytext` package which is part of the tidyverse family of packages. Central to the logic of tidyverse packages is that `tidytext` does not rely on a document term matrix but represents the data in a long format (Welbers, Van Atteveldt, and Benoit 2017, 252). Although this makes it less memory efficient, such data arrangement lends itself to effective visualisation. The whole logic of these packages is that it works with data which has columns (variables) and rows with single observations. While this is the logic most people know, but it is not always the quickest (and is also not used by `quanteda`). Yet, it always allows you to look at your data in a way most will understand. First, we run the LDA and have a look at the first 10 terms:

```
inaugural_lda10 <- LDA(inaugural_dtm, k = 10, method = "Gibbs",
    control = list(burnin = burnin, iter = iter, thin = thin,
        seed = seed, nstart = nstart, best = best))

terms(inaugural_lda10, 10)
```

```
##       Topic 1   Topic 2   Topic 3     Topic 4     Topic 5   Topic 6
##  [1,] "peace"   "us"      "business"  "americans" "every"   "never"
##  [2,] "world"   "new"     "may"       "citizens"  "great"   "must"
##  [3,] "nations" "people"  "congress"  "freedom"   "nation"  "republic"
##  [4,] "free"    "america" "policy"    "country"   "men"     "civilization"
##  [5,] "freedom" "must"    "states"    "president" "life"    "order"
##  [6,] "can"     "world"   "executive" "never"     "good"    "war"
##  [7,] "shall"   "can"     "made"      "common"    "part"    "concern"
##  [8,] "life"    "nation"  "necessary" "courage"   "upon"    "understanding"
##  [9,] "may"     "one"     "trade"     "day"       "action"  "tasks"
## [10,] "hope"    "time"    "hope"      "across"    "purpose" "production"
##       Topic 7      Topic 8       Topic 9      Topic 10
##  [1,] "change"     "united"      "government" "first"
##  [2,] "generation" "liberty"     "upon"       "need"
##  [3,] "journey"    "human"       "can"        "love"
##  [4,] "hands"      "democracy"   "people"     "days"
##  [5,] "weapons"    "believe"     "country"    "things"
##  [6,] "forth"      "states"      "progress"   "back"
##  [7,] "powerful"   "alone"       "must"       "hand"
##  [8,] "enduring"   "security"    "law"        "friends"
##  [9,] "greatness"  "millions"    "system"     "unity"
```

```
## [10,] "words"      "opportunity" "political"  "president"
```

Here, we can see that the first topic is most concerned with words referring to peace and freedom, the second with references to the people, the third with businesses, as so on. While we can interpret our topics this way, a better way might be to visualise the results. For this, we will use the `tidy` command to prepare the dataset for visualisation. Then, we tell the command to use the information from the *beta* column, which contains the probability of a word occurring in a certain topic:

```
library(tidytext)
library(dplyr)
library(ggplot2)

inaugural_lda10_topics <- tidy(inaugural_lda10, matrix = "beta")
```

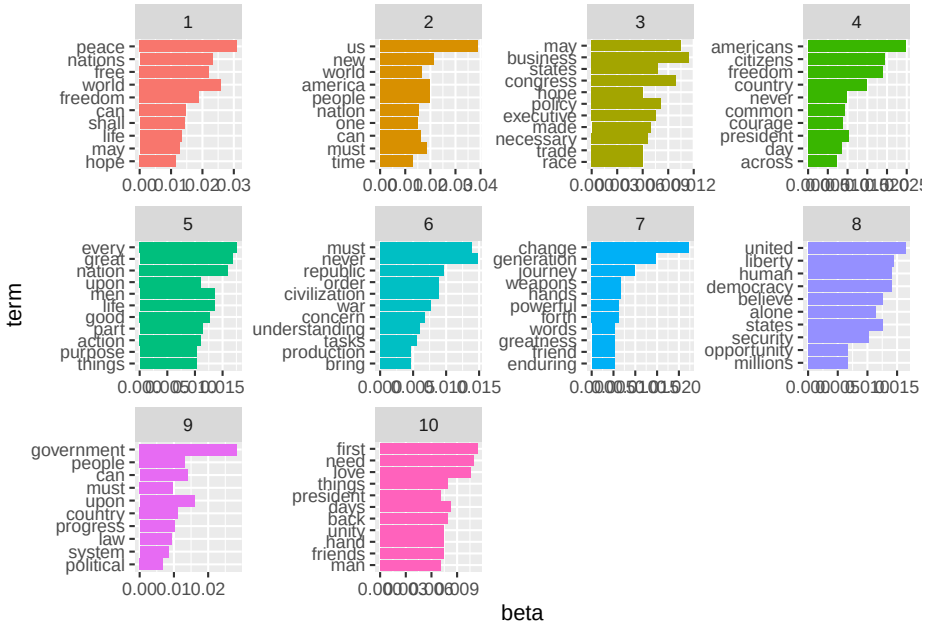
If we would look into the dataset now, we would see that it has 63130 observations with 3 variables. These are the number of the topic, the word (the term) and the **beta** - the chance that the word occurs in that topic. We now want to visualise only the top ten words for each topic in a bar-plot. Also, we want the graphs of each of these ten topics to appear in a single graph. To make this happen, we first have to select the top ten words for each topic. We do so gain using a pipe (which is the `%>%` command). This pipe transports an output of a command to another one before saving it. So here, we take our data-set and group it by topic using the `group_by` command. This command groups the dataset into 10 groups, each for every topic. What this allows us is to calculate things that we otherwise calculate for the whole data-set but here calculate for the groups instead. We then do so and select the top 10 terms (based on their beta value), using `top_n`. We then ungroup again (to make R view it as a single data-set), and use the `arrange` function to ensure the data-set has the topics sorted in an increasing fashion and the beta values in a decreasing fashion. Finally, we save this into a new object:

```
inaugural_lda10_topterms <- inaugural_lda10_topics %>%
    group_by(topic) %>%
    top_n(10, beta) %>%
    ungroup() %>%
    arrange(topic, -beta)
```

If we now look at the data-set, we see that it is much smaller and has the topics ordered. Yet, before we can plot this we have to ensure that (seen from top to bottom), all the beta for the first topic come first, then for the second topic, and so on. To do so, we use the `mutate` command, and redefine the term variable so that it is re-ordered based first on the term and then on the beta value. The result is a data frame with first the first topic, then the second topic etc., and with the beta values ordered within each topic. We then make the figure, with the terms on the horizontal axis and the beta values and the vertical axes, and

have the bars this generates coloured by topic. Also, we switch off the legend (which we do not need) and use the `facet_wrap` command to split up the total graph (which would have 107 bars otherwise - 107 bars and not a 100 because some terms had the same value for beta). We set the options for the scales to be **free** as it might be that the beta values for some topics are larger or smaller than for the others. Finally, we flip the graphs and make the x-axis the y-axis and vice versa, as this makes the picture more clear:

```
inaugural_lda10_topterms %>%
    mutate(term = reorder(term, beta)) %>%
    ggplot(aes(term, beta, fill = factor(topic))) + geom_col(show.legend = FALSE) +
    facet_wrap(~topic, scales = "free") + coord_flip()
```



What is clear here is that looking at only the words in each topic only says so much. In the first topic, the term "peace" is more important than anything else, and so is "us" in topic number 2. Also, in topic number ten, we see that both "first" and "need" are equally important.

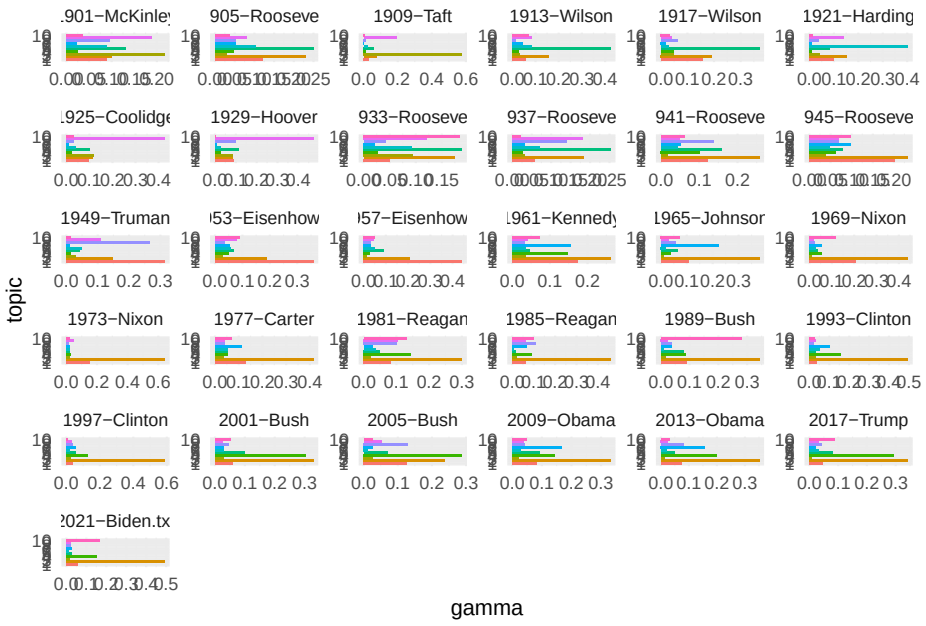
Another question we can ask is how much of each topic is in each of the documents. Put in another way: do certain documents talk more about certain topics than others? To see this, we first generate a new data frame with this information, known as the *gamma* value for each document:

```
inaugural_lda10_documents <- tidy(inaugural_lda10, matrix = "gamma")
```

We then go through similar steps to make the data-set ready for use and prepare the graph. For the graph, the only steps we do different are to force R to label each topic on the axis (as otherwise it will treat it as a continuous variable and

come up with useless values such as 7.5), and to give it a different look (using the `theme_classic()` command):

```
inaugural_lda10_toptopics <- inaugural_lda10_documents %>%
    group_by(document) %>%
    top_n(10, gamma) %>%
    ungroup() %>%
    arrange(topic, -gamma)

inaugural_lda10_toptopics %>%
    mutate(term = reorder(topic, gamma)) %>%
    ggplot(aes(topic, gamma, fill = factor(topic))) + geom_col(show.legend = FALSE) +
    scale_x_continuous(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9,
        10)) + facet_wrap(~document, scales = "free") + coord_flip() +
    theme_minimal()
```



Here, we see that in 1929 Hoover talked mostly about topic 9 (focusing on government), Biden in 2021 focused on words like "us" and "people," while in 1945 Roosevelt seemed to favour both the people and topics referring to peace. Again, our exact conclusions of course depend on how we interpret the topics.

9.2 Seeded Latent Dirichlet Allocation

An alternative to the above approach is one known as seeded-LDA. This approach uses seed words that can steer the LDA into the right direction. One origin of these seed words can be a dictionary that tells the algorithm which

words belong together in various categories. To use it, we will first load the packages and set a seed:

```
library(seededlda)
```

```
##
## Caricamento pacchetto: 'seededlda'
```

```
## I seguenti oggetti sono mascherati da 'package:topicmodels':
##
##     terms, topics
```

```
## Il seguente oggetto è mascherato da 'package:stats':
##
##     terms
```

```
library(quanteda.dictionaries)

set.seed(42)
```

Next, we need to specify a selection of seed words in a dictionary form. While we can construct a dictionary ourselves, here we choose to use the Laver and Garry dictionary we saw earlier. We then use this dictionary to run our seeded LDA:

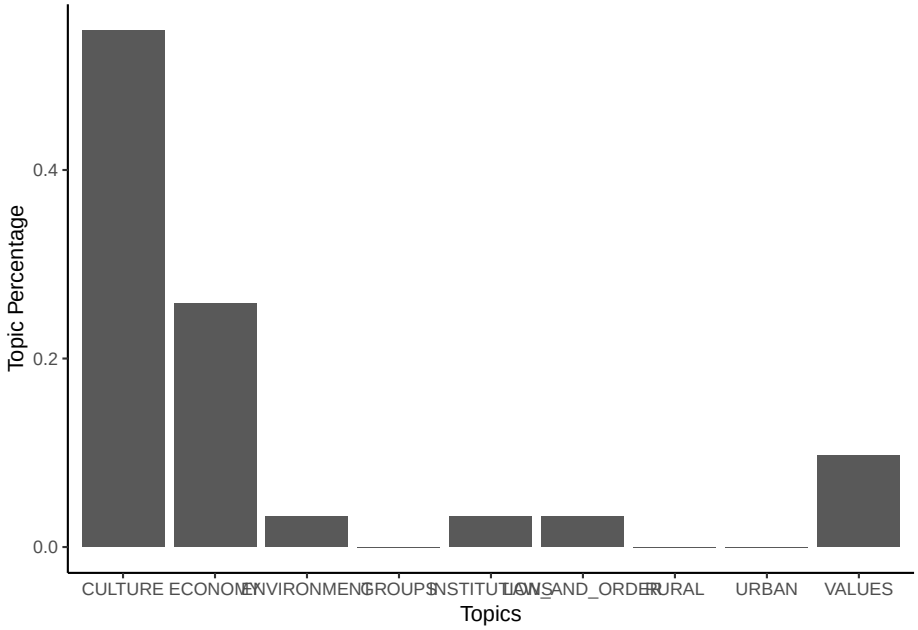
```
dict <- dictionary(data_dictionary_LaverGarry)
seededmodel <- textmodel_seededlda(data_inaugural_dfm, dictionary=dict)
terms(seededmodel, 20)
```

```
##       CULTURE     ECONOMY       ENVIRONMENT      GROUPS    INSTITUTIONS
##  [1,] "people"    "work"        "production"     "women"   "president"
##  [2,] "us"        "economic"    "productive"     "race"    "administration"
##  [3,] "art"       "opportunity" "planet"         "racial"  "continue"
##  [4,] "music"     "children"    "population"     "woman"   "office"
##  [5,] "operating" "economy"     "products"       "racism"  "executive"
##  [6,] "operation" "industrial"  "environment"    "ethnic"  "rule"
##  [7,] "new"       "trade"       "clean"          "racing"  "voices"
##  [8,] "america"   "confidence"  "productivity"   "day"     "legislation"
##  [9,] "nation"    "equal"       "produce"        "body"    "authority"
## [10,] "let"       "cost"        "product"        "built"   "democratic"
## [11,] "american"  "poverty"     "cleaner"        "fire"    "modern"
## [12,] "today"     "care"        "productions"    "task"    "agencies"
## [13,] "must"      "welfare"     "produced"       "evil"    "rules"
## [14,] "world"     "education"   "cleanse"        "mind"    "election"
## [15,] "one"       "commerce"    "productiveness" "spirit"  "sovereignty"
## [16,] "know"      "age"         "car"            "whether" "reforms"
## [17,] "americans" "health"      "chemical"       "put"     "voice"
## [18,] "now"       "private"     "warming"        "speaks"  "elected"
## [19,] "time"      "equality"    "depletion"      "serve"   "process"
```

```
## [20,] "can"       "jobs"        "republic"       "carried" "reform"
##       LAW_AND_ORDER   RURAL           URBAN      VALUES
##  [1,] "force"         "agriculture"  "town"     "history"
##  [2,] "forces"        "feed"         "towns"    "human"
##  [3,] "determined"    "agricultural" "story"    "rights"
##  [4,] "determination" "farm"         "thank"    "principles"
##  [5,] "conviction"    "farms"        "friends"  "past"
##  [6,] "court"         "forests"      "young"    "leadership"
##  [7,] "determine"     "farmers"      "hear"     "humanity"
##  [8,] "terror"        "villages"     "learned"  "maintain"
##  [9,] "forced"        "horseback"    "bless"    "preserve"
## [10,] "courts"        "farmer"       "prayer"   "defend"
## [11,] "dealing"       "countryside"  "must"     "leaders"
## [12,] "seize"         "village"      "together" "principle"
## [13,] "drugs"         "lanes"        "days"     "proud"
## [14,] "officers"      "landscape"    "lost"     "threat"
## [15,] "penalties"     "man"          "dreams"   "pride"
## [16,] "convictions"   "change"       "protect"  "heritage"
## [17,] "guarding"      "light"        "hearts"   "historic"
## [18,] "lawless"       "greatness"    "right"    "preserved"
## [19,] "illegal"       "hands"        "back"     "integrity"
## [20,] "victims"       "friends"      "personal" "preservation"
```

Note that using the dictionary has ensured that only the categories of the dictionary are used. We can therefore have a look which topics each inaugural speech and which terms were most likely for each of the topics. Let us start with the topics first:

```
topics <- topics(seededmodel)
topics_table <- ftable(topics)
topics_prop_table <- as.data.frame(prop.table(topics_table))

ggplot(data=topics_prop_table, aes(x=topics, y=Freq))+
  geom_bar(stat="identity")+
  labs(x="Topics", y="Topic Percentage")+
  theme(axis.text = element_text(size=10, angle=45,hjust = 1))+
  theme_classic()
```



Here, we find that Culture was the most favoured topic, followed by the Economy and Values. Finally, we can then have a look at the most likely terms for each topic, sorted by each of the categories in the dictionary:

```
terms <- terms(seededmodel)
terms_table <- ftable(terms)
terms_df <- as.data.frame(terms_table)
head(terms_df)
```

```
##   Var1    Var2      Freq
## 1    A CULTURE    people
## 2    B CULTURE        us
## 3    C CULTURE       art
## 4    D CULTURE     music
## 5    E CULTURE operating
## 6    F CULTURE operation
```

Here, we find that in the first cluster (denoted as "A"), the word "people" was most likely (from all words that belonged to Culture). Thus, within this cluster, talking about culture often contained references to the people, and we can make similar observations about the other categories.

9.3 Structural Topic Model

Besides LDA, various other methods for unsupervised classification exist, such as hierarchical clustering, k-means, and various other mixed membership models.

Each of them have their specific advantages and problems, and it often depends on the goal of the researcher to decide which method to use. One (relatively) new and flexible method is known as the Structural Topic Model or STM. In R, this method is implemented in the `stm` package (Roberts, Stewart, and Tingley 2019).

One of the outstanding features of stm is known as topical prevalence. This refers to the idea that we have the possibility to include any amount of supplemental information in the form of covariates that can help to identify the correct model or help to better understand the topics the model generates (Roberts et al. 2014). For example, information on time can be added to study how topics change over the years; actors on how they differ between different authors; and any other possible variable to see how they differ between them. One of the main advantages of STM is that, unlike in LDA, we are not required to set any parameters in advance. In LDA, these parameters - α (the degree of mixture of topics a document has) and β (the degree of mixture of words that a topic has) - have to be set beforehand based on previous knowledge. Yet, this knowledge is not always present and several iterations might be required before a correct number is settled upon. In STM, we use the metadata to set these parameters.

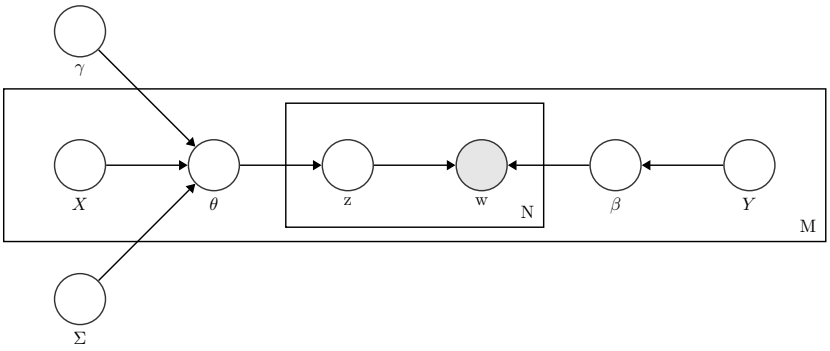


Figure 9.1: Plate diagram for a Structucal Topic Model.

In this plate diagram, X refers to the prevalence metadata; γ, the metadata weights; Σ, the topic covariances; θ, the document prevalence; z, the per-word topic; w, the observed word; Y, the content metadata; β, the topic content; N, the number of words in a document; and M, the number of documents in the corpus. 'So, to begin with, we load the package, set a seed, convert our dfm to the stm format and place our documents, vocabulary (the tokens) and any other data in three separate objects (for later convenience):

```
library(stm)
```

```
## stm v1.3.6 successfully loaded. See ?stm for help.
##  Papers, resources, and other materials at structuraltopicmodel.com
```

```
##
```

```
## Caricamento pacchetto: 'stm'

## Il seguente oggetto è mascherato da 'package:lattice':
##
##     cloud
```

```
library(quanteda)

set.seed(42)

data_inaugural_stm   <- convert(data_inaugural_dfm, to = "stm")

documents  <- data_inaugural_stm$documents
vocabulary <- data_inaugural_stm$vocab
meta       <- data_inaugural_stm$meta
```

The first thing we have to do is find the number of topics we need. In the `stm` package, we can do this by using a function called `searchK`. Here, we specify a range of values we think could include the "correct" number of topics, which are then run and collected. Subsequently, we can have a look at multiple goodness-of-fit measures to assess which number of topics (which k) has the best fit for the data. These measures include the exclusivity, semantic coherence, held-out likelihood, bound, lbound, and residual dispersion. Here, we run this for 2 to 15 possible topics.

In our code, we specify our documents, our tokens (the vocabulary), and our meta-data. Moreover, as our prevalence, we include parameters for Year and Party, as we expect the content of the topics to differ between both the Republican and Democratic party, as well as over time:

```
k <- c(3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15)
findingk <- searchK(documents, vocabulary, k, prevalence = ~Party +
    s(Year), data = meta, verbose = TRUE)
findingk_results <- as.data.frame(matrix(unlist(findingk$results),
    nrow = length(unlist(findingk$results[1]))))
names <- names(findingk$results)
names(findingk_results) <- names
```

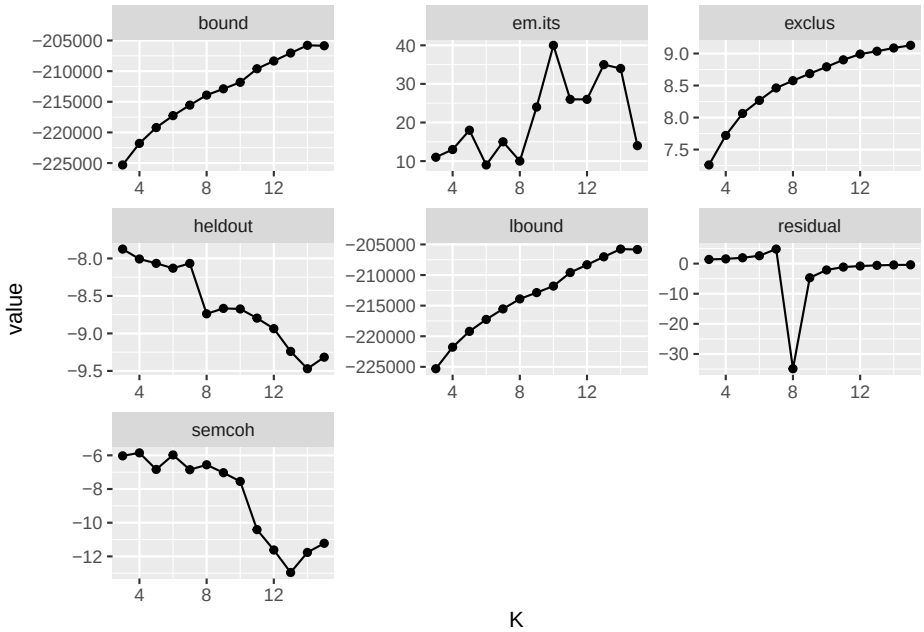
Looking at `findingk_results` we find various values. The first, exclusivity, refers to the occurrence when words have a high probability under one topic, they have a low probability under others. Related to this is semantic coherence which happens when the most probable words in a topic should occur in the same document. Held-out (or held-out log-likelihood) is the likelihood of our model on data that was not used in the initial estimation (the lower the better), residuals refers to the difference between a data point and the mean value that the model predicts for that data point (which we want to be 1, indicating a standard distribution). Finally, bound and lbound refer to a model's internal measure of fit. Here, we will be looking for the number of topics, that balance

the exclusivity and the semantic coherence, have a residual around 1, and a low held-out. To make this simpler, we visualise our data. In the first graph we plot all the values, while in the second, we only look at the exclusivity and the semantic coherence (as they are the most important):

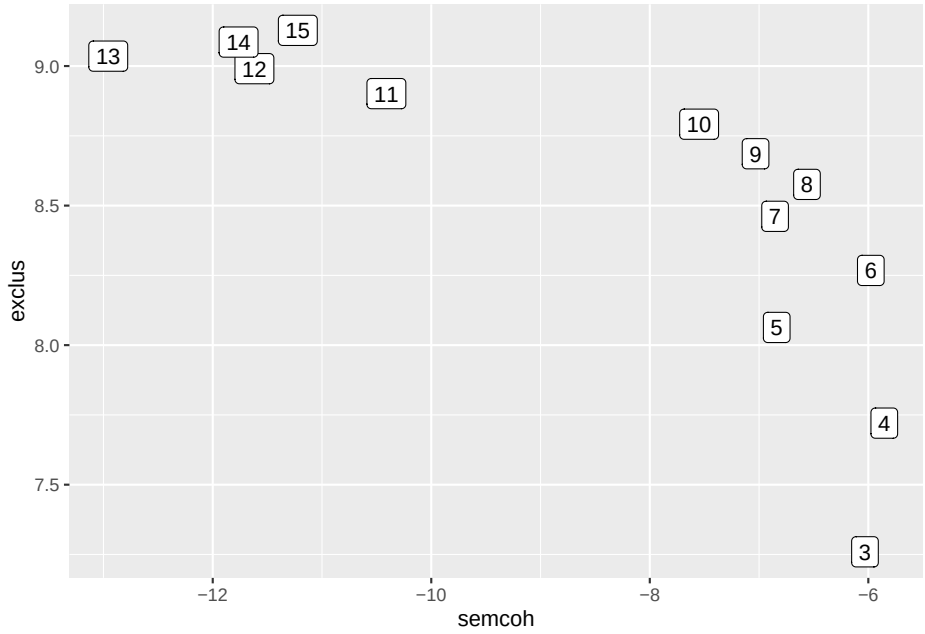
```
library(reshape2)
```

```
##
## Caricamento pacchetto: 'reshape2'

## Il seguente oggetto è mascherato da 'package:tidyr':
##
##     smiths
```

```
findingk_melt          <- melt(findingk_results, id="K")
findingk_melt$variable <- as.character(findingk_melt$variable)
findingk$K       <- as.factor(findingk_results$K)

ggplot(findingk_melt, aes(K, value)) +
  geom_point()+
  geom_line()+
  facet_wrap(~ variable, scales = "free")
```




```
ggplot(findingk_results, aes(semcoh, exclus)) +
  geom_point()+
  geom_label(data=findingk_results, label=findingk$K)
```



Based on these graphs, we decide upon 8 topics. The main reason for this is that for this number of topics, there is a relatively high semantic coherence given the exclusivity. We can now run our stm model, using spectral initialization and a topical prevalence including both the Party and the Year of the inauguration. Also, we have a look at the topics, and the words with the highest probability attached to them:

```
n_topics <- 8
output_stm <- stm(documents, vocabulary, K = n_topics, prevalence = ~Party +
    s(Year), data = meta, init.type = "Spectral", verbose = TRUE)
```

```
labelTopics(output_stm)
```

```
## Topic 1 Top Words:
##       Highest Prob: free, peace, world, shall, freedom, must, faith
##       FREX: strive, free, peoples, everywhere, truth, man's, learned
##       Lift: abhorring, absorbing, abstractions, acquire, aggressor, amass, andes
##       Score: anguished, productivity, strive, trial, learned, europe, defines
## Topic 2 Top Words:
##       Highest Prob: us, new, world, let, can, people, america
##       FREX: let, century, together, new, weapons, voices, abroad
##       Lift: 200th, 20th, dawn, explore, micah, moon, music
##       Score: attempting, nuclear, let, celebrate, voices, abroad, dawn
## Topic 3 Top Words:
##       Highest Prob: us, must, world, government, people, america, can
##       FREX: civilization, republic, experiment, normal, relationship, order, industi
```

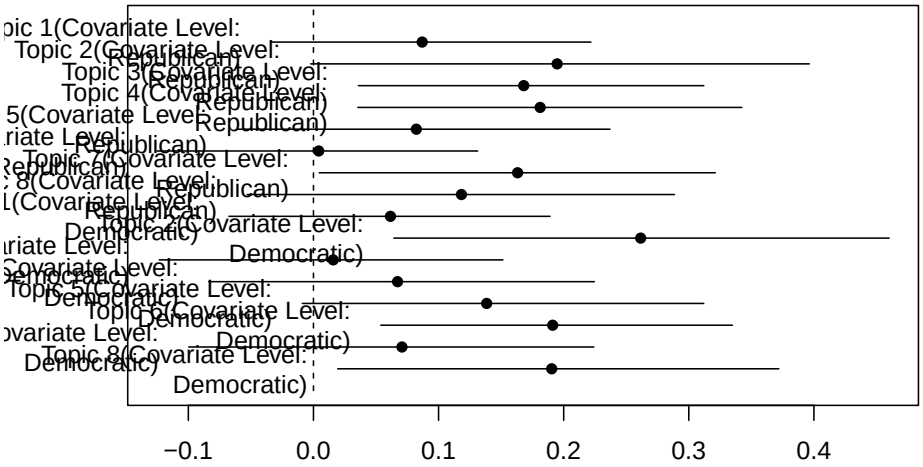
```
##       Lift: abnormal, acclaim, accompanied, accord, accumulation, acknowledgment, addressing
##       Score: accompanied, supreme, regards, deliberate, inspiration, unshaken, righteousness
## Topic 4 Top Words:
##       Highest Prob: us, america, nation, can, must, new, people
##       FREX: story, thank, president, defend, everyone, children, america
##       Lift: blowing, breeze, democracy's, january, obama, other's, page
##       Score: allowing, story, breeze, talk, crucial, everyone, virus
## Topic 5 Top Words:
##       Highest Prob: freedom, nation, people, america, government, know, democracy
##       FREX: speaks, mind, democracy, liberty, seen, came, millions
##       Lift: abreast, absence, admiration, agent, amount, aspires, attempts
##       Score: charta, speaks, paint, disaster, mind, defended, seen
## Topic 6 Top Words:
##       Highest Prob: us, must, nation, people, can, new, every
##       FREX: generation, journey, union, change, covenant, creed, enduring
##       Lift: demanded, mastery, span, storms, absolutism, abundantly, afghanistan
##       Score: abundantly, covenant, journey, mastery, storms, demanded, span
## Topic 7 Top Words:
##       Highest Prob: can, world, people, peace, nations, must, government
##       FREX: settlement, enforcement, countries, desire, party, international, property
##       Lift: aided, eighteenth, abilities, abound, abounding, absurd, acceptance
##       Score: abound, enforcement, contributed, settlement, property, major, eighteenth
## Topic 8 Top Words:
##       Highest Prob: upon, government, shall, can, must, great, may
##       FREX: army, interstate, negro, executive, tariff, business, proper
##       Lift: affected, amendments, antitrust, army, attention, avail, banking
##       Score: tariff, interstate, army, negro, policy, proper, business
```

Here, we see that the word "us" is relatively dominant in all our topics, making it a candidate for removal as a stop words in a future analysis. Looking at is more closely, we find that the first topic broadly refers to peace, the second, third and seventh to the world, the fourth and sixth to america, and the eighth to the government.

Finally, we can see whether there is any relation between these topics and any of the parameters we included. Here, let us look at any existing differences between the two parties:

```
est_assoc_effect <- estimateEffect(~Party, output_stm, metadata = meta,
    prior = 1e-05)
```

```
plot.estimateEffect(est_assoc_effect, "Party", method = "pointestimate",
    model = output_stm)
```



Here, we find that while the average for the topic do seem to differ a little between both of the parties, all the intervals are overlapping, thus indicating that they are not significantly different.

Albaugh, Quinn, Julie Sevenans, Stuart Soroka, and Peter John Loewen. 2013. "The Automated Coding of Policy Agendas: A Dictionary-Based Approach." In *6th Annual Comparative Agendas Conference, Antwerp, Belgium*.

Bakker, Ryan, Catherine de Vries, Erica Edwards, Liesbet Hooghe, Seth Jolly, Gary Marks, Jonathan Polk, Jan Rovny, Marco R. Steenbergen, and Milada Anna Vachudova. 2012. "Measuring Party Positions in Europe: The Chapel Hill Expert Survey Trend File, 1999-2010: The Chapel Hill Expert Survey Trend File, 1999–2010." *Party Politics* 21 (1): 1–15. https://doi.org/10.1177/1354068812462931.

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov. 2009. "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions." *American Journal of Political Science* 53 (2): 495–513. https://doi.org/10.1111/j.1540-5907.2009.00383.x.

Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "Quanteda: An r Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 3 (30): 774.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.

Clarke, Isobelle, and Jack Grieve. 2019. "Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted Between

2009 and 2018." *PLOS ONE* 14 (9): 1–27. https://doi.org/10.1371/journal.pone.0222062.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35 (4): 665–68.

Griffiths, Thomas L, and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences* 101 (suppl 1): 5228–35.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfals of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Haselmayer, Martin, and Marcelo Jenny. 2017. "Sentiment Analysis of Political Communication: Combining a Dictionary Approach with Crowdcoding." *Quality & Quantity* 51 (6): 2623–46.

Hayes, Andrew F., and Klaus Krippendorff. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures* 1 (1): 77–89. https://doi.org/10.1080/19312450709336664.

Krippendorff, Klaus. 2004. *Content Analysis - an Introduction to Its Methodology*. 2nd ed. Thousand Oaks, CA: SAGE Publications.

Lamprianou, Iasonas. 2020. "Measuring and Visualizing Coders' Reliability: New Approaches and Guidelines from Experimental Data." *Sociological Methods & Research*, 0049124120926198.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *The American Political Science Review* 97 (2): 311–31. https://doi.org/10.1017/S0003055403000698.

Laver, Michael, and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44 (3): 619–34. https://doi.org/10.2307/2669268.

Lin, L. 1989. "A Concordance Correlation Coefficient to Evaluate Reproducibility." *Biometrics* 45: 255–68.

Lind, Fabienne, Jakob-Moritz Eberl, Tobias Heidenreich, Hajo G Boomgaarden, Eva Luisa Gómez Montero, Beatriz Herrero, Rosa Berganza, Will Allen, and Peter Bajomi-Lazar. 2019. "Multilingual Dictionary Construction: A Roadmap to Measuring Migration Frames in European Media Discourse."

Lowe, Will. 2011. *JFreq: Count Words, Quickly*. http://www.conjugateprior.org/software/jfreq/.

Lowe, Will, and Kenneth Benoit. 2011. "Estimating Uncertainty in Quantitative Text Analysis." *Paper Prepared for Presentation at the Annual Conference of the Midwest Political Science Association*, 1–34.

Martin, Lanny W., and Georg Vanberg. 2008. "Reply to Benoit and Laver." *Political Analysis* 16 (1): 112–14. https://doi.org/10.1093/pan/mpm018.

Mikhaylov, Slava, Michael Laver, and Kenneth Benoit. 2012. "Coder Reliability and Misclassification in the Human Coding of Party Manifestos." *Political Analysis* 20 (1): 78–91. https://doi.org/10.1093/pan/mpr047.

Munzert, Simon, Christian Rubba, Peter Meißner, and Dominic Nyhuis. 2014. *Automated Data Collection with r: A Practical Guide to Web Scraping and Text Mining*. John Wiley & Sons.

Perriam, Jessamy, Andreas Birkbak, and Andy Freeman. 2020. "Digital Methods in a Post-API Environment." *International Journal of Social Research Methodology* 23 (3): 277–90.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. 2019. "stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91 (2). https://doi.org/10.18637/jss.v091.i02.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–82. https://doi.org/10.1111/ajps.12103.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52 (3): 705–22.

Volkens, Andrea, Werner Krause, Pola Lehmann, Theres Matthieß, Nicolas Merz, Sven Regel, and Bernhard Weßels. 2019. "The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR)." Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). 2019. https://doi.org/10.25522/manifesto.mpds.2019b.

Welbers, Kasper, Wouter Van Atteveldt, and Kenneth Benoit. 2017. "Text Analysis in r." *Communication Methods and Measures* 11 (4): 245–65.

Young, Lori, and Stuart Soroka. 2012. "Lexicoder Sentiment Dictionary."