Text Analysis with R

Claudia Engel, Scott Bailey

Last updated: May 12, 2022

Contents

Prerequisites

- You should have a **basic knowledge** of R, and be familiar with the topics covered in the Introduction to R.
- It is also recommended you have a **recent** version of R and RStudio installed.
- Packages needed:
  - `tidyverse`
  - `tidytext`
  - `readtext`
  - `sotu`
  - `SnowballC`
  - `widyr`
  - `igraph`
  - `ggraph`
  - `tm`

Make sure that you not only install, but also load the packages, to confirm the respective versions get along with your R version.

References

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text Mining Infrastructure in R. Journal of Statistical Software, 25(5), 1 - 54. doi: dx.doi.org/10.18637/jss.v025.i05

Gries, Stefan Thomas, 2009: Quantitative Corpus Linguistics with R: A Practical Introduction. Routledge.

Silge, J and D. Robinson, 2017: Text Mining with R: A Tidy Approach

Niekler, A. and G. Wiedemann 2020: Text mining in R for the social sciences and digital humanities

Kasper Welbers, Wouter Van Atteveldt & Kenneth Benoit (2017) Text Analysis in R. Communication Methods and Measures, 11:4, 245-265 doi: 10.1080/19312458.2017.1387238

Scott Chamberlain (2019). fulltext: Full Text of 'Scholarly' Articles Across Many Data Sources

CRAN Task View: Natural Language Processing

Chapter 1

Preparing Textual Data

Learning Objectives

- read textual data into R using `readtext`
- use the `stringr` package to prepare strings for processing
- use `tidytext` functions to tokenize texts and remove stopwords
- use `SnowballC` to stem words

We'll use several R packages in this section:

- `sotu` will provide the metadata and text of State of the Union speeches ranging from George Washington to Barack Obama.
- `tidyverse` is a collection of R packages designed for data science, including `dplyr` with a set of verbs for common data manipulations and `ggplot2` for visualization.
- `tidytext` provides specific functions for a "tidy" approach to working with textual data, where one row represents one "token" or meaningful unit of text, for example a word.
- `readtext` provides a function well suited to reading textual data from a large number of formats into R, including metadata.

```
library(sotu)
library(tidyverse)
library(tidytext)
library(readtext)
```

1.1 Reading text into R

First, let's look at the data in the `sotu` package. The metadata and texts are contained in this package separately in `sotu_meta` and `sotu_text` respectively. We can take a look at those by either typing the names or use funnctions like `glimpse()` or `str()`. Below, or example is what the metadata look like. Can you tell how many speeches there are?

```
# Let's take a look at the state of the union metadata
str(sotu_meta)
```

```
#> tibble [236 x 5] (S3: tbl_df/tbl/data.frame)
#>  $ president   : chr [1:236] "George Washington" "George Washington" "George Washington" "George W
#>  $ year        : int [1:236] 1790 1790 1791 1792 1793 1794 1795 1796 1797 1798 ...
#>  $ years_active: chr [1:236] "1789-1793" "1789-1793" "1789-1793" "1789-1793" ...
#>  $ party       : chr [1:236] "Nonpartisan" "Nonpartisan" "Nonpartisan" "Nonpartisan" ...
#>  $ sotu_type   : chr [1:236] "speech" "speech" "speech" "speech" ...
```

In order to work with the speech texts and to later practice reading text files from disk we use the function `sotu_dir()` to write the texts out. This function by default writes to a temporary directory with one speech in each file. It returns a character vector where each element is the name of the path to the individual speech file. We save this vector into the `file_paths` variable.

```
# sotu_dir writes the text files to disk in a temporary dir,
# but you could also specify a location.
file_paths <- sotu_dir()
head(file_paths)
```

```
#> [1] "/var/folders/5y/9x92pjcx2xd2h7qxqx39vpmc0000gn/T//Rtmpi4gRYm/file84cc50999363/george-washington-1790a.tx
#> [2] "/var/folders/5y/9x92pjcx2xd2h7qxqx39vpmc0000gn/T//Rtmpi4gRYm/file84cc50999363/george-washington-1790b.tx
#> [3] "/var/folders/5y/9x92pjcx2xd2h7qxqx39vpmc0000gn/T//Rtmpi4gRYm/file84cc50999363/george-washington-1791.txt
#> [4] "/var/folders/5y/9x92pjcx2xd2h7qxqx39vpmc0000gn/T//Rtmpi4gRYm/file84cc50999363/george-washington-1792.txt
#> [5] "/var/folders/5y/9x92pjcx2xd2h7qxqx39vpmc0000gn/T//Rtmpi4gRYm/file84cc50999363/george-washington-1793.txt
#> [6] "/var/folders/5y/9x92pjcx2xd2h7qxqx39vpmc0000gn/T//Rtmpi4gRYm/file84cc50999363/george-washington-1794.txt
```

Now that we have the files on disk and a vector of filepaths, we can pass this vector directly into `readtext` to read the texts into a new variable.

```
# let's read in the files with readtext
sotu_texts <- readtext(file_paths)
```

`readtext()` generated a dataframe for us with 2 colums: the doc_id, which is the name of the document and the actual text:

```
glimpse(sotu_texts)
```

```
#> Rows: 236
#> Columns: 2
#> $ doc_id <chr> "abraham-lincoln-1861.txt", "abraham-lincoln-1862.txt", "abraha~
#> $ text   <chr> "\n\n Fellow-Citizens of the Senate and House of Representative~
```

To work with a single table, we combine the text and metadata. Our `sotu_texts` are organized by alphabetical order, so we sort our metadata in `sotu_meta` to match that order and then bind the columns.

```
sotu_whole <-
  sotu_meta %>%
  arrange(president) %>% # sort metadata
  bind_cols(sotu_texts) # combine with texts

glimpse(sotu_whole)
```

```
#> Rows: 236
#> Columns: 7
#> $ president    <chr> "Abraham Lincoln", "Abraham Lincoln", "Abraham Lincoln", ~
#> $ year         <int> 1861, 1862, 1863, 1864, 1829, 1830, 1831, 1832, 1833, 183~
#> $ years_active <chr> "1861-1865", "1861-1865", "1861-1865", "1861-1865", "1829~
#> $ party        <chr> "Republican", "Republican", "Republican", "Republican", "~
#> $ sotu_type    <chr> "written", "written", "written", "written", "written", "w~
#> $ doc_id       <chr> "abraham-lincoln-1861.txt", "abraham-lincoln-1862.txt", "~
#> $ text         <chr> "\n\n Fellow-Citizens of the Senate and House of Represen~
```

Now that we have our data combined, we can start looking at the text. Typically quite a bit of effort goes into pre-processing the text for further analysis. Depending on the quality of your data and your goal, you might for example need to:

- replace certain characters or words,
- remove urls or certain numbers, such as phone numbers,
- clean up misspellings or errors,
- etc.

There are several ways to handle this sort of cleaning, we'll show a few examples below.

1.2 String operations

R has many functions available to manipulate strings including functions like `grep` and `paste`, which come with the R base install.

Here we will here take a look at the `stringr` package, which is part of the `tidyverse`. It refers to a lot of functionality from the `stringi` package which is perhaps one of the most comprehensive string manipulation packages.

Below are examples for a few functions that might be useful.

1.2.1 Counting ocurrences

`str_count` takes a character vector as input and by default counts the number of pattern matches in a string.

How man times does the word "citizen" appear in each of the speeches?

```
sotu_whole %>%
    mutate(n_citizen = str_count(text, "citizen"))
```

```
#> # A tibble: 236 x 8
#>    president        year years_active party     sotu_type doc_id text  n_citizen
#>    <chr>           <int> <chr>        <chr>     <chr>     <chr>  <chr>     <int>
#>  1 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ "\n\~         9
#>  2 Abraham Lincoln  1862 1861-1865    Republic~ written   abrah~ "\n\~         7
#>  3 Abraham Lincoln  1863 1861-1865    Republic~ written   abrah~ "\n\~        15
#>  4 Abraham Lincoln  1864 1861-1865    Republic~ written   abrah~ "\n\~         3
#>  5 Andrew Jackson   1829 1829-1833    Democrat~ written   andre~ "\n\~        19
#>  6 Andrew Jackson   1830 1829-1833    Democrat~ written   andre~ "\n\~        14
#>  7 Andrew Jackson   1831 1829-1833    Democrat~ written   andre~ "\n\~        23
#>  8 Andrew Jackson   1832 1829-1833    Democrat~ written   andre~ "\n\~        19
#>  9 Andrew Jackson   1833 1833-1837    Democrat~ written   andre~ "\n\~        14
#> 10 Andrew Jackson   1834 1833-1837    Democrat~ written   andre~ "\n\~        25
#> # ... with 226 more rows
```

It is possible to use regular expressions, for example, this is how we would check how many times either "citizen" or "Citizen" appear in each of the speeches:

```
sotu_whole %>%
    mutate(n_citizen = str_count(text, "citizen"),
           n_cCitizen = str_count(text, "[C|c]itizen"))
```

```
#> # A tibble: 236 x 9
#>    president        year years_active party     sotu_type doc_id text  n_citizen
#>    <chr>           <int> <chr>        <chr>     <chr>     <chr>  <chr>     <int>
#>  1 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ "\n\~         9
#>  2 Abraham Lincoln  1862 1861-1865    Republic~ written   abrah~ "\n\~         7
#>  3 Abraham Lincoln  1863 1861-1865    Republic~ written   abrah~ "\n\~        15
#>  4 Abraham Lincoln  1864 1861-1865    Republic~ written   abrah~ "\n\~         3
#>  5 Andrew Jackson   1829 1829-1833    Democrat~ written   andre~ "\n\~        19
#>  6 Andrew Jackson   1830 1829-1833    Democrat~ written   andre~ "\n\~        14
#>  7 Andrew Jackson   1831 1829-1833    Democrat~ written   andre~ "\n\~        23
#>  8 Andrew Jackson   1832 1829-1833    Democrat~ written   andre~ "\n\~        19
#>  9 Andrew Jackson   1833 1833-1837    Democrat~ written   andre~ "\n\~        14
#> 10 Andrew Jackson   1834 1833-1837    Democrat~ written   andre~ "\n\~        25
#> # ... with 226 more rows, and 1 more variable: n_cCitizen <int>
```

Going into the use of regular expressions is beyond this introduction. However we want to point out the `str_view()` function which can help you to create your correct expression. Also see RegExr, an online tool to learn, build, &

test regular expressions.

When used with the `boundary` argument `str_count()` can count different entities like "character", "line_break", "sentence", or "word". Here we add a new column to the dataframe indicating how many words are in each speech:

```
sotu_whole %>%
  mutate(n_words = str_count(text, boundary("word")))
```

```
#> # A tibble: 236 x 8
#>    president        year years_active party      sotu_type doc_id  text  n_words
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>   <chr>   <int>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraha~ "\n\~    6998
#>  2 Abraham Lincoln  1862 1861-1865    Republican written   abraha~ "\n\~    8410
#>  3 Abraham Lincoln  1863 1861-1865    Republican written   abraha~ "\n\~    6132
#>  4 Abraham Lincoln  1864 1861-1865    Republican written   abraha~ "\n\~    5975
#>  5 Andrew Jackson   1829 1829-1833    Democratic written   andrew~ "\n\~   10547
#>  6 Andrew Jackson   1830 1829-1833    Democratic written   andrew~ "\n\~   15109
#>  7 Andrew Jackson   1831 1829-1833    Democratic written   andrew~ "\n\~    7198
#>  8 Andrew Jackson   1832 1829-1833    Democratic written   andrew~ "\n\~    7887
#>  9 Andrew Jackson   1833 1833-1837    Democratic written   andrew~ "\n\~    7912
#> 10 Andrew Jackson   1834 1833-1837    Democratic written   andrew~ "\n\~   13472
#> # ... with 226 more rows
```

> CHALLENGE: Use the code above and add another column `n_sentences` where you calculate the number of sentences per speech. Then create a third column `avg_word_per_sentence`, where you calculate the number of words per sentence for each speech. Finally use `filter` to find which speech has shortest/longest average sentences length and what is the average length.

1.2.2 Detecting patterns

`str_detect` also looks for patterns, but instead of counts it returns a logical vector (TRUE/FALSE) indiciating if the pattern is or is not found. So we typically want to use it with the `filter` "verb" from `dplyr`.

What are the names of the documents where the words "citizen" and "Citizen" do **not** occur?

```
sotu_whole %>%
  filter(!str_detect(text, "[C|c]itizen")) %>%
  select(doc_id)
```

```
#> # A tibble: 11 x 1
#>    doc_id
#>    <chr>
#>  1 dwight-d-eisenhower-1958.txt
#>  2 gerald-r-ford-1975.txt
#>  3 richard-m-nixon-1970.txt
#>  4 richard-m-nixon-1971.txt
#>  5 richard-m-nixon-1972a.txt
#>  6 ronald-reagan-1988.txt
#>  7 woodrow-wilson-1916.txt
#>  8 woodrow-wilson-1917.txt
#>  9 woodrow-wilson-1918.txt
#> 10 woodrow-wilson-1919.txt
#> 11 woodrow-wilson-1920.txt
```

1.2.3 Extracting words

The `word` function extracts specific words from a character vector of words. By default it returns the first word. If for example we wanted to extract the first 5 words of each speech by Woodrow Wilson we provide the `end` argument like this:

```
sotu_whole %>%
  filter(president == "Woodrow Wilson") %>%  # sample a few speeches as demo
  pull(text) %>% # we pull out the text vector only
  word(end = 5)
```

```
#> [1] "\n\nGentlemen of the Congress:\n\nIn pursuance"
#> [2] "\n\nGENTLEMEN OF THE CONGRESS: \n\nThe"
#> [3] "GENTLEMEN OF THE CONGRESS: \n\nSince"
#> [4] "\n\nGENTLEMEN OF THE CONGRESS: \n\nIn"
#> [5] "Gentlemen of the Congress:\n\nEight months"
#> [6] "\n\nGENTLEMEN OF THE CONGRESS: \n\nThe"
#> [7] "\n\nTO THE SENATE AND HOUSE"
#> [8] "\n\nGENTLEMEN OF THE CONGRESS:\n\nWhen I"
```

1.2.4 Replacing and removing characters

Now let's take a look at text 'cleaninng'. We will first remove the newline characters (`\n`). We use the `str_replace_all` function to replace all the ocurrences of the `\n` pattern with a white space `" "`. We need to add the escape character `\` in front of our pattern to be replaced so the backslash before the `n` is interpreted correctly.

```
sotu_whole %>%
  filter(president == "Woodrow Wilson") %>%
  pull(text) %>%
  str_replace_all("\\n", " ") %>% # replace newline
  word(end = 5)
```

```
#> [1] "  Gentlemen of the"          "  GENTLEMEN OF THE"
#> [3] "GENTLEMEN OF THE CONGRESS: " "  GENTLEMEN OF THE"
#> [5] "Gentlemen of the Congress: " "  GENTLEMEN OF THE"
#> [7] "  TO THE SENATE"             "  GENTLEMEN OF THE"
```

This looks better, but we still have a problem to extract exactly 5 words because the too whitespaces are counted as a word. So let's get rid of any whitespaces before and after, as well as repeated whitespaces within the string with the `str_squish()` function.

```
sotu_whole %>%
  filter(president == "Woodrow Wilson") %>%
  pull(text) %>%
  str_replace_all("\\n", " ") %>%
  str_squish() %>%  # remove whitespaces
  word(end = 5)
```

```
#> [1] "Gentlemen of the Congress: In"    "GENTLEMEN OF THE CONGRESS: The"
#> [3] "GENTLEMEN OF THE CONGRESS: Since" "GENTLEMEN OF THE CONGRESS: In"
#> [5] "Gentlemen of the Congress: Eight" "GENTLEMEN OF THE CONGRESS: The"
#> [7] "TO THE SENATE AND HOUSE"          "GENTLEMEN OF THE CONGRESS: When"
```

(For spell checks take a look at https://CRAN.R-project.org/package=spelling or https://CRAN.R-project.org/package=hunspell)

1.3 Tokenize

A very common part of preparing your text for analysis involves *tokenization*. Currently our data contains in each each row a single text with metadata, so the entire speech text is the unit of observation. When we tokenize we break down the text into "tokens" (most commonly single words), so each row contains a single word with its metadata as unit of observation.

`tidytext` provides a function `unnest_tokens()` to convert our speech table into one that is tokenized. It takes three arguments:

- a tibble or data frame which contains the text;
- the name of the newly created column that will contain the tokens;
- the name of the column within the data frame which contains the text to be tokenized.

In the example below we name the new column to hold the tokens `word`. Remember that the column that holds the speech is called `text`.

```
tidy_sotu <- sotu_whole %>%
  unnest_tokens(word, text)

tidy_sotu
```

```
#> # A tibble: 1,965,212 x 7
#>    president        year years_active party      sotu_type doc_id          word
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>           <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ fell~
#>  2 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ citi~
#>  3 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ of
#>  4 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ the
#>  5 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ sena~
#>  6 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ and
#>  7 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ house
#>  8 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ of
#>  9 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ repr~
#> 10 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ in
#> # ... with 1,965,202 more rows
```

Note that the `unnest_tokens` function didn't just tokenize our texts at the word level. It also lowercased each word and stripped off the punctuation. We can tell it not to do this, by adding the following parameters:

```
# Word tokenization with punctuation and no lowercasing
sotu_whole %>%
  unnest_tokens(word, text, to_lower = FALSE, strip_punct = FALSE)
```

```
#> # A tibble: 2,157,777 x 7
#>    president        year years_active party      sotu_type doc_id          word
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>           <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ Fell~
#>  2 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ -
#>  3 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ Citi~
#>  4 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ of
#>  5 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ the
#>  6 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ Sena~
#>  7 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ and
#>  8 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ House
#>  9 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ of
#> 10 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ Repr~
#> # ... with 2,157,767 more rows
```

We can also tokenize the text at the level of ngrams or sentences, if those are the best units of analysis for our work.

```
# Sentence tokenization
sotu_whole %>%
  unnest_tokens(sentence, text, token = "sentences", to_lower = FALSE) %>%
  select(sentence)
```

```
#> # A tibble: 69,158 x 1
#>    sentence
#>    <chr>
#>  1 Fellow-Citizens of the Senate and House of Representatives:   In the midst o~
```

```
#>  2 You will not be surprised to learn that in the peculiar exigencies of the ti~
#>  3 A disloyal portion of the American people have during the whole year been en~
#>  4 A nation which endures factious domestic division is exposed to disrespect a~
#>  5 Nations thus tempted to interfere are not always able to resist the counsels~
#>  6 The disloyal citizens of the United States who have offered the ruin of our ~
#>  7 If it were just to suppose, as the insurgents have seemed to assume, that fo~
#>  8 If we could dare to believe that foreign nations are actuated by no higher p~
#>  9 The principal lever relied on by the insurgents for exciting foreign nations~
#> 10 Those nations, however, not improbably saw from the first that it was the Un~
#> # ... with 69,148 more rows
```

```
# N-gram tokenization as trigrams
sotu_whole %>%
  unnest_tokens(trigram, text, token = "ngrams", n = 3) %>%
  select(trigram)
```

```
#> # A tibble: 1,964,740 x 1
#>    trigram
#>    <chr>
#>  1 fellow citizens of
#>  2 citizens of the
#>  3 of the senate
#>  4 the senate and
#>  5 senate and house
#>  6 and house of
#>  7 house of representatives
#>  8 of representatives in
#>  9 representatives in the
#> 10 in the midst
#> # ... with 1,964,730 more rows
```

(Take note that the trigrams are generated by a moving 3-word window over the text.)

1.4 Stopwords

Another common task of preparing text for analysis is to remove stopwords. Stopwords are highly common words that are considered to provide non-relevant information about the content of a text.

Let's look at the stopwords that come with the `tidytext` package to get a sense of what they are.

```
stop_words
```

```
#> # A tibble: 1,149 x 2
#>    word        lexicon
#>    <chr>       <chr>
#>  1 a           SMART
#>  2 a's         SMART
#>  3 able        SMART
#>  4 about       SMART
#>  5 above       SMART
#>  6 according   SMART
#>  7 accordingly SMART
#>  8 across      SMART
#>  9 actually    SMART
#> 10 after       SMART
#> # ... with 1,139 more rows
```

These are English stopwords, pulled from different lexica ("onix", "SMART", or "snowball"). Depending on the type of analysis you're doing, you might leave these words in or alternatively use your own curated list of stopwords.

Stopword lists exist for many languages, see for examle the `stopwords` package in R. For now we will remove the English stopwords as suggested here.

For this we use `anti_join` from `dplyr`. We join and return all rows from our table of tokens `tidy_sotu` where there are **no matching values** in our list of stopwords. Both of these tables have one column name in common: `word` so by default the join will be on that column, and dplyr will tell us so.

```
tidy_sotu_words <- tidy_sotu %>%
  anti_join(stop_words)

tidy_sotu_words
```

```
#> # A tibble: 778,161 x 7
#>    president        year years_active party      sotu_type doc_id           word
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>            <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ fell~
#>  2 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ citi~
#>  3 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ sena~
#>  4 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ house
#>  5 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ repr~
#>  6 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ midst
#>  7 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ unpr~
#>  8 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ poli~
#>  9 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ trou~
#> 10 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ grat~
#> # ... with 778,151 more rows
```

If we compare this with `tidy_sotu` we see that the records with words like "of", "the", "and", "in" are now removed.

We also went from 1965212 to 778161 rows, which means we had a lot of stopwords in our corpus. This is a huge removal, so for serious analysis, we might want to scrutinize the stopword list carefully and determine if this is feasible.

1.5 Word Stemming

Another way you may want to clean your data is to stem your words, that is, to reduce them to their word stem or root form, for example reducing *fishing*, *fished*, and *fisher* to the stem *fish*.

`tidytext` does not implement its own word stemmer. Instead it relies on separate packages like `hunspell` or `SnowballC`.

We will give an example here for the `SnowballC` package which comes with a function `wordStem`. (`hunspell` appears to run much slower, and it also returns a list instead of a vector, so in this context `SnowballC` seems to be more convenient.)

```
library(SnowballC)
tidy_sotu_words %>%
        mutate(word_stem = wordStem(word))
```

```
#> # A tibble: 778,161 x 8
#>    president        year years_active party     sotu_type doc_id word  word_stem
#>    <chr>           <int> <chr>        <chr>     <chr>     <chr>  <chr> <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ fell~ fellow
#>  2 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ citi~ citizen
#>  3 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ sena~ senat
#>  4 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ house hous
#>  5 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ repr~ repres
#>  6 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ midst midst
#>  7 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ unpr~ unpreced
#>  8 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ poli~ polit
```

```
#>  9 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ trou~ troubl
#> 10 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ grat~ gratitud
#> # ... with 778,151 more rows
```

Lemmatization takes this another step further. While a stemmer operates on a single word without knowledge of the context, lemmatization attempts to discriminate between words which have different meanings depending on part of speech. For example, the word "better" has "good" as its lemma, something a stemmer would not detect.

For lemmatization in R, you may want to take a look a the `koRpus` package, another comprehensive R package for text analysis. It allows to use TreeTagger, a widely used part-of-speech tagger. For full functionality of the R package a local installation of TreeTagger is recommended.

Chapter 2

Analyzing Texts

Learning Objectives

- perform frequency counts and generate plots
- use the `widyr` package to calculate co-ocurrance
- use `igraph` and `ggraph` to plot a co-ocurrance graph
- import and export a Document-Term Matrix into `tidytext`
- use the `sentiments` dataset from `tidytext` to perform a sentiment analysis

---

Now that we've read in our text and metadata, tokenized and cleaned it a little, let's move on to some analysis.

First, we'll make sure we have loaded the libraries we'll need.

```
library(tidyverse)
library(tidytext)
```

Let's remind ourselves of what our data looks like.

```
tidy_sotu_words
```

```
#> # A tibble: 778,161 x 7
#>    president        year years_active party      sotu_type doc_id          word
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>           <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ fell~
#>  2 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ citi~
#>  3 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ sena~
#>  4 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ house
#>  5 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ repr~
#>  6 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ midst
#>  7 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ unpr~
#>  8 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ poli~
#>  9 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ trou~
#> 10 Abraham Lincoln  1861 1861-1865    Republican written   abraham-lincol~ grat~
#> # ... with 778,151 more rows
```

2.1 Frequencies

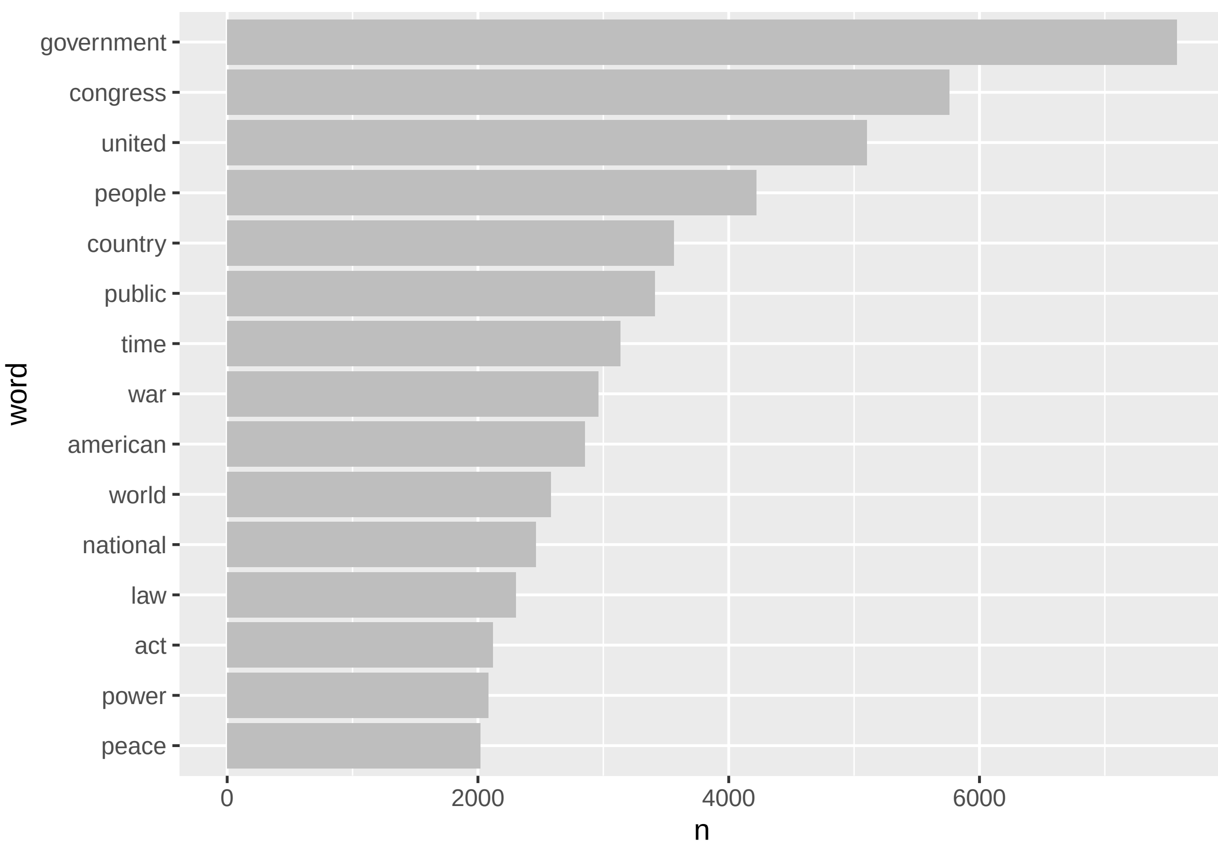
Since our unit of analysis at this point is a word, let's count to determine which words occur most frequently in the corpus as a whole.

```
tidy_sotu_words %>%
  count(word, sort = TRUE)
```

```
#> # A tibble: 29,558 x 2
#>    word           n
#>    <chr>      <int>
#>  1 government  7573
#>  2 congress    5759
#>  3 united      5102
#>  4 people      4219
#>  5 country     3564
#>  6 public      3413
#>  7 time        3138
#>  8 war         2961
#>  9 american    2853
#> 10 world       2581
#> # ... with 29,548 more rows
```

We can pipe this into `ggplot` to make a graph of the words that occur more that 2000 times. We count the words and use `geom_col` to represent the n values.

```
tidy_sotu_words %>%
  count(word) %>%
  filter(n > 2000) %>%
  mutate(word = reorder(word, n)) %>%  # reorder values by frequency
  ggplot(aes(word, n)) +
     geom_col(fill = "gray") +
     coord_flip()  # flip x and y coordinates so we can read the words better
```



Now let's look at a different question: In any given year, how often is the word 'peace' used and how often is the word 'war' used?

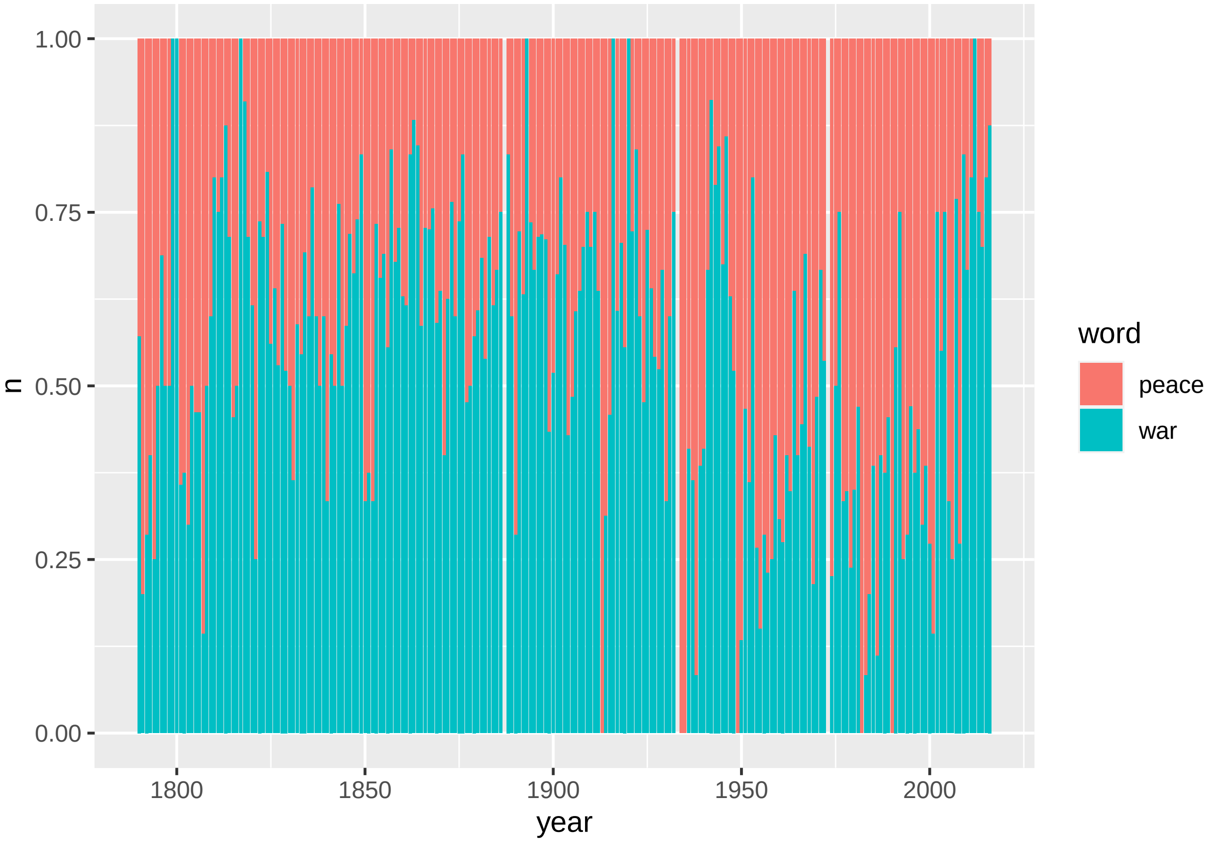
```
# steps:
# Select only the words 'war' and 'peace'.
```

```
# count ocurrences of each per year

tidy_sotu_words %>%
  filter(word %in% c("war", "peace")) %>%
  count(year, word)
```

```
#> # A tibble: 435 x 3
#>     year word      n
#>    <int> <chr> <int>
#>  1  1790 peace     3
#>  2  1790 war       4
#>  3  1791 peace     4
#>  4  1791 war       1
#>  5  1792 peace     5
#>  6  1792 war       2
#>  7  1793 peace     6
#>  8  1793 war       4
#>  9  1794 peace     3
#> 10  1794 war       1
#> # ... with 425 more rows
```

Now we can plot this as a bar chart that shows for each year the proportion of each of these two words out of the total of how often both words are used.

```
# plot n by year, and use position 'fill' to show the proportion

tidy_sotu_words %>%
  filter(word %in% c("war", "peace")) %>%
  count(year, word) %>%
  ggplot(aes(year, n, fill = word)) +
    geom_col(position = "fill")
```



As another example let us calculate the average number of words per speech for each president: How long was the average speech of each president and who are the most 'wordy' presidents?

First we summarize the words per president per speech:

```
tidy_sotu_words %>%
  count(president, doc_id)
```

```
#> # A tibble: 236 x 3
#>    president       doc_id                        n
#>    <chr>           <chr>                     <int>
#>  1 Abraham Lincoln abraham-lincoln-1861.txt   2578
#>  2 Abraham Lincoln abraham-lincoln-1862.txt   3088
#>  3 Abraham Lincoln abraham-lincoln-1863.txt   2398
#>  4 Abraham Lincoln abraham-lincoln-1864.txt   2398
#>  5 Andrew Jackson  andrew-jackson-1829.txt    3849
#>  6 Andrew Jackson  andrew-jackson-1830.txt    5428
#>  7 Andrew Jackson  andrew-jackson-1831.txt    2612
#>  8 Andrew Jackson  andrew-jackson-1832.txt    2881
#>  9 Andrew Jackson  andrew-jackson-1833.txt    2869
#> 10 Andrew Jackson  andrew-jackson-1834.txt    4952
#> # ... with 226 more rows
```

Then we use the output table and group it by president. That allows us to calculate the average number of words per speech.

```
tidy_sotu_words %>%
  count(president, doc_id)  %>%
  group_by(president) %>%
  summarize(avg_words = mean(n)) %>%
  arrange(desc(avg_words))
```

```
#> # A tibble: 41 x 2
#>    president           avg_words
#>    <chr>                   <dbl>
#>  1 William Howard Taft     9126.
#>  2 William McKinley        7797
#>  3 Jimmy Carter            7673.
#>  4 Theodore Roosevelt      7356
#>  5 James K. Polk           6920.
#>  6 Grover Cleveland        5736.
#>  7 James Buchanan          5409
#>  8 Benjamin Harrison       5308.
#>  9 Rutherford B. Hayes     4411
#> 10 Martin Van Buren        4286.
#> # ... with 31 more rows
```

2.2 Term frequency

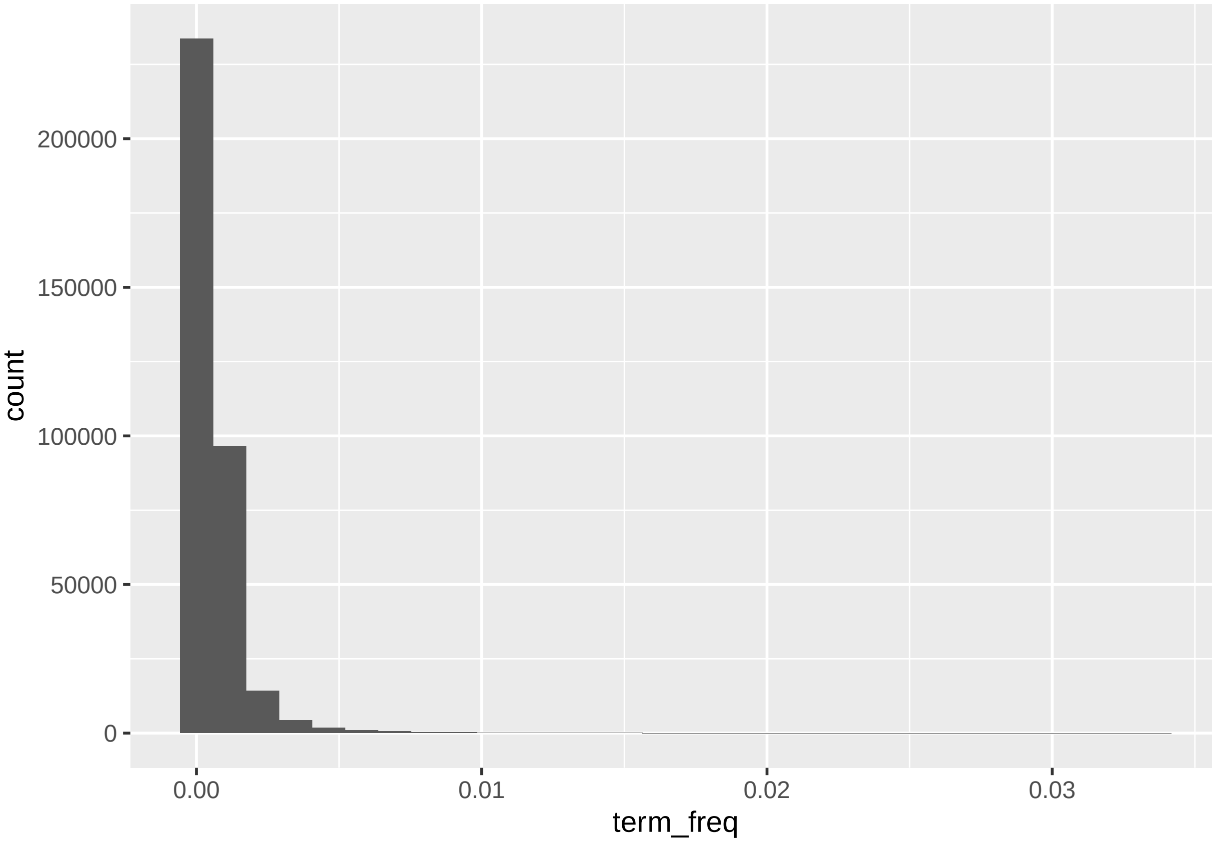
Often a raw count of a word is less important than understanding how often that word appears *relative to the total number* of words in a text. This ratio is called the **term frequency**. We can use `dplyr` to calculate it like this:

```
tidy_sotu_words %>%
  count(doc_id, word, sort = T)  %>%  # count occurrence of word and sort descending
  group_by(doc_id) %>%
  mutate(n_tot = sum(n),              # count total number of words per doc
         term_freq = n/n_tot)
```

```
#> # A tibble: 352,846 x 5
#> # Groups:   doc_id [236]
#>    doc_id                        word               n n_tot term_freq
#>    <chr>                         <chr>          <int> <int>     <dbl>
#>  1 harry-s-truman-1946.txt       dollars          207 12614   0.0164
#>  2 jimmy-carter-1980b.txt        congress         204 16128   0.0126
#>  3 harry-s-truman-1946.txt       war              201 12614   0.0159
#>  4 william-howard-taft-1910.txt government       164 11178   0.0147
#>  5 james-k-polk-1846.txt         mexico           158  7023   0.0225
#>  6 richard-m-nixon-1974b.txt     federal          141  9996   0.0141
#>  7 harry-s-truman-1946.txt       million          138 12614   0.0109
#>  8 harry-s-truman-1946.txt       fiscal           129 12614   0.0102
#>  9 jimmy-carter-1981.txt         administration   129 16595   0.00777
#> 10 william-howard-taft-1912.txt government       129 10215   0.0126
#> # ... with 352,836 more rows
```

Let's plot the distribution of the term frequency for the speeches:

```
tidy_sotu_words %>%
  count(doc_id, word)  %>%  # count n for each word
  group_by(doc_id) %>%
  mutate(n_tot = sum(n), # count total number of words per doc
         term_freq = n/n_tot) %>%
  ggplot(aes(term_freq)) +
    geom_histogram()
```



This distribution makes sense. Many words are used relatively rarely in a text. Only a few have a high term frequency.

Assuming that terms with high relative frequency are an indicator of significance we can find the term with the highest term frequency for each president:

```
tidy_sotu_words %>%
  count(president, word)  %>%  # count n for each word
  group_by(president) %>%
  mutate(n_tot = sum(n), # count total number of words per doc
         term_freq = n/n_tot) %>%
  arrange(desc(term_freq)) %>% # sort by term frequency
  top_n(1) %>%  # take the top for each president
  print(n = Inf) # print all rows
```

```
#> # A tibble: 43 x 5
#> # Groups:   president [41]
#>    president             word           n n_tot term_freq
#>    <chr>                 <chr>      <int> <int>     <dbl>
#>  1 John Adams            united        49  2768    0.0177
#>  2 John Tyler            government   209 12596    0.0166
#>  3 Martin Van Buren      government   256 17145    0.0149
#>  4 William J. Clinton    people       336 22713    0.0148
#>  5 Franklin D. Roosevelt war          283 19311    0.0147
#>  6 William McKinley      government   452 31188    0.0145
#>  7 Andrew Jackson        government   436 31031    0.0141
#>  8 Andrew Johnson        government   207 14968    0.0138
#>  9 George Washington     united        86  6226    0.0138
#> 10 Calvin Coolidge       government   274 20518    0.0134
#> 11 James K. Polk         mexico       360 27679    0.0130
#> 12 James Buchanan        government   279 21636    0.0129
```

```
#> 13 Zachary Taylor        congress      38  2948   0.0129
#> 14 Ulysses S. Grant      united       359 27933   0.0129
#> 15 William Howard Taft   government   461 36506   0.0126
#> 16 Grover Cleveland      government   574 45889   0.0125
#> 17 Franklin Pierce       united       200 16240   0.0123
#> 18 George Bush           world         82  6706   0.0122
#> 19 James Monroe          united       184 15157   0.0121
#> 20 George W. Bush        america      209 17265   0.0121
#> 21 Millard Fillmore      government   135 11986   0.0113
#> 22 John Quincy Adams     congress     131 11788   0.0111
#> 23 Harry S Truman        war          308 27819   0.0111
#> 24 Gerald R. Ford        federal       65  5879   0.0111
#> 25 Herbert Hoover        government   121 10947   0.0111
#> 26 Rutherford B. Hayes   congress     194 17644   0.0110
#> 27 Chester A. Arthur     government   185 16961   0.0109
#> 28 Lyndon B. Johnson     congress     115 11207   0.0103
#> 29 James Madison         war           85  8327   0.0102
#> 30 Barack Obama          america      204 20529   0.00994
#> 31 Benjamin Harrison     government   209 21230   0.00984
#> 32 Richard M. Nixon      federal      232 23701   0.00979
#> 33 Jimmy Carter          congress     518 53710   0.00964
#> 34 John F. Kennedy       world         68  7302   0.00931
#> 35 Theodore Roosevelt    government   528 58848   0.00897
#> 36 Ronald Reagan         government   133 15005   0.00886
#> 37 Ronald Reagan         people       133 15005   0.00886
#> 38 Woodrow Wilson        government   105 11982   0.00876
#> 39 Warren G. Harding     public        39  4583   0.00851
#> 40 Dwight D. Eisenhower  world        204 24410   0.00836
#> 41 Thomas Jefferson      country       58  7418   0.00782
#> 42 Abraham Lincoln       congress      81 10462   0.00774
#> 43 Abraham Lincoln       united        81 10462   0.00774
```

> CHALLENGE: Pick one president. For each of his speeches, which is the term with highest term frequency? Create a table as output. (Hint: `top_n`might be useful)

2.3 Tf-idf

So far we've been looking at term frequency per document. What if we want to know about words that seem more important based on the contents of the *entire* corpus?

For this, we can use **term-frequency according to inverse document frequency**, also callled **tf-idf**. Tf-idf measures how important a word is within a corpus by scaling term frequency per document according to the inverse of the term's document frequency (number of documents within the corpus in which the term appears divided by the number of documents).

The tf-idf value will be:

- lower for words that appear frequently in many documents of the corpus, and lowest when the word occurs in virtually all documents.
- higher for words that appear frequently in just a few documents of the corpus, this lending high discriminatory power to those few documents.

The intuition here is that if a term appears frequently in a document, we think that it is important but if that word appears in too many other documents, it is not that unique and thus perhaps not that important.

The `tidytext` package includes a function `bind_tf_idf`. It takes a table that contains one-row-per-term-per-document, the name of the column that contains the words (terms), the name of the column which contains the doc-id, and the name of the column that contains the document-term counts.

So below we aggregate our tibble with the word tokens to create the one-row-per-term-per-document table and then pipe it into the `bind_tf_idf` function.

```
tidy_sotu_words %>%
  count(doc_id, word, sort = TRUE)  %>%  # aggregate to count n for each word
  bind_tf_idf(word, doc_id, n)
```

```
#> # A tibble: 352,846 x 6
#>    doc_id                         word                n      tf     idf    tf_idf
#>    <chr>                          <chr>           <int>   <dbl>   <dbl>     <dbl>
#>  1 harry-s-truman-1946.txt        dollars           207 0.0164  0.612   0.0100
#>  2 jimmy-carter-1980b.txt         congress          204 0.0126  0.00425 0.0000537
#>  3 harry-s-truman-1946.txt        war               201 0.0159  0.0345  0.000550
#>  4 william-howard-taft-1910.txt   government        164 0.0147  0.00425 0.0000623
#>  5 james-k-polk-1846.txt          mexico            158 0.0225  0.810   0.0182
#>  6 richard-m-nixon-1974b.txt      federal           141 0.0141  0.293   0.00414
#>  7 harry-s-truman-1946.txt        million           138 0.0109  0.728   0.00796
#>  8 harry-s-truman-1946.txt        fiscal            129 0.0102  0.494   0.00505
#>  9 jimmy-carter-1981.txt          administration    129 0.00777 0.282   0.00219
#> 10 william-howard-taft-1912.txt   government        129 0.0126  0.00425 0.0000536
#> # ... with 352,836 more rows
```

Our function added three columns to the aggregated table which contain term frequency (`tf`), inverse document frequency (`idf`) and Tf-idf (`tf_idf`).

Let's look at some of the words in the corpus that have the highest tf-idf scores, which means words that are particularly distinctive for their documents.

```
tidy_sotu_words %>%
  count(doc_id, word, sort = TRUE)  %>%
  bind_tf_idf(word, doc_id, n) %>%
  arrange(desc(tf_idf))
```

```
#> # A tibble: 352,846 x 6
#>    doc_id                          word          n      tf   idf tf_idf
#>    <chr>                           <chr>     <int>   <dbl> <dbl>  <dbl>
#>  1 lyndon-b-johnson-1966.txt       vietnam      32 0.0152   2.42 0.0367
#>  2 jimmy-carter-1980a.txt          soviet       31 0.0218   1.47 0.0321
#>  3 george-w-bush-2003.txt          hussein      19 0.00811  3.85 0.0313
#>  4 george-w-bush-2003.txt          saddam       19 0.00811  3.67 0.0298
#>  5 franklin-d-roosevelt-1943.txt   1942         13 0.00758  3.85 0.0292
#>  6 dwight-d-eisenhower-1961.txt    1953         23 0.00747  3.85 0.0288
#>  7 john-adams-1800.txt             gentlemen     8 0.0153   1.80 0.0275
#>  8 benjamin-harrison-1892.txt      1892         40 0.00741  3.52 0.0261
#>  9 franklin-d-roosevelt-1942.txt   hitler        7 0.00527  4.77 0.0251
#> 10 herbert-hoover-1930.txt         1928         14 0.00711  3.52 0.0250
#> # ... with 352,836 more rows
```

To understand the occurrence of the years as being particularly distinctive we might need to look more closely at the speeches themselves, and determine whether the years are significant or whether they need to be removed from the text either permanently in the clean up or temporarily using `filter()`.

> CHALLENGE: Pick the same president you chose above. For each of his speeches, which is the term with highest tf-idf? Create a table as output. (Hint: Remember to group by doc_id before you use top_n)

2.4 N-Grams

We mentioned n-grams in the intro, but let's revisit them here and take a look at the most common bigrams in the speeches. Remember we can use the `unnest_token()` function on our texts and explicitly tell it to generate bigrams:

```
sotu_whole %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) # create bigram
```

```
#> # A tibble: 1,964,976 x 7
#>    president        year years_active party      sotu_type doc_id          bigram
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>           <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ fello~
#>  2 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ citiz~
#>  3 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ of the
#>  4 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ the s~
#>  5 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ senat~
#>  6 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ and h~
#>  7 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ house~
#>  8 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ of re~
#>  9 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ repre~
#> 10 Abraham Lincoln  1861 1861-1865    Republican written   abraham-linco~ in the
#> # ... with 1,964,966 more rows
```

Let's see the most common bigrams:

```
sotu_whole %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  count(bigram, sort = TRUE) # count occurrences and sort descending
```

```
#> # A tibble: 469,092 x 2
#>    bigram            n
#>    <chr>         <int>
#>  1 of the        33610
#>  2 in the        12499
#>  3 to the        11643
#>  4 for the        6892
#>  5 and the        6224
#>  6 by the         5606
#>  7 of our         5172
#>  8 the united     4767
#>  9 united states  4760
#> 10 it is          4756
#> # ... with 469,082 more rows
```

Ok, so we again need to remove the stopwords. First let us separate the two words into two columns "word1" and "word2" with `separate` from the `tidyr` package:

```
sotu_whole %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ")
```

```
#> # A tibble: 1,964,976 x 8
#>    president        year years_active party      sotu_type doc_id    word1 word2
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>     <chr> <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ fell~ citi~
#>  2 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ citi~ of
#>  3 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ of    the
#>  4 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ the   sena~
#>  5 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ sena~ and
```

```
#>  6 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ and   house
#>  7 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ house of
#>  8 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ of    repr~
#>  9 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ repr~ in
#> 10 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ in    the
#> # ... with 1,964,966 more rows
```

Now we use dplyr's `filter()` function to select only the words in each column that are not in the stopwords.

```
sotu_whole %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>% # separate into cols
  filter(!word1 %in% stop_words$word, # remove stopwords
         !word2 %in% stop_words$word)
```

```
#> # A tibble: 215,992 x 8
#>    president        year years_active party      sotu_type doc_id    word1 word2
#>    <chr>           <int> <chr>        <chr>      <chr>     <chr>     <chr> <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ fell~ citi~
#>  2 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ unpr~ poli~
#>  3 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ poli~ trou~
#>  4 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ abun~ harv~
#>  5 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ pecu~ exig~
#>  6 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ fore~ nati~
#>  7 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ prof~ soli~
#>  8 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ soli~ chie~
#>  9 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ dome~ affa~
#> 10 Abraham Lincoln  1861 1861-1865    Republican written   abraham-~ disl~ port~
#> # ... with 215,982 more rows
```

Lastly, we re-unite the two word columns into back into our bigrams and save it into a new table `sotu_bigrams`.

```
sotu_bigrams <- sotu_whole %>%
  unnest_tokens(bigram, text, token = "ngrams", n = 2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>% # separate into cols
  filter(!word1 %in% stop_words$word, # remove stopwords
         !word2 %in% stop_words$word) %>%
  unite(bigram, word1, word2, sep = " ")  # combine columns


sotu_bigrams %>%
  count(bigram, sort = TRUE)
```

```
#> # A tibble: 129,622 x 2
#>    bigram                n
#>    <chr>             <int>
#>  1 federal government   479
#>  2 american people      428
#>  3 june 30              325
#>  4 fellow citizens      296
#>  5 public debt          283
#>  6 public lands         256
#>  7 health care          240
#>  8 social security      232
#>  9 post office          202
#> 10 annual message       200
#> # ... with 129,612 more rows
```

A bigram can also be treated as a term in a document in the same way that we treated individual words. That means we can look at tf-idf values in the same way. For example, we can find out the most distinct bigrams that the presidents uttered in all their respective speeches taken together.

We count per president and bigram and then bind the tf-idf value with the `bind_tf_idf` function. In order to get the top bigram for each president we then group by president, and sort and retrieve the highest value for each.

```
sotu_bigrams %>%
  count(president, bigram) %>%
  bind_tf_idf(bigram, president, n) %>%
  group_by(president) %>%
  arrange(desc(tf_idf)) %>%
  top_n(1)
```

```
#> # A tibble: 44 x 6
#> # Groups:   president [41]
#>    president          bigram              n      tf   idf tf_idf
#>    <chr>              <chr>           <int>   <dbl> <dbl>  <dbl>
#>  1 George W. Bush     al qaida           35 0.00628 3.02  0.0190
#>  2 John Adams         john adams          3 0.00510 3.71  0.0189
#>  3 William J. Clinton 21st century       59 0.00830 1.77  0.0147
#>  4 Thomas Jefferson   gun boats           7 0.00462 3.02  0.0140
#>  5 Thomas Jefferson   port towns          7 0.00462 3.02  0.0140
#>  6 Thomas Jefferson   sea port            7 0.00462 3.02  0.0140
#>  7 Zachary Taylor     german empire       5 0.00789 1.63  0.0129
#>  8 Lyndon B. Johnson  south vietnam      13 0.00424 3.02  0.0128
#>  9 James Madison      james madison       8 0.00412 3.02  0.0124
#> 10 Harry S Truman     million dollars   119 0.0129  0.941 0.0121
#> # ... with 34 more rows
```

CHALLENGE: Again, pick the same president you chose above. For each of his speeches, which is the bigram with highest tf-idf? Create a table as output.

2.5 Co-occurrence

Co-occurrences give us a sense of words that appear in the same text, but not necessarily next to each other.

For this section we will make use of the `widyr` package. The function which helps us do this is the `pairwise_count()` function. It lets us count common pairs of words co-appearing within the same speech.

Behind the scenes, this function first turns our table into a wide matrix. In our case that matrix will be made up of the individual words and the cell values will be the counts of in how many speeches they co-occur, like this:

```
#>      we thus have
#> we   NA    4    5
#> thus  4   NA    2
#> have  5    2   NA
```

It then will turn the matrix back into a tidy form, where each row contains the word pairs and the count of their co-occurrence. Since we don't care about the order of the words, we will not count the upper triangle of the wide matrix, which leaves us with:

```
#>
#>    we thus 4
#>    we have 5
#>  thus have 2
```

Since processing the entire corpus would take too long here, we will only look at the last 100 words of each speech: which words occur most commonly together at the end of the speeches?

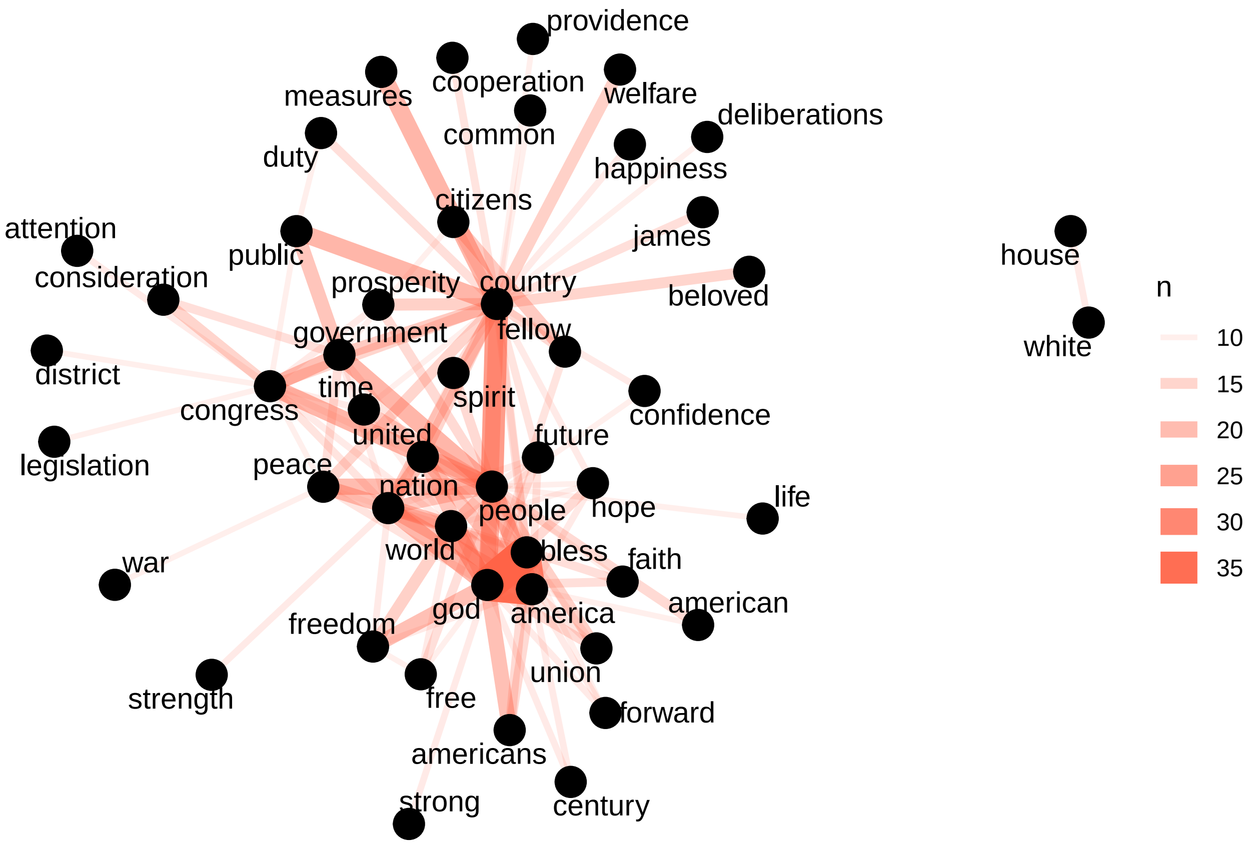
```
library(widyr)
```

```r
sotu_word_pairs <- sotu_whole %>%
  mutate(speech_end = word(text, -100, end = -1)) %>%  # extract last 100 words
  unnest_tokens(word, speech_end) %>%   # tokenize
  filter(!word %in% stop_words$word) %>%  # remove stopwords
  pairwise_count(word, doc_id, sort = TRUE, upper = FALSE) # don't include upper triangle of matrix

sotu_word_pairs
```

```
#> # A tibble: 125,576 x 3
#>    item1      item2       n
#>    <chr>      <chr>   <dbl>
#>  1 god        bless      37
#>  2 god        america    35
#>  3 bless      america    30
#>  4 people     country    26
#>  5 world      god        22
#>  6 god        people     22
#>  7 government people     21
#>  8 congress   people     21
#>  9 public     country    21
#> 10 god        nation     21
#> # ... with 125,566 more rows
```

To visualize the co-occurrence network of words that occur together at the end of 10 or more speeches, we use the **igraph** package to convert our table into a network graph and the **ggraph** package which adds functionality to ggplot to make it easier to plot a network.

```r
library(igraph)
library(ggraph)

sotu_word_pairs %>%
  filter(n >= 10) %>%  # only word pairs that occur 10 or more times
  graph_from_data_frame() %>% #convert to graph
  ggraph(layout = "fr") + # place nodes according to the force-directed algorithm of Fruchterman and Reingold
  geom_edge_link(aes(edge_alpha = n, edge_width = n), edge_colour = "tomato") +
  geom_node_point(size = 5) +
  geom_node_text(aes(label = name), repel = TRUE,
                 point.padding = unit(0.2, "lines")) +
  theme_void()
```



There are alternative approaches for this as well. See for example the `findAssocs` function in the `tm` package.

2.6 Document-Term Matrix

A document-term matrix (DTM) is a format which is frequently used in text analysis. It is a matrix where we can see the counts of each term per document. In a DTM each row represents a document, each column represents a term, and the cell values are the counts of the occurrences of the term for the particular document.

`tidytext` provides functionality to convert to and from DTMs, if for example, your analysis requires specific functions from a different R package which only works with DTM object types.

The `cast_dtm` function can be used to create a DTM object from a tidy table.

Let's assume that for some reason we want to use the `findAssoc()` function from the `tm` package.

First we use dplyr to create a table with the document name, the term, and the count.

```r
# make a table with document, term, count
tidy_sotu_words %>%
  count(doc_id, word)
```

```
#> # A tibble: 352,846 x 3
#>    doc_id                   word                n
#>    <chr>                    <chr>           <int>
#>  1 abraham-lincoln-1861.txt 1,470,018           1
#>  2 abraham-lincoln-1861.txt 1,500               1
#>  3 abraham-lincoln-1861.txt 100,000             1
#>  4 abraham-lincoln-1861.txt 102,532,509.27      1
#>  5 abraham-lincoln-1861.txt 12,528,000          1
#>  6 abraham-lincoln-1861.txt 13,606,759.11       1
#>  7 abraham-lincoln-1861.txt 1830                1
#>  8 abraham-lincoln-1861.txt 1859                1
```

```
#>  9 abraham-lincoln-1861.txt 1860                2
#> 10 abraham-lincoln-1861.txt 1861                6
#> # ... with 352,836 more rows
```

Now we cast it as a DTM.

```
sotu_dtm <- tidy_sotu_words %>%
  count(doc_id, word) %>%
  cast_dtm(doc_id, word, n)

class(sotu_dtm)
```

```
#> [1] "DocumentTermMatrix"    "simple_triplet_matrix"
```

Finally, let's use it in the `tm` package:

```
library(tm)

# look at the terms with tm function
Terms(sotu_dtm) %>% tail()
```

```
#> [1] "queretaro"    "refreshments" "schleswig"    "sedulous"     "subagents"
#> [6] "transcript"
```

```
# most frequent terms
findFreqTerms(sotu_dtm, lowfreq = 5000)
```

```
#> [1] "congress"   "government" "united"
```

```
# find terms associated with "citizen"
findAssocs(sotu_dtm, "citizen", corlimit = 0.5)
```

```
#> $citizen
#>        laws citizenship  protection   contained    entitled  government
#>        0.62        0.59        0.56        0.55        0.53        0.53
#>    citizens  postmaster     careful    question      report       suits
#>        0.52        0.52        0.51        0.51        0.51        0.51
```

Conversely, `tidytext` implements the `tidy` function (originally from the `broom` package) to import DocumentTermMatrix objects. Note that it only takes the cells from the DTM that are not 0, so there will be no rows with 0 counts.

2.7 Sentiment analysis

`tidytext` comes with a dataset `sentiments` which contains several sentiment lexicons, where each word is attributed a certain sentiment, like this:

```
sentiments
```

```
#> # A tibble: 6,786 x 2
#>    word        sentiment
#>    <chr>       <chr>
#>  1 2-faces     negative
#>  2 abnormal    negative
#>  3 abolish     negative
#>  4 abominable  negative
#>  5 abominably  negative
#>  6 abominate   negative
#>  7 abomination negative
#>  8 abort       negative
#>  9 aborted     negative
#> 10 aborts      negative
```

```
#> # ... with 6,776 more rows
```

Here we will take a look at how the sentiment of the speeches change over time. We will use the lexicon from Bing Liu and collaborators, which assigns positive/negative labels for each word:

```
bing_lex <- get_sentiments("bing")
bing_lex
```

```
#> # A tibble: 6,786 x 2
#>    word        sentiment
#>    <chr>       <chr>
#>  1 2-faces     negative
#>  2 abnormal    negative
#>  3 abolish     negative
#>  4 abominable  negative
#>  5 abominably  negative
#>  6 abominate   negative
#>  7 abomination negative
#>  8 abort       negative
#>  9 aborted     negative
#> 10 aborts      negative
#> # ... with 6,776 more rows
```

We can use these sentiments attached to each word and join them to the words of our speeches. We will use `inner_join` from `dplyr`. It will take all rows with words from `tidy_sotu_words` that match words in `bing_lex`, eliminating rows where the word cannot be found in the lexicon. Since our columns to join on have the same name (`word`) we don't need to explicitly name it.

```
sotu_sentiments <- tidy_sotu_words %>%
  inner_join(bing_lex)  # join to add semtinemt column

sotu_sentiments
```

```
#> # A tibble: 105,206 x 8
#>    president        year years_active party     sotu_type doc_id word  sentiment
#>    <chr>           <int> <chr>        <chr>     <chr>     <chr>  <chr> <chr>
#>  1 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ trou~ negative
#>  2 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ grat~ positive
#>  3 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ unus~ negative
#>  4 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ abun~ positive
#>  5 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ pecu~ negative
#>  6 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ prof~ positive
#>  7 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ soli~ negative
#>  8 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ disl~ negative
#>  9 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ dest~ negative
#> 10 Abraham Lincoln  1861 1861-1865    Republic~ written   abrah~ disr~ negative
#> # ... with 105,196 more rows
```

Finally we can visualize the proportion of positive sentiment (out of the total of positive and negative) in US State of the Union Addresses over time like this:

```
sotu_sentiments %>%
  count(year, sentiment) %>% # count by year and sentiment
  pivot_wider(names_from = "sentiment", values_from = "n") %>% # create column for positive
                                                               # and negative sentiment
  mutate(positive_ratio = positive/(negative + positive)) %>% # calculate positive ratio
  # plot
  ggplot(aes(year, positive_ratio)) +
    geom_line(color="gray") +
    geom_smooth(span = 0.3, se = FALSE) + # smooth for easier viewing
```

```
    geom_hline(yintercept = .5, linetype="dotted", color = "orange", size = 1) + # .5 as reference
    scale_x_continuous(breaks = seq(1790, 2016, by = 10)) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```