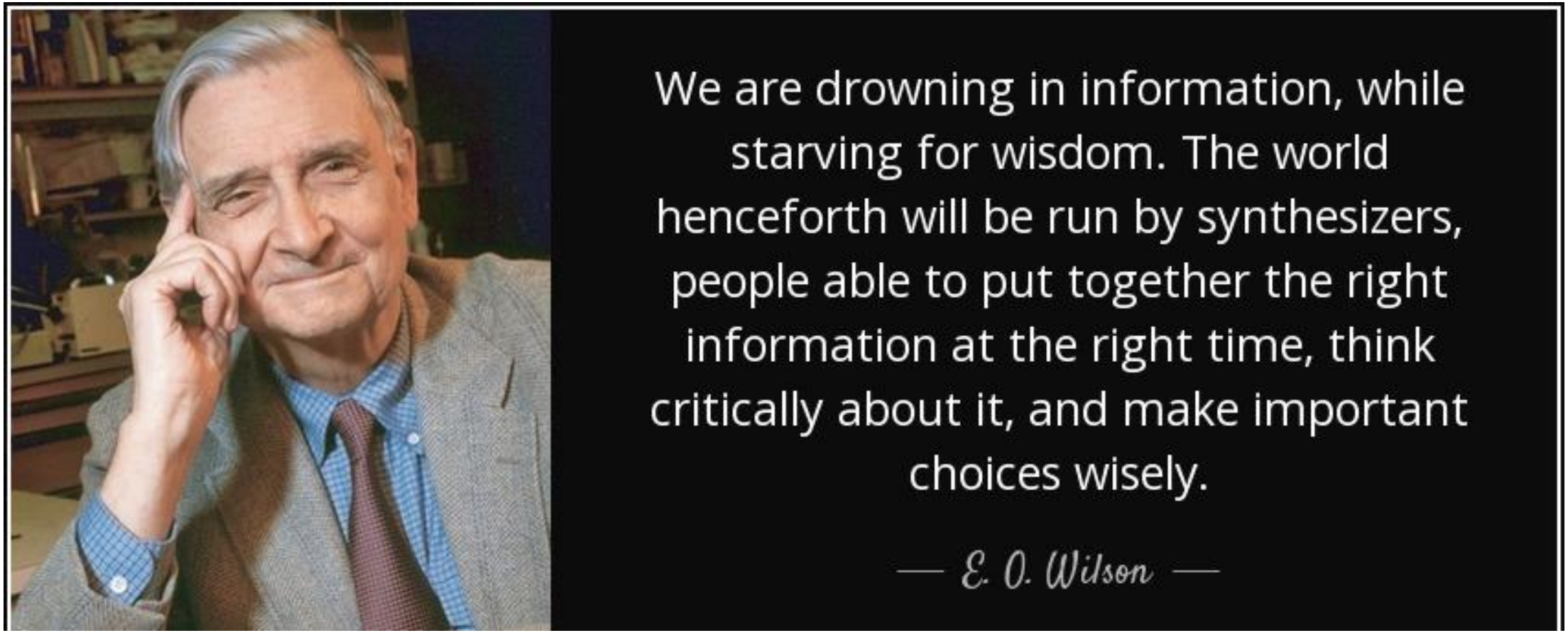
More than words

Text analytics in R


We are drowning in information, while starving for wisdom. The world henceforth will be run by synthesizers, people able to put together the right information at the right time, think critically about it, and make important choices wisely.
— E. O. Wilson —

Why are we here?

- Why text analytics
- How do we do it?

  Steps in data analytics
- Can you show it again but slooooowly?

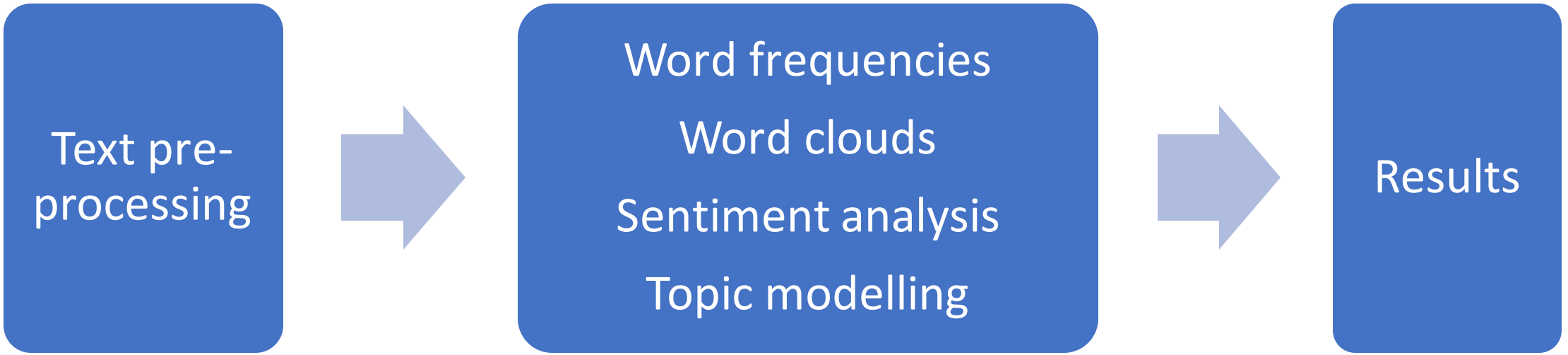
  R packages to work with data mining
- Oh! The places we will go!

  Application of text analytics

Why?

- Woolworths wants to know the magnitude of the technical outage on Monday on his customers' attitude
- George Clooney is about to run his first presidential election campaign for US and wants to know the chances of winning
- Did Shakespeare really write his own plays?
- What are trendy teenage topics?

Text analysis



R packages:
stringi
dplyr
readr
readtext

R packages:
tidytext
tm

wordcloud
topicmodels

Steps in text analysis

making choices that can affect the accuracy, validity, and findings

Steps:

- importing text
- Pre-processing: remove punctuation signs, stop words
- creating a document-term matrix (DTM), and
- filtering and weighting the DTM.

From alleged political bias to his personal character, here were the big issues from Facebook CEO Mark Zuckerberg's hearings on Capitol Hill.
(Jhaan Elker/The Washington Post)

Facebook chief executive Mark Zuckerberg appeared before the House Energy and Commerce Committee Wednesday for his second day of questioning on the Hill. Below is a partial transcript of the hearing.

[Transcript of Mark Zuckerberg's Senate hearing]

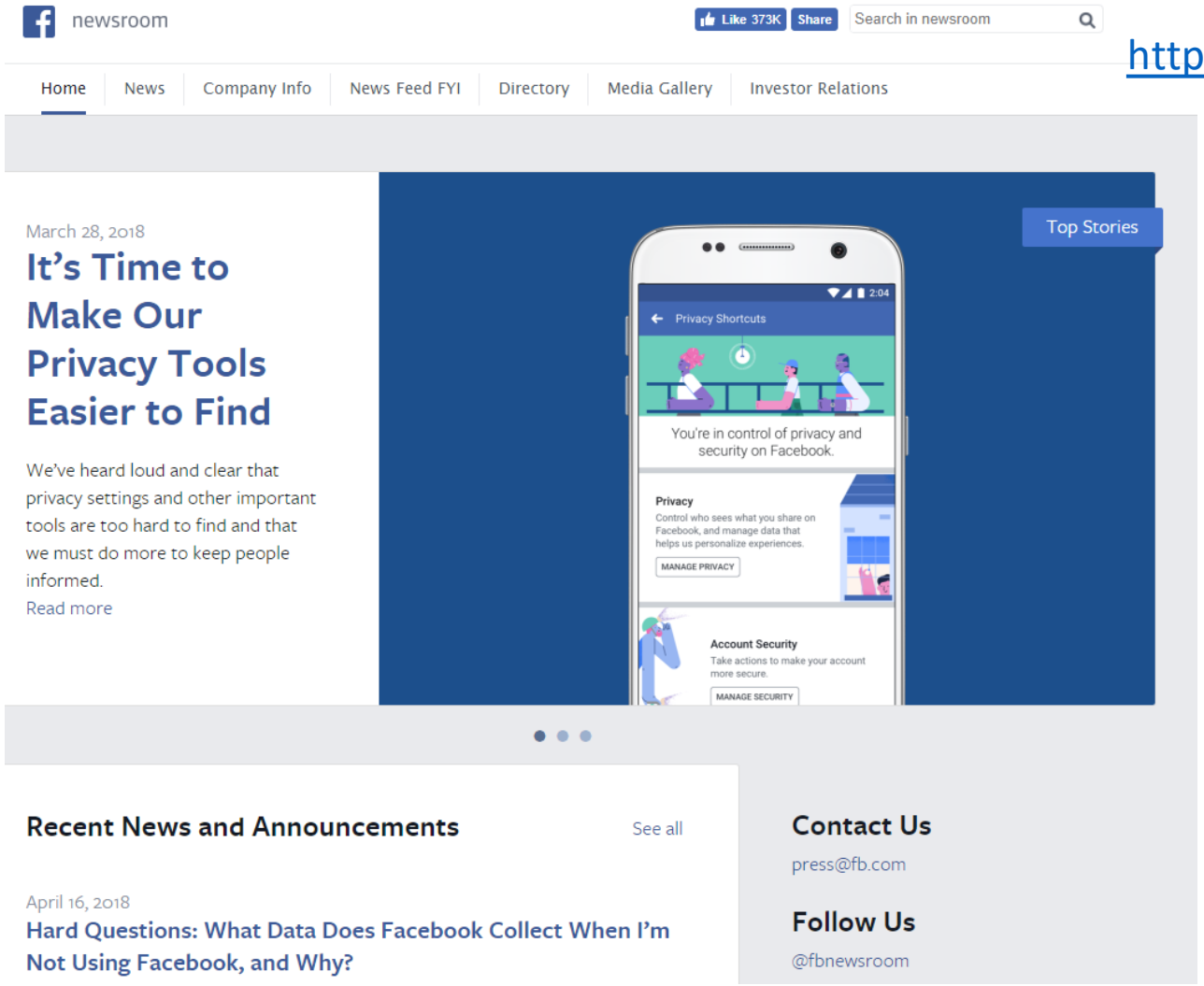
REP. GREG WALDEN (R-ORE.): Okay. I'd ask our guests to please take their seats so we can get started. The Committee on Energy And Commerce will now come to order.

WALDEN: Before my opening statement, just as a reminder to our committee members on both sides, it's another busy day at Energy and Commerce. In addition, as you will recall, to this morning's Facebook hearing, later today, our Health Subcommittee will hold its third in the series of legislative hearings on solutions to combat the opioid crisis.

And, remember, Oversight and Investigations Subcommittee will hold a hearing where we will get an update on the restoration of Puerto Rico's electric infrastructure following last year's hurricane season

Do we need every single word? Every character?

Sources



https://newsroom.fb.com/

From alleged political bias to his personal character, here were the big issues from Facebook CEO Mark Zuckerberg's hearings on Capitol Hill.
(Jhaan Elker/The Washington Post)

Facebook chief executive Mark Zuckerberg appeared before the House Energy and Commerce Committee Wednesday for his second day of questioning on the Hill. Below is a partial transcript of the hearing.

[Transcript of Mark Zuckerberg's Senate hearing]

REP. GREG WALDEN (R-ORE.): Okay. I'd ask our guests to please take their seats so we can get started. The Committee on Energy And Commerce will now come to order.

WALDEN: Before my opening statement, just as a reminder to our committee members on both sides, it's another busy day at Energy and Commerce. In addition, as you will recall, to this morning's Facebook hearing, later today, our Health Subcommittee will hold its third in the series of legislative hearings on solutions to combat the opioid crisis.

And, remember, Oversight and Investigations Subcommittee will hold a hearing where we will get an update on the restoration of Puerto Rico's electric infrastructure following last year's hurricane season

Transcript of Mark Zuckerberg's Senate hearing

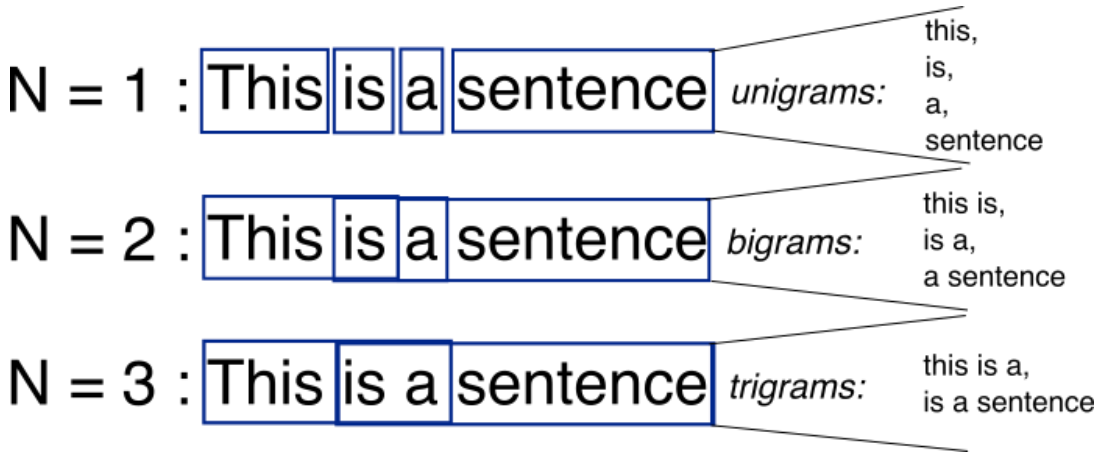
Data preparation: Importing text

- Flat text files (e.g. CSV and TXT)
- Formatted text files (JSON, HTML, XML, Word, Excel, PDF)
  - individual packages or
  - readtext package (combines everything in one place!)
- ! Character encodings:
  - all texts are encoded as UTF-8 - read as UTF-8 texts, or convert them to UTF-8
- Manipulations:
  - Parsing: joining, splitting, and extracting parts of strings
  - Use regular expressions to find or replace patterns
  - low-level string operations: e.g. removing html tags (stringi !)

Data preparation:

Tokenization: unstructured data -> structured data

- Split the text into smaller linguistic units, e.g. words, punctuation, numbers (“Los Angeles” and “rock 'n' roll”?)
  - words are separated by whitespace, punctuation marks or line breaks
- Ngrams: consecutive sequences of words
  - “hello world”

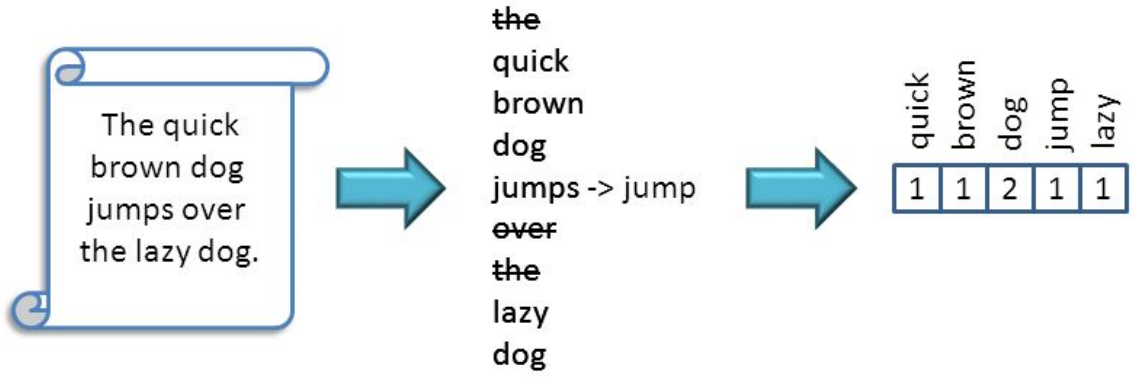


Data preparation:

The quick brown dog jumps over the lazy dog

Lowercasing and stemming:

reducing words to their root form



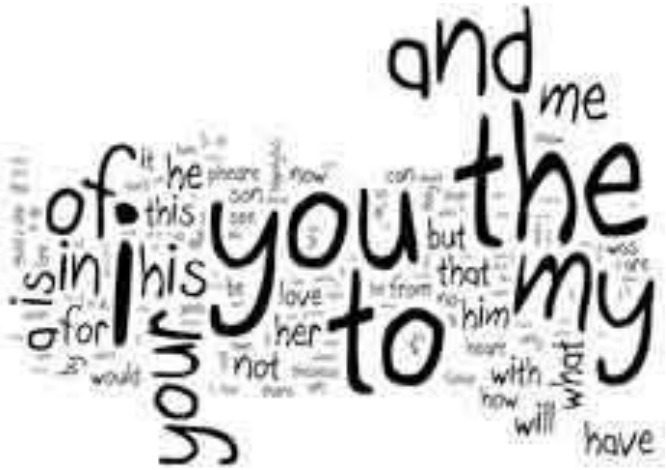
Regular expressions:

- Link , regex cheatsheet
- E.g. Remove html text, but keep retweets

- Stopwords: most common words in a language

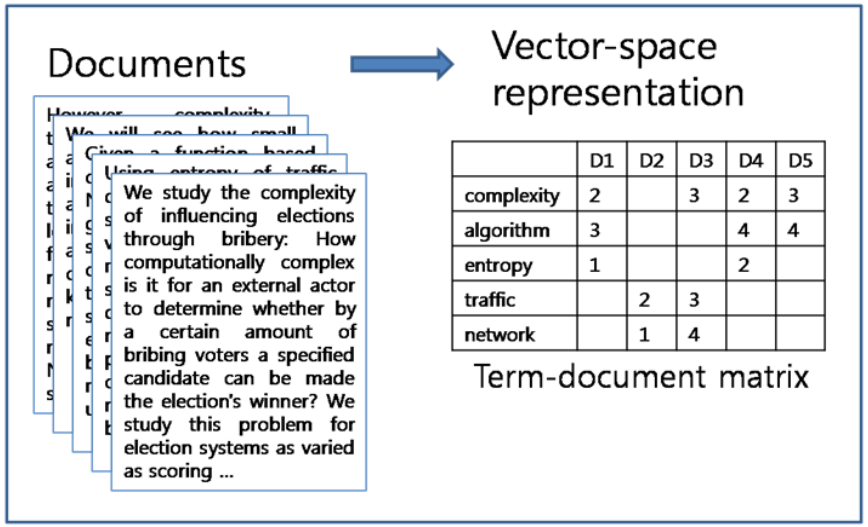
  the, *is*, *at*, *which*, and *on*

  Be careful! "The Who", "The The", or "Take That"



Text corpus: bag-of-words approach

- Document-term matrix (DTM): bag-of-words format
- a matrix in which:
  - rows are documents
  - columns are terms
  - cells: counts how often each term occurred in each document
- based on vector and matrix algebra
- very memory efficient, highly optimized operations.
- tm and quanteda package



	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

Term-document matrix

bangladesh	base	believ	benefit	better	blog	book	broader	calen
1	1	3	1	3	1	1	1	
0	1	0	0	1	0	0	0	
0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	
0	1	1	0	0	0	0	0	
0	0	1	0	0	0	0	0	
0	0	1	0	0	0	0	0	
0	1	2	0	3	0	0	1	
0	0	0	0	1	0	0	0	
0	1	0	0	3	0	0	0	
0	0	0	0	1	0	0	0	

Text corpus: tidy data

- Tidy data: **a table with one-token-per-row**
  - Each variable (=token) is a column
  - Each observation is a row
  - Each type of observational unit is a table
- A token: a meaningful unit of text (e.g. a word, n-gram, sentence, or paragraph)
- Tokenization: splitting the text into commonly used units of interest.
- can be less memory efficient and make matrix algebra less easily applicable

Tidyverse

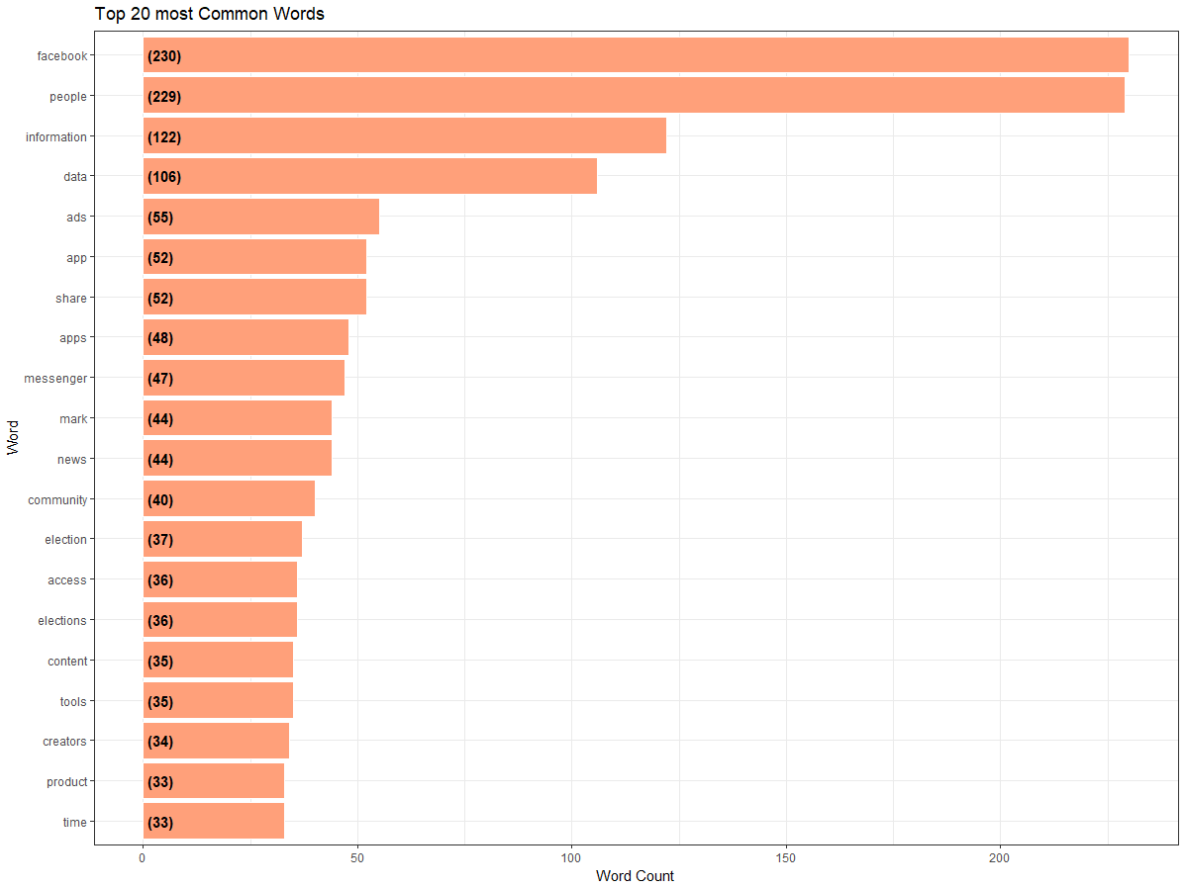
- Tidyverse is a collection of R packages for data science (text analytics!)
- Package list
- Dplyr: data manipulation made easy:
  - mutate() adds new variables that are functions of existing variables
  - select() picks variables based on their names.
  - filter() picks cases based on their values.
  - summarise() reduces multiple values down to a single summary.
  - arrange() changes the ordering of the rows.
- Piping %>%

Most common use

  - x %>% f is equivalent to f(x)
  - x %>% f(y) is equivalent to f(x, y)
  - x %>% f %>% g %>% h is equivalent to h(g(f(x)))
- Tutorial

Visualising text data

- Word Frequency:
  - see trends and patterns
  - check accuracy of pre-processing



- Word Clouds: A picture is worth 1,000 words.

Sentiment analysis

What are the emotions behind the words? What is the public opinion behind certain topics?

- Opinions are key influencers of our behaviors
- Whenever we need to make a decision, we often seek out the opinions of others.
  - Individuals: seek opinions from friends and family
  - Organizations: use surveys, focus groups, opinion polls, consultants.

Sentiment analysis: uses

- Multiple!
  - Stock exchange market: change in social media sentiment predict share prices
  - Politics: new policy announcements, election, public moods
  - Marketing: consumer opinion, benchmark products and services; market intelligence
  - Neuro-linguistic programming (NLP): "the conscious mind is the goal setter, and the unconscious mind is the goal getter"
    - Identifying and using the thinking strategies and emotional states
  - Individuals: Make decisions to buy products or to use services, find public opinions about political candidates and issues

What are your emotions?

- constructed via either crowdsourcing (using, for example, Amazon Mechanical Turk) or
- by one of the authors,
- validated using some combination of crowdsourcing again, restaurant or movie reviews, or Twitter data.
- Most general-purpose lexicons are
  - AFINN from Finn Årup Nielsen
    - negative, positive
    - assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.
  - bing from Bing Liu
    - positive and negative categories.
  - Loughran lexicon
    - "litigious", "uncertainty", "constraining", and "superfluous".
  - nrc from Saif Mohammad and Peter Turney:

    Emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, trust

Topic Modelling

- classification of documents (similar to clustering on numeric data!)
- unsupervised - we do not know what to look for before we start!
- put items into “natural” groups

Latent Dirichlet allocation (LDA)

- each document as a mixture of topics, and each topic as a mixture of words.
- documents can “overlap” each other- do not need to separate them into discrete groups (very natural language like!)

The Challenge!