Introduction to Text-as-Data & Word Embeddings

Dr. Fabienne Lind (University of Vienna)

Wednesday | 23.07.2025

Summer School for Women in Political Methodology (2025)

Text

Workshop objectives

- Getting to know (basic) procedures of computational text analysis
- **Apply** basic text as data methods
- Insights into **practical challenges**
- **Critical reflection** on the method
- **Inspiration** for your own projects

Course philosophie

Topics are covered with

- Lecture style input
- Group discussions
- Guided coding sessions

Interrupt me, ask all kinds of questions

Course materials

https://github.com/fabiennelind/text-as-data-in-R

Agenda

09:00 - 09:15	Orga
09:15 - 12:30	Introduction to Text-as-Data
12:30 - 13:30	Lunch Break
13:30 - 17:00	Word Embeddings
17:00 - 18:00	Feedback on your projects (Optional)

Feedback on your projects (optional)

When:

- Today after class 17:00 to 18:00

What:

- Informal opportunities to talk about your text analysis use case and (initial) design (plan), Research question, Data, Collection, Current struggles
- Receive feedback

About me

Fabienne Lind

- Communication Scientist
- Methods development: Multilingual computational text analysis, comparative analysis
- Substantial focus: Political communication, environmental communication
- fabienne.lind@univie.ac.at

</Computational> Communication Science Lab
Department of Communication | University of Vienna

universität wien

CiDAPE
Climate, Inequality & Democratic Action: The Force of Political Emotions

Your turn :)

https://www.menti.com/al7epayw63gd

Introduction to Text-as-Data

Agenda

09:15 - 10:45	Dictionary methods Overview + R Session
10:45 - 11:15	Coffee break
11:15 - 12:30	Supervised machine learning Overview + R Session

Motivations to analyse text automatically

Limitations of Manual Coding:

- Slow
- Limited scale
- Reliability issues

Motivations to analyse text

- Huge volumes of digital available information
- These data can be useful to research media contents, communicators, effects, and user interactions
- Explain, understand, predict feelings, attitudes and behaviour of individuals, groups, societies

Barberá, P., & Steinert-Threlkeld, Z. C. (2020). How to use social media data for political science research. *The SAGE handbook of research methods in political science and international relations*, *2*, 404-423.

Example 1: Contents and Effects

Research Question: How prevalent is each frame in NGOs' tweets?

- **Approach:** Dictionary

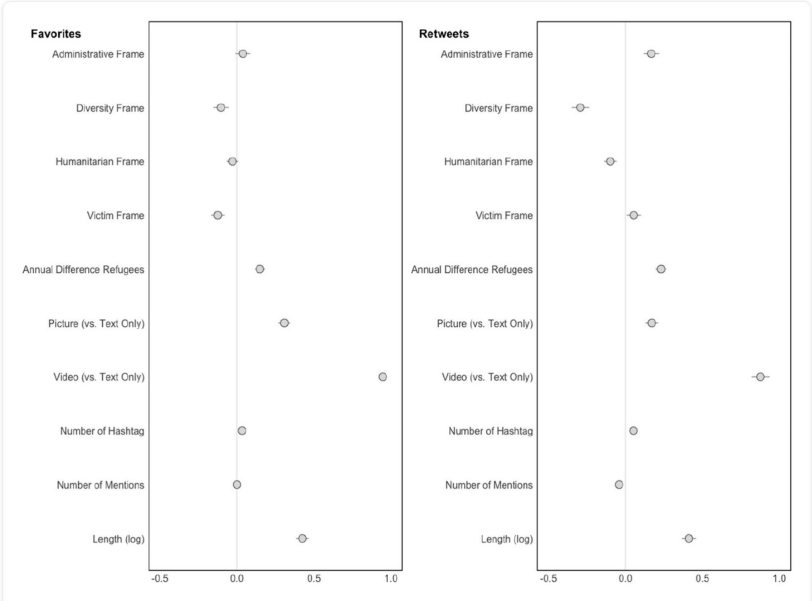
Table 3. Frame frequencies in Twitter communication of top 10 humanitarian NGOs.

Frame	Frequency (percent)
Administrative frame	2301 (11.78)
Diversity frame	2028 (10.39)
Humanitarian frame	4411 (22.59)
Victim frame	3075 (15.75)

Note. N = 19,528. Part of the sample did not contain any of the four frames measured.

Dimitrova, D., Heidenreich, T., & Georgiev, T. A. (2022). The relationship between humanitarian NGO communication and user engagement on Twitter. *New Media & Society*, 14614448221088970.

- **Research Question:** What reactions do the frames provoque?

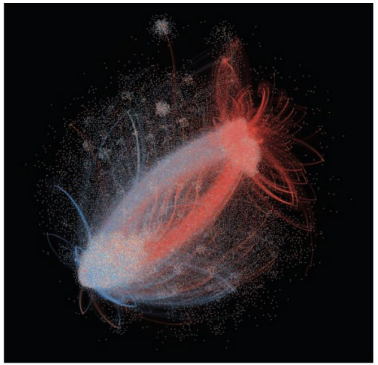


Example 2: Data selection for network analysis

Data selection requirement: Which of the 12 defined topics was mentioned in which tweet?

- **Approach:** Dictionary

Dark red = conservative
dark blue = liberal



2012 Presidential Election

2013 Super Bowl

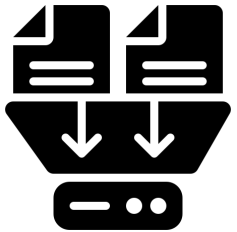
Barbera et al., 2015, Fig. 3

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber?. *Psychological science*, *26*(10), 1531-1542.

Purpose of obtaining measures for a large number of documents

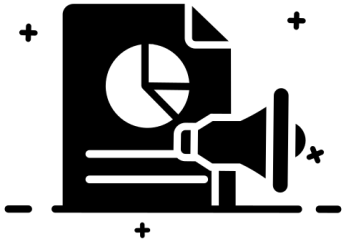
With rich data, more advanced filtering becomes possible

- Examples: linkage studies combination with media usage data; speech data for various actors



Evidence-based policy making

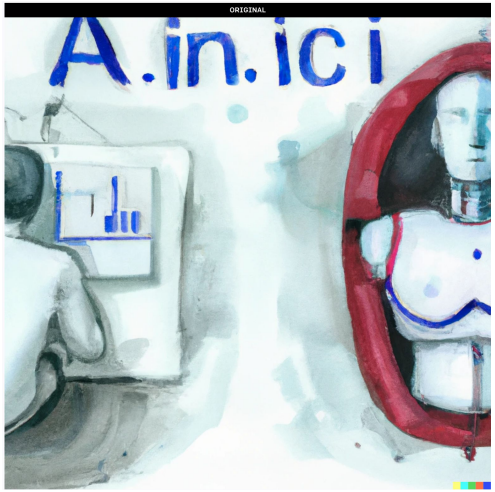
- Making the opinion of populations visible,
- Holding those in power accountable



González-Bailón, S. (2013). Social science in the era of big data. *Policy & Internet*, *5*(2), 147-160.

The end of manual coding?

- Augmenting not replacing (Grimmer & Steward, 2013)
- Human input for quality control:
  - select, monitor, and test on the level of data, inputs, process, outputs



Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Some major challenges when working with large corpora

Big data, big bias?

The end of theory?

Generalizing from online to offline behavior

Ethical concerns

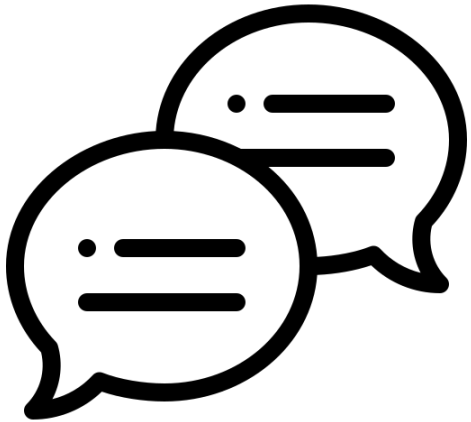
Barbera, 2021; González-Bailón, 2013

Text types (Examples)

- Social media posts and comments
- Transcribed speeches
- Wikipedia articles
- Open answer questions in surveys
- CVs
- Transcripts of interviews
- AI generated texts
- Etc.

Your projects

- What are your applications of text analysis?
- What text types do you study?
- Where do you see challenges?



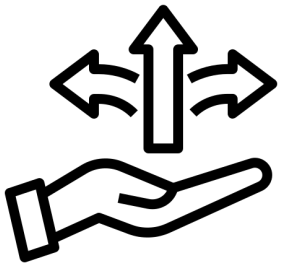
Dictionaries

Dictionary method

Definition: List of terms (words or phrases) that indicate a category or concept

Application: Count occurrences of dictionary terms within documents to score them on the concept

Approach: Theory-driven, use when a category is known (= defined by researcher)



Example: Dictionary for the concept “EU countries”

```
eu_countries <-c("Austria", "Belgium", "Bulgaria", "Croatia", "Cyprus",
"Czech Republic", "Denmark", "Estonia", "Finland", "France",
"Germany", "Greece", "Hungary", "Ireland", "Italy", "Latvia",
"Lithuania", "Luxembourg", "Malta", "Netherlands", "Poland",
"Portugal", "Romania", "Slovakia", "Slovenia", "Spain", "Sweden")
```

Is the list exhaustive to match all mentions of EU countries in text?

Example: Dictionary for the concept “EU countries”

```
eu_countries <-c("Austria", "Belgium", "Bulgaria", "Croatia", "Cyprus",
"Czech Republic", "Denmark", "Estonia", "Finland", "France",
"Germany", "Greece", "Hungary", "Ireland", "Italy", "Latvia",
"Lithuania", "Luxembourg", "Malta", "Netherlands", "Poland",
"Portugal", "Romania", "Slovakia", "Slovenia", "Spain", "Sweden")
```

Is the list exhaustive to match all mentions of EU countries in text?

- Abbreviations (AT), different common spellings (e.g., Czechia), historical data (UK)

Dictionary use case 1: Filter Data

Example szenario:

- You have a text corpus about inflation.
- You are interested in analysing articles that discuss the inflation topic in the context of the next national election.
- You construct a dictionary to filter the data.

Dictionary use case 1: Filter Data

- All texts are about inflation. What words indicate that some are also related to a national election?

The Swiss National Bank became the first central bank among advanced economies to declare victory over the post-pandemic surge in inflation, cutting its key interest rate by 0.25 percentage point to 1.5 percent and hinting at more cuts in the coming months.

Politico.eu, March 21, .2024

"Even as incomes are going up and the economy is doing well and inflation is coming down, people can't buy homes," said Daryl Fairweather, chief economist at the brokerage Redfin. "That's like the biggest problem for Biden because it's not one that he can solve."

AP News, March 15.2024

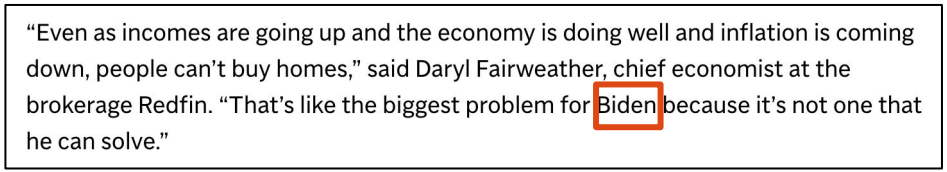
Global Fuel Prices Are Surging With Supply Risks Ahead

- US pump prices 60% higher than 2020, when Biden was elected
- Global oil demand and refinery runs are both forecast to rise

Bloomberg, March 12.2024

Dictionary use case 1: Filter Data

- Keep only articles where your words appear



AP News, March 15.2024

Bloomberg, March 12.2024

Dictionary use case 2: Measure classes of category

Example szenario:

- You have a corpus about inflation and the next national election
- You design dictionaries to know which article is about housing and which is about energy

Dictionary use case 2: Measure classes of category

- What words indicate that a text is about the housing topic?
- What words indicate that a text is about the energy topic?

"Even as incomes are going up and the economy is doing well and inflation is coming down, people can't buy homes," said Daryl Fairweather, chief economist at the brokerage Redfin. "That's like the biggest problem for Biden because it's not one that he can solve."

AP News, March 15.2024

Global Fuel Prices Are Surging With Supply Risks Ahead

- US pump prices 60% higher than 2020, when Biden was elected
- Global oil demand and refinery runs are both forecast to rise

Bloomberg, March 12.2024

Dictionary use case 2: Measure classes of category

- What words indicate that a text is about the housing topic?
- What words indicate that a text is about the energy topic?

"Even as incomes are going up and the economy is doing well and inflation is coming down, people can't buy homes," said Daryl Fairweather, chief economist at the brokerage Redfin. "That's like the biggest problem for Biden because it's not one that he can solve."

AP News, March 15.2024

Global Fuel Prices Are Surging With Supply Risks Ahead

- US pump prices 60% higher than 2020, when Biden was elected
- Global oil demand and refinery runs are both forecast to rise

Bloomberg, March 12.2024

Application: Classify actor salience

Objective: salience of women and men migrant's in the news across time and outlet

- Dictionary approach to measure mentions of women and men migrants in German news articles

Lind, F., & Meltzer, C. E. (2021). Now you see me, now you don't: Applying automated content analysis to track migrant women's salience in German news. *Feminist Media Studies*, *21*(6), 923-940.

Example: Classify actor salience

Keyword list more complex than initially thought

```
person_f_migrant_endIn_regex  = c("[Aa]sylantin","[Aa]sylbewerberin", "[Aa]
                                 "[Zz]uwanderin", "[Ee]inwanderin", "[Gg]a
                                 "[Aa]usländische\\w{0,2}\\s[Bb]ürgerin",

f_relation_regex = c("[Mm]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Mm]ein\\
                     "[Dd]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Dd]ein\\w{
                     "[Ss]eine\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ss]ein\\w{
                     "[Ii]hr\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ii]hr\\w{0,2
                     "[Uu]ns\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Uu]ns(er|re)
                     "[Ee]u(er|re)\\w{0,2}\\s[Ff]rau(en|)(\\s|\\W)", "[Ee]u(e

f_relation_regex <- paste(apos_closed, f_relation_regex, "", sep = "") #add
df_f_relation_regex <- data.frame(f_relation_regex) #as. dataframe
```

Table 2. Dictionary subcategories for the measurement of the concepts: migrant women's and migrant men's salience.

Subcategory name	Description	Measured concepts (examples)	
		Migrant women's salience	Migrant men's salience
1. Impersonal designation	General person nominations for migrants	(immigrant, asylum seeker, etc.) with female or gender-neutral word endings (-In,/-in, etc.)	(immigrant, asylum seeker, etc.) with generic masculine form or gender-neutral word endings (-In,/-in, etc.)
2. Origin: Single words			
	Nominations that refer to the territorial origin of a person [a]	(French, African, Syrian, etc.) with female or gender-neutral word endings (-In, -/in, etc.)	(French, African, Syrian, etc.) with generic masculine form or gender-neutral word endings (-In, -/in, etc.)
3. Origin: Combinations of different word groups [b]			
a	General gender-related person nominations + from + general territorial denominations [c]	(women, girl, mother, sister) + from + (France, Africa, Syria, etc.)	(man, boy, father, brother) + from + (France, Africa, Syria, etc.)
b	General territorial denominations (adjectives) [c] + General gender-related person nominations	(French, African, Syrian, etc.) + (women, girl, mother, sister)	(French, African, Syrian) + (man, boy, father, brother)
4. In relation	Relational expressions: possessive pronouns + General gender-related person nominations	(my, your, her, his, our, yours, their) + (women, girl, mother, sister)	(my, your, her, his, our, yours, their) + (man, boy, father, brother)
5. As phrase	Phrases	e.g., "women and children"	e.g., "men and children"
6. Other expressions		e.g., "women from abroad"	e.g., "men from abroad"

As general notes, the dictionary includes the singular and plural version of all words, word endings (e.g., for prepositions) consider the different cases used in the German language.

[a]Downloaded from the CLDR (Unicode Common Locale Data Repository) http://cldr.unicode.org/, which holds standard name translations of countries and regions (version v33.1).

[b]Measured at the sentence level.

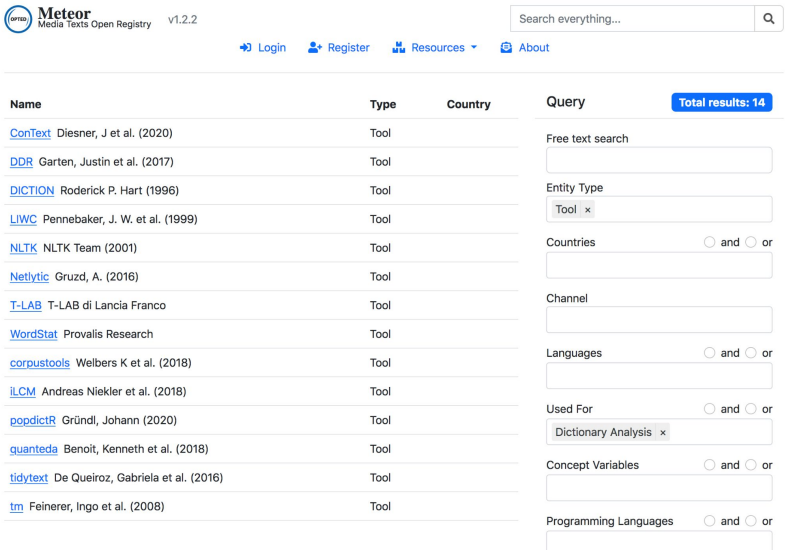
[c]Manually compiled by a native speaker for all CLDR territorial denominations, which the German language allows (e.g., no separate word for many smaller islands, e.g., Isle of Man, Curaçao). Assisted by the preeminent German language dictionary duden.de.

Lind, F., & Meltzer, C. E. (2021). Now you see me, now you don't: Applying automated content analysis to track migrant women's salience in German news. *Feminist Media Studies*, *21*(6), 923-940.

Many existing dictionaries

You find a list at

https://meteor.opted.eu/



Dictionaries = Search strings?

- When dictionaries are used to search for data in a repository they are called search strings

Table 3 Boolean search strings used for retrieval of migration-related news articles

Country	Language	Search string
Spain	Spanish	asilo* OR inmigra* OR refugiad* OR migrante* OR migratori* OR "sin papeles" OR "campo de desplazados" OR patera* OR emigra* OR "libre circulación" OR "fuga de cerebros"
UK	English	asyl* OR immigrant* OR immigrat* OR migrant* OR migrat* OR refugee* OR foreigner* OR "undocumented worker*" OR "guest worker*" OR "foreign worker*" OR emigrat* OR "freedom of movement" OR "free movement"
Germany	German	asyl* OR immigrant* OR immigriert* OR immigrat* OR migrant* OR migrat* OR flüchtling* OR ausländer* OR zuwander* OR zugewander* OR einwander* OR eingewander* OR gastarbeiter* OR "ausländische arbeitnehmer*" OR emigr* OR auswander* OR ausgewander* OR personenfreizügigkeit* OR arbeitnehmerfreizügigkeit* OR "freier personenverkehr*"

Heidenreich et al., 2020; Lind et al., 2020

Heidenreich et al., 2020; Lind et al., 2020

Validate, validate, validate

Table 1 Four principles of quantitative text analysis

(1) All quantitative models of language are wrong—but some are useful.
(2) Quantitative methods for text amplify resources and augment humans.
(3) There is no globally best method for automated text analysis.
(4) Validate, Validate, Validate.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. Political analysis, 21(3), 267-297.

Dictionary validation

- Face validity: Do terms intuitively make sense?
- External validity: Compare dictionary scores against manual codes or other external measures.

External validity

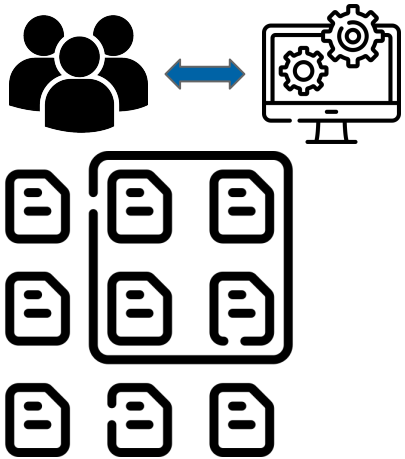
- **Construct validity:** How close is an automated measurement to a more trusted measurement

- The more trusted measurement is **typically human understanding of text**
    - Why? As social scientists, we typically care about how readers understand a text, or what communicators intended to say

Validation with manually created baseline

Steps

1. Select a subset of the documents
2. Label the subset manually
3. Classify the subset automatically (with dictionary)
4. Compare manual decisions with automated decisions (via metrics calculation: recall, precision, F1)



Low metrics? → improve the dictionary and calculate metrics again

Creation of a manual baseline

- Codebook: Rules, Examples
- Coder Selection: Expert coders vs. Crowdcoders; Coder Training
- Quality Assessment: Inter-coder reliability, majority vote, *valid disagreement* (Baden et al., 2023)
- Document Selection: Representative for target discourse (e.g., random selection or artificial week)

Baden, C., Boxman-Shabtai, L., Tenenboim-Weinblatt, K., Overbeck, M., & Aharoni, T. (2023). Meaning multiplicity and valid disagreement in textual measurement: A plea for a revised notion of reliability. *SCM Studies in Communication and Media*, *12*(4), 305-326.

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures*, *11*(3), 191-209.

Tools for manual coding

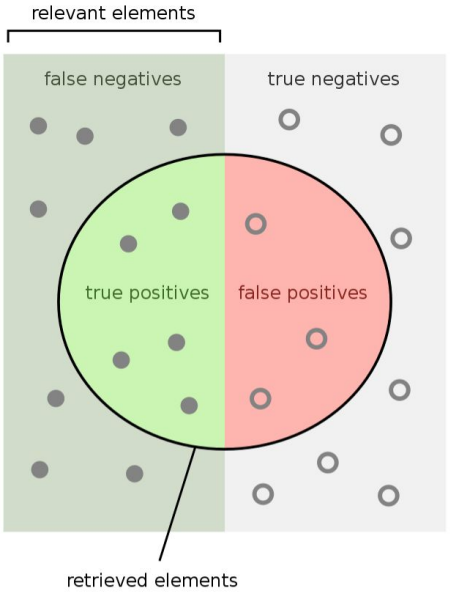
- Google sheets
- Excel
- Label Studio: https://labelstud.io/academic/
- AmCAT: https://vu.amcat.nl/accounts/login/?next=/
- AnnoTinder: https://github.com/ccs-amsterdam/CCS_annotator

Recall, precision?

Metrics frequently used to express the validity of a dictionary & other automated classification methods

Example category: Inflation (yes/no)

- Recall: Of all the actual texts about “Inflation” cases, how many did the model find?
- Precision: Of all the times the model predicted “Inflation”, how often was it correct?



Scharkow, M. (2011). Zur Verknüpfung manueller und automatischer Inhaltsanalyse durch maschinelles Lernen. *M&K Medien & Kommunikationswissenschaft*, *59*(4), 545-562.

F1

The F1-Score is the harmonic mean of Precision and Recall. It provides a single, balanced score.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

What to report? F1 alone is not sufficient, can hide problems with recall or precision

Summary Dictionaries

Pros

- Often needed to select data (search strings)
- High reliability and control
- High transparency and reproducibility

Cons

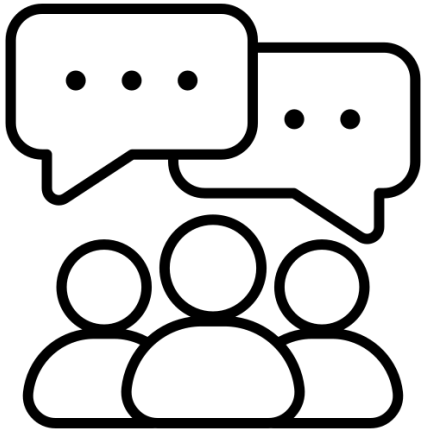
- Difficulty increases with the latency of the construct
- May miss language nuances

What do we see here?

"#\\w+"

"\\b\\d{4}\\b"

"(?<=Speaker:)\\b[A-Z][a-z]+\\s[A-Z][a-z]+"



What do we see here?

Regular expressions:

"#\\w+" Extracts hashtags, often used for social media analysis

"\\b\\d{4}\\b" Extracts 4-digit numbers, often used to find years.

"(?<=Speaker:)\\b[A-Z][a-z]+\\s[A-Z][a-z]+" Extracts speaker names following "Speaker: " in text.

Regular Expressions (regex)

- Formal language to specify search strings

Regular Expressions (regex)

- Formal language to specify search strings
- Insanely difficult

Regular Expressions (regex)

- Formal language to specify search strings
- INSANELY difficult

Regular Expressions (regex)

- Formal language to specify search strings
- ***INSANELY*** difficult

Regular Expressions (regex)

- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything

Regular Expressions (regex)

- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything
- Different ***flavours***

Regular Expressions (regex)

- Formal language to specify search strings
- ***INSANELY*** difficult
- Nobody can remember anything
- Different ***flavours***

- “Some people, when confronted with a problem, think “I know, I’ll use regular expressions.” Now they have two problems.” *Jamie Zawinski*

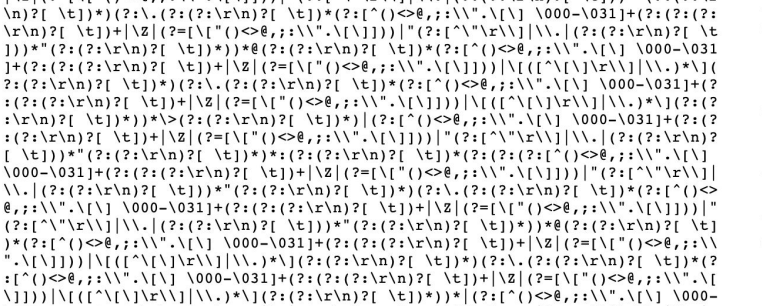
How difficult to regex an email

How difficult to regex an email

Rather

How difficult to regex an email

```
(?:[a-z0-9!#$%&'*+/=?^_`{|}~-]+(?:\.[a-z0-9!#$%&'*+/=?^_`{|}~-]+)*|"(?:[\x01-
\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:
(?:[a-z0-9](?:[a-z0-9-]*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]*[a-z0-9])?|\[(?:(?:(2(5[0-
5]|[0-4][0-9])|1[0-9][0-9]|[1-9]?[0-9]))\.){3}(?:(2(5[0-5]|[0-4][0-9])|1[0-9][0-9]|[1-
9]?[0-9])|[a-z0-9-]*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\x7f]|\\
[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])
```




Grimmer / Jurafsky
Cheat-sheet

- Disjunctions

RE	Match	Example Patterns Matched
`[mM]oney`	Money or money	"Money"
`[abc]`	'a', 'b', *or* 'c'	"Investing in Iran"
		"is dangerous business"
`[1234567890]`	any digit	"sitting on $7.5 billion dollars"
		"2005 and 2006, more than "
		"$150 million dollars"
`[\.]`	A period	" 'Run!', he screamed."

- Ranges

RE	Match	Example Patterns Matched
`[A-Z]`	an upper case letter	"Rep. Anthony Weiner
		(D-Brooklyn & Queens)"
`[a-z]`	a lower case letter	"ACORN's"
`[0-9]`	a single digit	"(9th CD) "

Grimmer / Jurafsky
Cheat-sheet

- Negations

RE	Match	Example Patterns Matched
`[^A-Z]`	not an upper case letter	"ACORN**'s**"
`[^Ss]`	neither 'S' nor 's'	"**ACORN'**s"
`[^\.]`	not a period	" 'Run!', he screamed."

- Optional Characters: `?`, `*`, `+`

RE	Match	Example Patterns Matched
`colou?r`	Words with `u` 0 or 1 times	"color" or "colour "
`oo*h!`	Words with `o` 0 or more times	"oh!" or "ooh!" or "oooh!"
`o+h!`	Words with `o` 1 or more times	"oh!" or "ooh!" or "oooooh!" or

Grimmer / Jurafsky
Cheat-sheet

- Start of the line anchor ^, end of the line anchor **$**

RE	Match	Example Patterns Matched
`^[A-Z]`	Upper case start of line	"Palo Alto" "the town of **P**alo Alto"
`^[^A-Z]`	Not upper case start of line	"the town of Palo Alto" "**P**alo Alto"
`^.`	Start of line	"Palo Alto" "the town of Palo Alto"
`.$`	Identify character that ends a line	"Wait**!**" "This is the end**.**"

- "Or"| statements, Useful short hand

RE	Match	Example Patterns Matched
yours|mine	Matches"yours" or "mine"	"it's either yours or mine"
\ d	Any digit	"1-Mississippi"
\ D	Any non-digit	"1-Mississippi"
\ s	Any whitespace character	"1, 2"
\ S	Any non-whitespace character	"1, 2"
\ w	Any alpha-numeric	"1-Mississippi "
\ W	Any non-alpha numeric	"1-Mississippi"

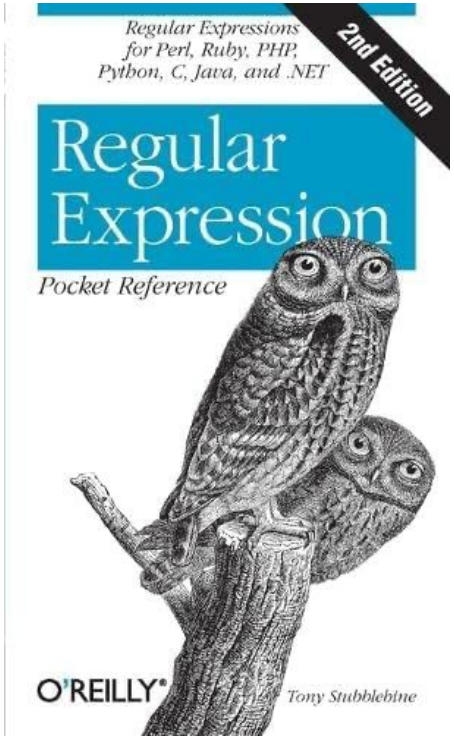
Helpers

GPT is great with regex

https://regex101.com

Recommended R tutorial (Wickham & Grolemund, 2017):
https://r4ds.had.co.nz/strings.html

To test regular expressions quickly:
https://spannbaueradam.shinyapps.io/r_regex_tester/



Coding session

Jointly:

1. Use an existing dictionary:
   sentiment_dictionary.Rdm
   LSDprep_dec2017.R
2. Create a dictionary + validate it:
   dictionary_creation_and_validation.RmD

Get scripts:

https://github.com/fabiennelind/text-as-data-in-R

Supervised classification

Supervised machine learning

Objectives:

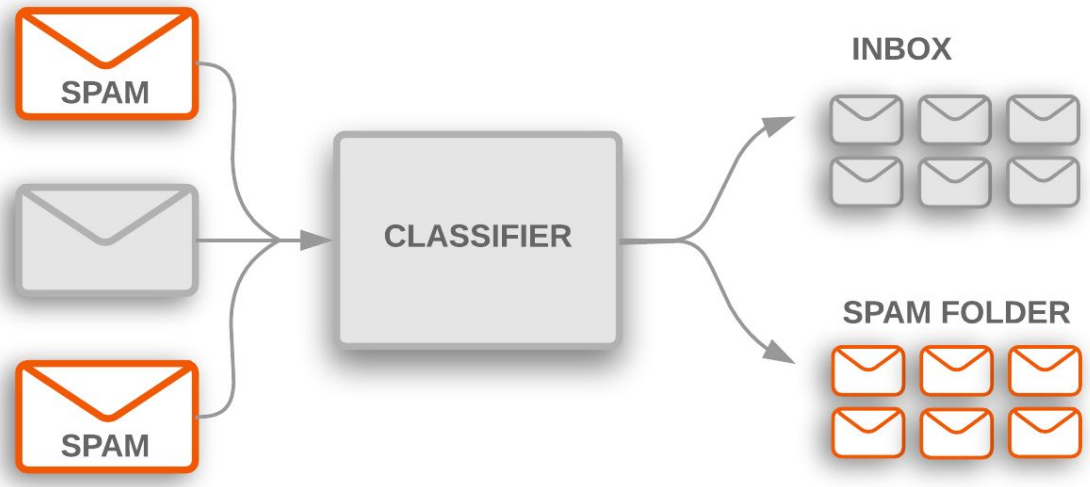
- Classification (for categorical variables)
  - E.g.: classify documents into pre existing categories
- Regression (for continuous variables)

Supervised machine learning

Objectives:

- **Classification (for categorical variables)**
  - **E.g.: classify documents into pre existing categories**
- Regression (for continuous variables)

Email classification

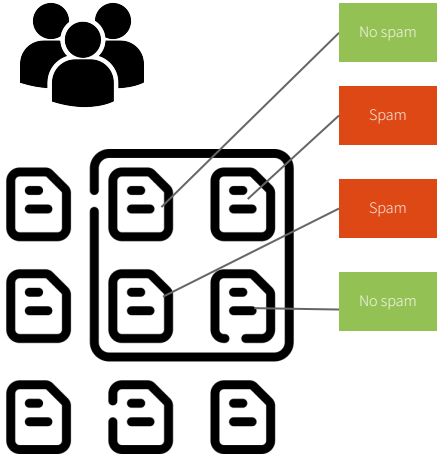


Steps

1. Create a labeled data set
2. Training phase
3. Check performance
4. Use the model

Grimmer et al., 2022

1. Create a labeled data set



How:

- Typically: Human coders label parts of the corpus

Considerations:

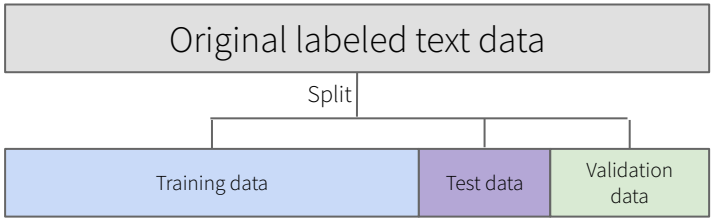
- Sampling should be representative for the corpus (e.g., Random, Stratified sample e.g., across time and source)
- Quality of human coding matters (intercoder reliability)
- Number of documents

Song, H., Tolochko, P., Eberl, J. M., Eisele, O., Greussing, E., Heidenreich, T., ... & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, *37*(4), 550-572.

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19-42.

1. Create a labeled data set

Split labeled data in training data, test data, validation set



Training data

- The subset that is used to learn the model parameters

Test data

- Another subset used to evaluate the model's predictive quality
- Not used for learning!

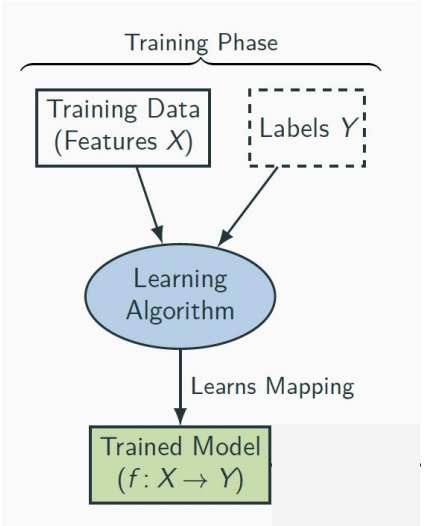
Validation data

- Only used to evaluate in the end, no further optimization allowed

2. Training phase

Classifier learns the mapping between (text) features and the labels in the training set

- We define a model $f(Y)=g(X)$
- And apply a learning algorithm to establish which features in X (features extracted from the training documents) matter to recover Y (i.e, the labels of the training documents)
- We fit the model



Grimmer et al., 2022

2. Training phase

Considerations:

- Feature representation: Bag of words representation or embeddings
- Feature selection: Remove irrelevant features
- Classifier selection: e.g., Naive Bayes, SVM, KNN, or ensemble methods

Grimmer et al., 2022

2. Training phase

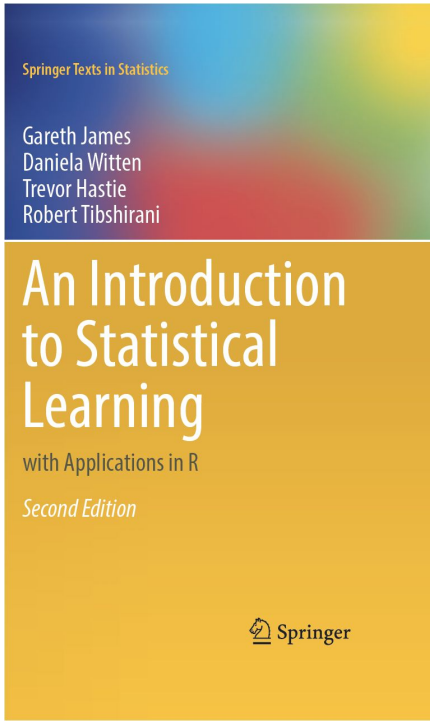
Naïve Bayes Classifier

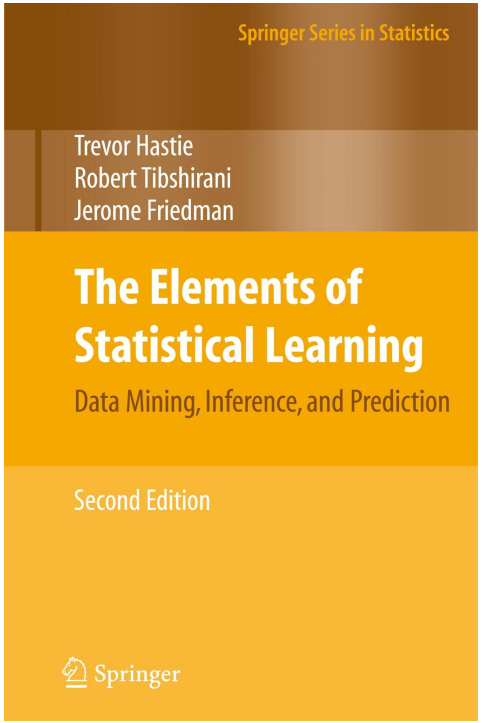
- Probabilistic classifier
- Simple
- Fast
- Good Accuracy

Support Vector Machine

- Comes from computer science
- Very good
- Rather difficult math
- Considered one of the best of-the-shelf classification algorithms

More details:





Performing supervised machine learning in R and Python

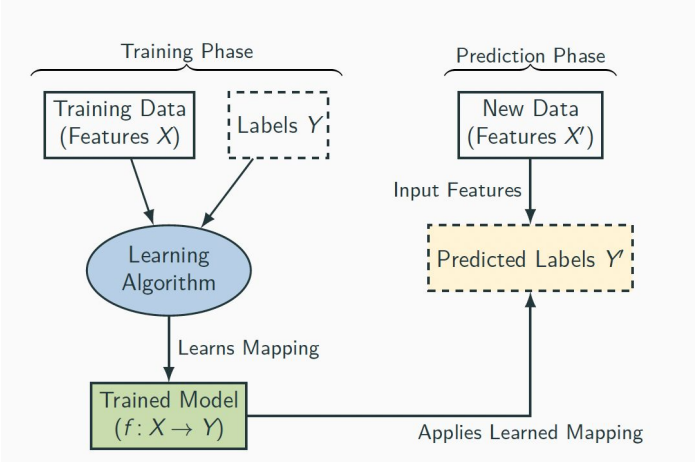
The classifiers are implemented in many stats/ML packages

R: quanteda, caret, e1071, klaR, C5.0, OneR

Python: scikit-learn

3. Check performance

The fitted model (the trained classifier) is applied to a held-out test set (was not used for training the model).



Grimmer et al., 2022

3. Check performance

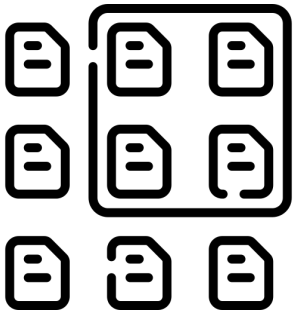
- Compare predicted labels with manual labels for the test set
  - Inspect Confusion matrix
  - Calculate recall, precision, F1
- Bad results for test set: Improve training phase
- Good results for test set: Repeat for validation set

Confusion matrix

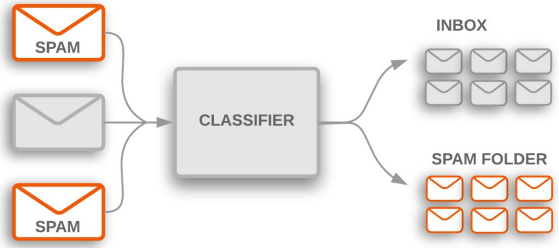
	Actual label	
Classification (algorithm)	Negative	Positive
Negative	True negative	False negative
Positive	False negative	True positive

4. Using the model

Remember that we labeled only parts of our original corpus.



When you are satisfied with the performance of the classifier, you can use it to classify all documents in the corpus

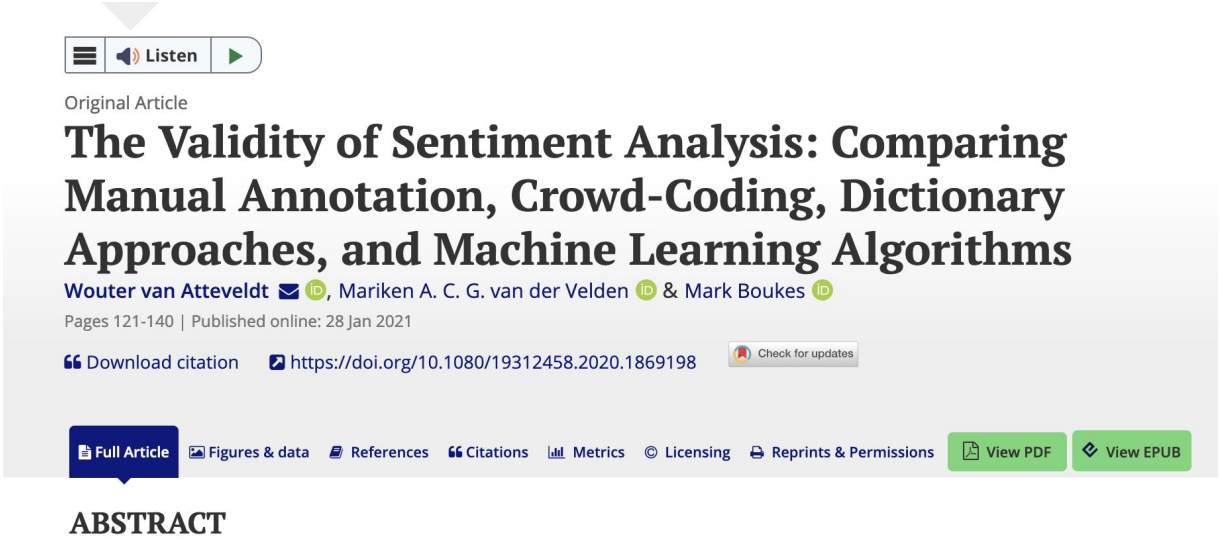


Grimmer et al., 2022

Dictionary vs. supervised machine learning

Category: Sentiment

Result: machine learning significantly outperformed dictionary coding



Van Atteveldt el al. (2021)

Dictionary vs. supervised machine learning

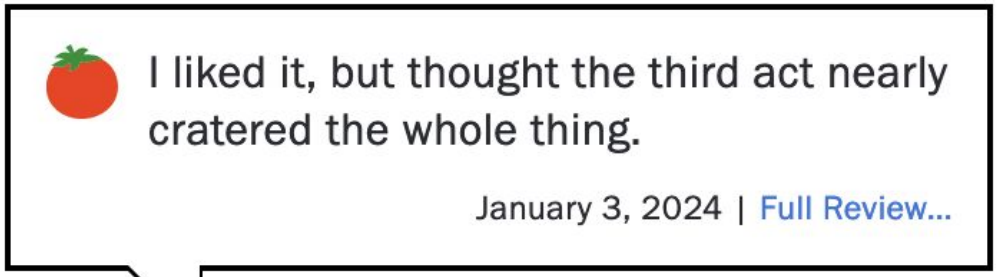
- Supervised machine learning requires (potentially larger amounts) of labeled data

- If the training sample is large enough supervised learning will outperform dictionaries

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 95-107.

Coding session

- Sentiment in movie reviews



Coding session

- Sentiment in movie reviews

I liked it, but thought the third act nearly cratered the whole thing.

January 3, 2024 | Full Review...

An intense, inventively filmed, and well-acted biopicture about the kinds of events that are hard on the heart.

Full Review | Original Score: 5/5 | Nov 20, 2023

In a movie about impending global catastrophe, he gives a close-up of a face, and a twitch of a lip the power of an atom bomb.

Full Review | Original Score: A | Nov 17, 2023

Coding session

Jointly:

Apply and validate classifier
supervised_nb_svm.RmD

Get script:

https://github.com/fabiennelind/text-as-data-in-R

What is next on the program

Lunch (12:30 – 1:30 PM)

Lunch

You can view the menu here:

Location:

Uni - Mensa
Universitäts-Boulevard 3
28359 Bremen

Word Embeddings

Agenda

13:30 - 13:45	Bag-of-words approach
13:45 - 15:00	Word embeddings
15:00 - 15:30	Coffee break
15:30 - 17:00	R Session

What is text?

- Data
- Unstructured
- Multidimensional (Highly)

In general, difficult to work with (if you’re not human)

Computers only understand numbers. How can we bridge this gap?

From Text to Structure

- We need to **“structure”** the text before we perform analyses
- Different ways to **represent** text so that computers “understand”
- Different ways to **model** text so that both (we and computer) “understand”

Two text representation approaches

- Bag-of-words
- Embeddings

Document-Term Matrix

$$X = \begin{bmatrix} 1 & 1 & \dots & 0 \\ 2 & 1 & \dots & 0 \\ \dots & \dots & \dots & \\ 2 & 0 & \dots & 3 \end{bmatrix}$$

X = N * K matrix
N = number of *documents*
K = number of *terms/features*

Example

Corpus (Collection of texts)

Document 1: “John loves icecream”
Document 2: “John loves oranges”
Document 3: “Marry hates icecream”

N? K?

	Terms					
Docs	icecream	john	loves	oranges	hates	marry
1	1	1	1	0	0	0
2	0	1	1	1	0	0
3	1	0	0	0	1	1

N = 3
K = 6

X = 3 * 6 matrix

Types & Tokens

Types: unique words in a text

Tokens: all words in the text

Types and tokens in our example corpus?

Types & Tokens

Types: unique words in a text

Tokens: all words in the text

Types and tokens in our example corpus?

6 types (unique words) / *9 tokens* (total length of the corpus)

Example Texts

Text_1 = "banana banana banana banana chocolate"

Text_3 = "chocolate chocolate chocolate banana fudge"

Text_2 = "banana banana"

Text_4 = "icecream icecream fudge ice-cream"

Text_5 = "fudge fudge fudge"

Text_6 = "ice-cream ice-cream fudge fudge"

How many documents? How many unique terms (types)?

Document-Term Matrix

banana	chocolate	fudge	icecream
4	1	0	0
2	0	0	0
1	3	1	0
0	0	1	3
0	0	3	0
0	0	2	2

Text 1: I like green apples but no green bananas

Bag-of-Words representation

Summary: Collection of words, each text is represented as a *count* of words contained in it

Advantages:

- Simple representation that is easy to understand, implement, and interpret

Disadvantage:

- For corpora with large vocabulary, memory constraints
- Ignores word order, grammar, context

Text id	I	like	green	apples	but	no	bananas
1							

Bag-of-Words representation

Text 1: I like green apples but no green bananas

Summary: Collection of words, each text is represented as a *count* of words contained in it

Advantages:

- Simple representation that is easy to understand, implement, and interpret

Disadvantage:

- For corpora with large vocabulary, memory constraints
- Ignores word order, grammar, context

Text id	I	like	green	apples	but	no	bananas
1	1	1	2	1	1	1	1

Using BoW representation

On its own:

- Helps understand and explore the corpus
- Enables analysis of word frequencies, vocabulary diversity, and common patterns

Combined with other methods:

- Converts text into a structured, quantifiable format
- Produces a word count matrix that feeds into:
  - Document classification
  - Topic modeling

Example: Vocabulary Diversity

Text A: "The cat sat on the mat."

Text B: "Innovation drives progress in technology and science."

Text A: 6 words total, 5 unique (cat, sat, on, the, mat)
Text B: 7 words total, 7 unique

Vocabulary diversity = (Unique words) / (Total words)

- Text A: 5 / 6 = 0.83
- Text B: 7 / 7 = 1.00

Text B has **higher vocabulary diversity**, more unique words relative to its length, which may indicate richer or more specialized content.

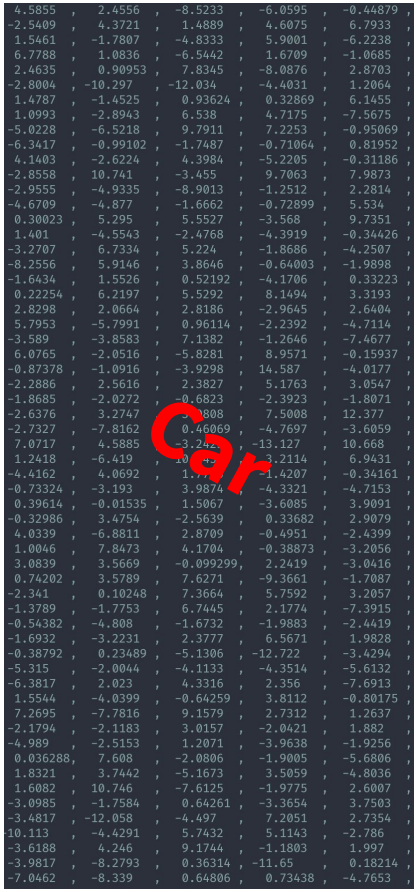
```
[ 0.20778 , -2.4151  ,  0.36605 ,  2.0139  , -0.23752 , -3.1952  ,
 -0.2952  ,  1.2272  , -3.4129  , -0.54969 ,  0.32634 , -1.0813  ,
  0.55626 ,  1.5195  ,  0.97797 , -3.1816  , -0.37207 , -0.86093 ,
  2.1509  , -4.0845  ,  0.035405,  3.5702  , -0.79413 , -1.7025  ,
 -1.6371  , -3.198   , -1.9387  ,  0.91166 ,  0.85409 ,  1.8039  ,
 -1.103   , -2.5274  ,  1.6365  , -0.82082 ,  1.0278  , -1.705   ,
  1.5511  , -0.95633 , -1.4702  , -1.865   , -0.19324 , -0.49123 ,
  2.2361  ,  2.2119  ,  3.6654  ,  1.7943  , -0.20601 ,  1.5483  ,
 -1.3964  , -0.50819 ,  2.1288  , -2.332   ,  1.3539  , -2.1917  ,
  1.8923  ,  0.28472 ,  0.54285 ,  1.2309  ,  0.26027 ,  1.9542  ,
  1.1739  , -0.40348 ,  3.2028  ,  0.75381 , -2.7179  , -1.3587  ,
 -1.1965  , -2.0923  ,  2.2855  , -0.3058  , -0.63174 ,  0.70083 ,
  0.16899 ,  1.2325  ,  0.97006 , -0.23356 , -2.094   , -1.737   ,
  3.6075  , -1.511   , -0.9135  ,  0.53878 ,  0.49268 ,  0.44751 ,
  0.6315  ,  1.4963  ,  4.1725  ,  2.1961  , -1.2409  ,  0.4214  ,
  2.9678  ,  1.841   ,  3.0133  , -4.4652  ,  0.96521 , -0.29787 ,
  4.3386  , -1.2527  , -1.7734  , -3.5637  , -0.20035 , -3.3013  ,
  0.99951 , -0.92888 , -0.94594 ,  1.5124  , -3.9385  ,  2.7935  ,
 -3.1042  ,  3.3382  ,  0.54513 , -0.37663 ,  2.5151  ,  0.51468 ,
 -0.88907 ,  1.011   ,  3.4705  , -3.6037  ,  1.3702  ,  2.3468  ,
  1.6674  ,  1.3904  , -2.8112  ,  2.237   , -1.0344  , -0.57164 ,
  1.0641  , -1.6919  ,  1.958   , -0.78305 ,  0.14741 ,  0.51083 ,
  1.8278  , -0.69638 ,  0.90548 ,  0.62282 , -1.8315  , -2.8587  ,
  0.48424 , -2.0527  , -0.53808 , -2.3472  ,  1.0354  , -1.8257  ,
 -0.3892  , -0.24943 ,  0.8651  , -1.5195  ,  1.2166  , -2.698   ,
 -0.96698 ,  2.2175  , -0.16089 , -0.49677 , -0.19646 ,  1.3284  ,
  4.0824  ,  1.3919  ,  0.80669 , -1.0316  , -0.28056 , -1.8632  ,
  0.47716 , -0.53628 ,  1.3853  , -2.1755  , -0.2354  ,  2.4933  ,
 -0.87255 ,  1.4493  , -0.10778 , -0.44159 ,  1.3462  ,  4.4211  ,
 -1.8385  ,  0.3985  ,  0.47637 , -0.60074 ,  3.3583  , -0.15006 ,
 -0.40495 ,  2.7225  , -1.6297  ,  0.86797 , -4.1445  , -2.7793  ,
  1.1535  , -0.011691,  0.9792  , -1.0141  ,  0.80134 ,  0.43642 ,
  1.4337  ,  2.8927  ,  0.82871 , -1.1827  , -1.3838  ,  2.3903  ,
 -0.89323 ,  1.1461  , -1.7435  ,  0.8654  , -0.27075 , -0.78698 ,
  1.5631  , -0.5923  ,  0.098082, -0.26682 ,  1.6282  , -0.77495 ,
  3.2552  ,  1.7964  , -1.4314  ,  1.2336  ,  2.3102  , -1.6328  ,
  2.8366  , -0.71384 ,  0.43967 ,  1.5627  ,  3.079   , -0.922   ,
 -0.43981 , -0.7659  ,  1.9362  , -2.2479  ,  1.041   ,  0.63206 ,
  1.5855  ,  3.4097  , -2.9204  , -1.4751  , -0.59534 , -1.688   ,
 -4.1362  ,  2.745   , -2.8515  ,  3.6509  , -0.66993 , -2.8794  ,
  2.0733  ,  1.1779  , -2.0307  ,  2.595   , -0.12246 ,  1.5844  ,
  1.1855  ,  0.022385, -2.2916  , -2.2684  , -2.7537  ,  0.34981 ,
 -4.6243  , -0.96521 , -1.1435  , -2.8894  , -0.12619 ,  2.9577  ,
 -1.7227  ,  0.24757 ,  1.2149  ,  3.5349  , -0.95802 ,  0.080346,
 -1.6553  , -0.6734  ,  2.2918  , -1.8229  , -1.1336  ,  1.8884  ,
  2.4789  , -0.66061 ,  2.0529  , -0.76687 ,  0.32362 , -2.2579  ,
  0.91278 ,  0.36231 ,  0.61562 , -0.15396 , -0.42917 , -0.89848 ,
  0.17298 , -0.76978 , -2.0222  , -1.7127  , -1.5632  ,  0.56631 ,
 -1.354   ,  2.6261  ,  1.9156  , -1.5651  ,  1.8315  , -1.4257  ,
 -1.6861  , -0.51953 ,  1.7635  , -0.50722 ,  1.388   , -1.1012  ],
```

```
[-7.5251e-01, -3.3480e+00, -2.9293e+00,  3.6773e+00,  6.7698e-01,
 -4.6221e+00,  1.7471e+00,  2.9072e+00, -1.1218e+00,  1.9050e+00,
  6.1861e+00, -1.5307e+00, -2.1315e-01, -4.9000e-01,  3.1568e+00,
 -3.2417e+00,  9.3068e-02, -1.6506e+00,  1.8947e+00, -3.6223e+00,
 -1.4505e+00,  2.8421e+00, -1.6908e+00, -4.7524e-01, -5.5192e+00,
 -1.5492e+00, -3.2481e+00,  4.3969e+00, -9.3570e-01,  1.1385e+00,
  2.3373e+00,  2.3882e+00, -1.5618e+00, -7.5315e-01, -5.9527e-01,
 -1.9020e+00,  1.0459e+00,  2.2420e-01, -9.4108e-01,  4.8844e-01,
 -2.2083e+00,  2.5857e+00,  2.2875e+00,  6.3069e-01,  3.4058e+00,
  2.0758e+00, -1.0663e+00, -8.4464e-01, -5.2534e-02, -7.9447e-01,
  3.0140e+00, -8.9454e-01,  2.1576e+00, -3.0407e+00,  1.3439e+00,
 -2.1920e-01, -6.2846e-02,  1.1478e+00,  2.8001e+00,  2.6444e+00,
  2.6262e+00,  2.2010e-02,  1.4596e+00, -1.1558e+00,  1.8789e-01,
  9.4600e-01, -2.9744e+00, -2.2531e+00,  7.7054e-01, -5.4315e-01,
 -2.2618e+00,  2.2210e+00, -1.2964e+00,  1.0105e+00,  5.8169e-01,
  3.5617e-01, -2.4568e-01, -2.0808e+00,  3.5410e+00, -5.2889e-01,
 -2.8393e-01,  4.8163e-01,  1.7635e+00,  7.4050e-01,  6.7875e-01,
 -2.2662e-01,  4.8440e+00,  8.9114e-01, -2.5486e+00, -6.9544e-01,
  6.2643e-01,  2.9279e-01,  3.1008e+00, -3.2464e+00,  1.5747e+00,
 -1.1993e+00,  3.0120e+00, -1.0923e+00, -1.1773e+00, -2.2735e+00,
 -1.2936e+00, -1.3023e+00,  1.0400e+00, -9.1724e-01,  1.0221e+00,
  8.9763e-01, -3.4229e+00,  2.7322e+00, -2.2374e-01,  4.8981e-01,
  5.6333e-01, -1.3467e+00,  9.6163e-01, -7.6177e-01, -3.1454e-01,
 -1.3337e+00,  2.9608e+00, -3.7193e+00, -1.1941e+00, -1.0349e+00,
 -2.5313e+00,  1.7521e+00, -1.5778e+00,  5.4771e-02,  4.6839e-01,
 -1.9399e+00, -1.3847e-01, -3.9830e+00,  4.9884e+00, -4.8193e-01,
 -1.1701e+00,  6.2994e-01,  2.9822e+00, -2.4728e-01, -2.6371e+00,
  1.0801e+00, -2.2238e+00, -5.2203e+00,  1.5099e+00, -1.8284e+00,
 -3.9196e-01,  1.7773e+00, -6.8698e-01, -1.0951e+00, -1.5319e+00,
  2.8311e+00,  2.9736e-01, -1.5198e+00,  1.3076e+00,  5.9841e-01,
 -5.2798e-01, -5.5499e-01, -3.4522e+00, -1.3115e+00, -1.0095e+00,
  3.3329e+00,  6.5440e+00, -1.3999e+00,  2.3499e+00, -2.4218e+00,
  6.9150e-01, -1.4240e+00,  3.3080e-01, -1.2254e+00,  3.7678e+00,
 -1.6502e+00, -1.6829e+00,  2.3409e-01,  8.3192e-01,  2.0174e+00,
 -2.6225e+00, -3.7696e-01, -2.1272e-01,  3.4416e+00, -3.6619e+00,
 -2.1298e+00,  9.7029e-01, -5.1133e-02,  8.2768e-01, -1.2364e+00,
  5.9028e-02, -5.1805e-01, -1.0821e+00, -1.7695e+00, -2.9489e-02,
 -2.5980e+00, -4.9045e-02, -6.9158e-01,  9.1347e-01, -4.6027e-01,
 -1.9500e+00,  2.0457e+00, -1.7526e+00,  2.7582e+00,  3.6836e-02,
 -2.3929e+00, -1.3635e+00,  2.1516e+00,  1.1975e+00, -1.2935e+00,
  1.4003e+00, -1.5616e+00, -7.1990e-01, -1.2839e-01,  1.5071e+00,
 -2.8197e+00, -5.9906e-01, -3.1609e-01,  4.8745e+00,  1.7453e+00,
  4.0927e+00, -5.4239e-01, -6.3825e-04,  3.3456e+00,  1.4135e+00,
  2.0070e-01,  1.8593e+00,  1.0568e+00, -2.4357e+00,  2.3165e+00,
  5.5872e-01, -1.6893e+00, -2.2931e+00,  1.6865e+00, -2.2543e+00,
 -1.6019e+00,  6.9584e-01,  2.5392e+00, -3.3192e-01,  3.2114e+00,
 -2.1623e+00, -9.7765e-01, -8.8937e-01, -5.1731e-01, -2.1990e+00,
  4.1397e+00,  4.2648e-01,  4.6854e+00,  1.0355e+00,  1.4013e+00,
 -1.0843e-01, -5.9694e-01, -4.0420e-01,  4.2305e+00, -5.2332e-01,
  4.1959e+00, -2.2805e-02, -6.3223e-01, -6.5072e-01, -1.7390e+00,
 -3.0675e+00,  3.9072e+00, -2.6396e+00, -2.4627e+00, -3.1164e-01,
 -2.5056e+00, -1.6382e+00,  3.2290e+00, -2.6652e+00, -7.3372e-01,
  9.6020e-01,  4.3881e+00, -1.2500e+00,  1.2498e+00,  1.9080e-01,
  1.9253e+00,  1.8284e+00, -2.3579e+00, -3.3646e+00,  6.8795e-01,
  1.2263e+00, -9.3136e-01,  5.5192e-01,  1.1171e+00, -2.8175e+00,
 -2.6307e+00,  1.4002e-01,  3.1652e-01, -5.7089e-01, -1.2883e+00,
  9.8610e-01, -1.0684e+00, -7.9920e-02, -2.6351e+00, -1.4276e+00,
 -5.3942e-01, -1.3570e+00, -6.0974e-01, -2.2030e+00,  2.0585e+00,
 -4.2257e-01, -1.5917e+00, -1.1557e+00, -2.8138e+00, -2.9554e+00,
```

```
 4.5855  ,  2.4556  , -8.5233  , -6.0595  , -0.44879 ,
-2.5409  ,  4.3721  ,  1.4889  ,  4.6075  ,  6.7933  ,
 1.5461  , -1.7807  , -4.8333  ,  5.9001  , -6.2238  ,
 6.7788  ,  1.0836  , -6.5442  ,  1.6709  , -1.0685  ,
 2.4635  ,  0.90953 ,  7.8345  , -8.0876  ,  2.8703  ,
-2.8004  , -10.297  , -12.034  , -4.4031  ,  1.2064  ,
 1.4787  , -1.4525  ,  0.93624 ,  0.32869 ,  6.1455  ,
 1.0993  , -2.8943  ,  6.538   ,  4.7175  , -7.5675  ,
-5.0228  , -6.5218  ,  9.7911  ,  7.2253  , -0.95069 ,
-6.3417  , -0.99102 , -1.7487  , -0.71064 ,  0.81952 ,
 4.1403  , -2.6224  ,  4.3984  , -5.2205  , -0.31186 ,
-2.8558  ,  10.741  , -3.455   ,  9.7063  ,  7.9873  ,
-2.9555  , -4.9335  , -8.9013  , -1.2512  ,  2.2814  ,
-4.6709  , -4.877   , -1.6662  , -0.72899 ,  5.534   ,
 0.30023 ,  5.295   ,  5.5527  , -3.568   ,  9.7351  ,
 1.401   , -4.5543  , -2.4768  , -4.3919  , -0.34426 ,
-3.2707  ,  6.7334  ,  5.224   , -1.8686  , -4.2507  ,
-8.2556  ,  5.9146  ,  3.8646  , -0.64003 , -1.9898  ,
-1.6434  ,  1.5526  ,  0.52192 , -4.1706  ,  0.33223 ,
 0.22254 ,  6.2197  ,  5.5292  ,  8.1494  ,  3.3193  ,
 2.8298  ,  2.0664  ,  2.8186  , -2.9645  ,  2.6404  ,
 5.7953  , -5.7991  ,  0.96114 , -2.2392  , -4.7114  ,
-3.589   , -3.8583  ,  7.1382  , -1.2646  , -7.4677  ,
 6.0765  , -2.0516  , -5.8281  ,  8.9571  , -0.15937 ,
-0.87378 , -1.0916  , -3.9298  ,  14.587  , -4.0177  ,
-2.2886  ,  2.5616  ,  2.3827  ,  5.1763  ,  3.0547  ,
-1.8685  , -2.0272  , -0.6823  , -2.3923  , -1.8071  ,
-2.6376  ,  3.2747  ,  3.0808  ,  7.5008  ,  12.377  ,
-2.7327  , -7.8162  ,  0.46069 , -4.7697  , -3.6059  ,
 7.0717  ,  4.5885  , -3.2425  , -13.127  ,  10.668  ,
 1.2418  , -6.419   ,  10.045  ,  3.2114  ,  6.9431  ,
-4.4162  ,  4.0692  ,  1.7755  , -1.4207  , -0.34161 ,
-0.73324 , -3.193   ,  3.9874  , -4.3321  , -4.7153  ,
 0.39614 , -0.01535 ,  1.5067  , -3.6085  ,  3.9091  ,
-0.32986 ,  3.4754  , -2.5639  ,  0.33682 ,  2.9079  ,
 4.0339  , -6.8811  ,  2.8709  , -0.4951  , -2.4399  ,
 1.0046  ,  7.8473  ,  4.1704  , -0.38873 , -3.2056  ,
 3.0839  ,  3.5669  , -0.099299,  2.2419  , -3.0416  ,
 0.74202 ,  3.5789  ,  7.6271  , -9.3661  , -1.7087  ,
-2.341   ,  0.10248 ,  7.3664  ,  5.7592  ,  3.2057  ,
-1.3789  , -1.7753  ,  6.7445  ,  2.1774  , -7.3915  ,
-0.54382 , -4.808   , -1.6732  , -1.9883  , -2.4419  ,
-1.6932  , -3.2231  ,  2.3777  ,  6.5671  ,  1.9828  ,
-0.38792 ,  0.23489 , -5.1306  , -12.722  , -3.4294  ,
-5.315   , -2.0044  , -4.1133  , -4.3514  , -5.6132  ,
-6.3817  ,  2.023   ,  4.3316  ,  2.356   , -7.6913  ,
 1.5544  , -4.0399  , -0.64259 ,  3.8112  , -0.80175 ,
 7.2695  , -7.7816  ,  9.1579  ,  2.7312  ,  1.2637  ,
-2.1794  , -2.1183  ,  3.0157  , -2.0421  ,  1.882   ,
-4.989   , -2.5153  ,  1.2071  , -3.9638  , -1.9256  ,
 0.036288,  7.608   , -2.0806  , -1.9005  , -5.6806  ,
 1.8321  ,  3.7442  , -5.1673  ,  3.5059  , -4.8036  ,
 1.6082  ,  10.746  , -7.6125  , -1.9775  ,  2.6007  ,
-3.0985  , -1.7584  ,  0.64261 , -3.3654  ,  3.7503  ,
-3.4817  , -12.058  , -4.497   ,  7.2051  ,  2.7354  ,
10.113   , -4.4291  ,  5.7432  ,  5.1143  , -2.786   ,
-3.6188  ,  4.246   ,  9.1744  , -1.1803  ,  1.997   ,
-3.9817  , -8.2793  ,  0.36314 , -11.65   ,  0.18214 ,
-7.0462  , -8.339   ,  0.64806 ,  0.73438 , -4.7653  ,
```


Banana


Chocolate


Car

Word Embeddings

- Definition: a dense vector (often 300 dimensions) capturing semantics
- Recent developments: Contextualized word embeddings leading to cutting-edge models like BERT and GPT2 (see session on Friday)
- We cover the basics today

Credit to Bastiaan Bruinsma; his course materials were a great ressource to build this set on embeddings

The Idea: Distributional Hypothesis

“In most cases, the meaning of a word is its use” (Wittgenstein, 1953)
“Difference of meaning correlates with difference of distribution” (Harris, 1954)
“A word is characterized by the company it keeps” (Firth, 1957)
“Words which are similar in meaning occur in similar contexts” (Rubenstein & Goodenough, 1965)

Distributional Hypothesis: words that are used in similar ways, surrounded by similar words, are likely to have similar semantic meanings.

Learning Meaning from Data

- The **narmel** met weekly to discuss neighborhood issues
- After the protest, several **narmels** were formed across the city
- Each **narmel** had a rotating facilitator and made decisions by consensus

What is a narmel?

Learning Meaning from Data

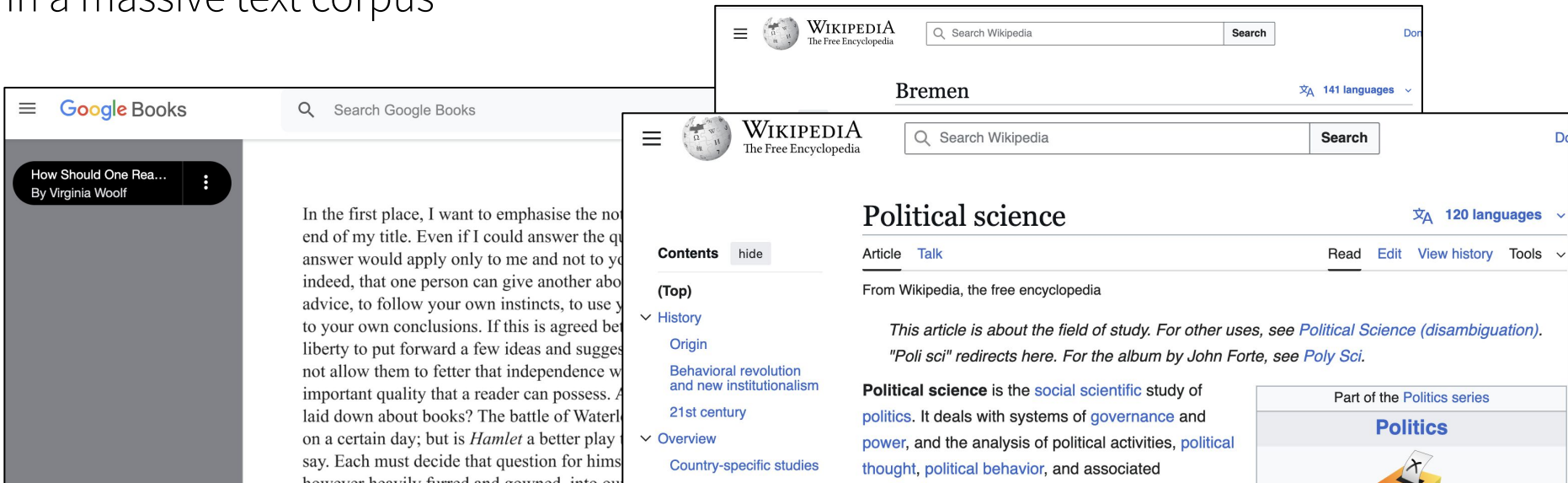
- The **narmel** met weekly to discuss neighborhood issues
- After the protest, several **narmels** were formed across the city
- Each **narmel** had a rotating facilitator and made decisions by consensus

What is a narmel? From *weekly meetings*, *neighborhood issues*, *formed after protests*, *rotating facilitator*, and *decisions by consensus*, we can infer that a narmel is likely a community group, grassroots organization, or local council.

Learning Meaning from Data

We don’t need to define meaning explicitly

Design an algorithm that learns a vector for each word based on its neighbours in a massive text corpus



Popular Word Embedding Models

Static models (One vector per word — context-independent)

Word2Vec (Mikolov, 2013)

- Type: Prediction-based (local context)
- Architectures: Skip-gram, CBOW
- Optimization: Negative Sampling, Hierarchical Softmax

GloVe (Pennington et al., 2014)

- Type: Count-based (global statistics)
- Method: Co-occurrence matrix construction, Weighted matrix factorization

Word2Vec: How it works

- Trains a shallow neural network
- The learned weights of the hidden (projection) layer become the word embeddings
- Self-supervised learning (No manual labels, predict from context)
- Two models to learn word embeddings: CBOW and Skip-Gram

CBOW

(Context → Target)

Input: [the, cat, on, the]
Output: predict → [mat]

Skip-Gram

Skip-Gram (Target → Context)

Input: [mat]
Output: predict → [the, cat, on, the]

Word2Vec: How it works

(Window size = 1)

The government passed new climate legislation

The government passed new climate legislation

Training examples generated:

- Skip-Gram: Given ‘climate’, predict ‘new’ and ‘legislation’
- CBOW: Given (‘new’, ‘legislation’), predict ‘climate’

To get good at these predictions across the entire corpus, the model must learn that words like ‘climate’, ‘environment’, and ‘emissions’ need to have similar vector representations — because they often appear in similar political contexts elsewhere, such as with ‘policy’, ‘government’, and ‘regulation’.

Further Reading: Illustrated Word2Vec

- Step-by-step explanation of CBOW and Skip-Gram
- Clear and visually engaging breakdown of the math and intuition

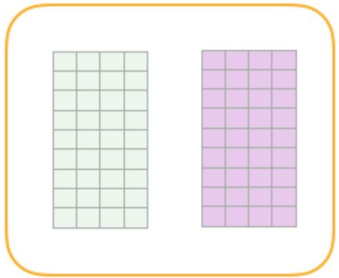
“The Illustrated Word2Vec” by Jay Alammar

https://jalammar.github.io/illustrated-word2vec/

dataset

input word	output word	target
not	thou	1
not	aaron	0
not	taco	0
not	shalt	1
not	mango	0
not	finglonger	0
not	make	1
not	plumbus	0
...	...	...

model



GloVe

- Word2Vec uses only local context windows.
- GloVe asks: *Can we use statistics from the entire corpus?*

GloVe builds a co-occurrence matrix:
→ *How often does each word appear near every other word in the corpus?*

It then learns word vectors that reflect these global relationships

GloVe: Building the Co-occurrence Matrix

Corpus:

1. I like deep learning.
2. I like NLP.
3. I enjoy learning NLP.

Co-occurrence Matrix X:

	I	like	deep	learning	enjoy	NLP
I	0	2	1	1	1	2
like	2	0	1	1	0	1
deep	1	1	0	1	0	0
learning	1	1	1	0	1	1
enjoy	1	0	0	1	0	1
NLP	2	1	0	1	1	0

Notes: Values represent how many times the words co-occur within a context window.

GloVe: Building the Co-occurrence Matrix

Corpus:

1. I like deep learning.
2. I like NLP.
3. I enjoy learning NLP.

Goal:
Learn word vectors where "like" and "enjoy" are closer to each other than to "deep", because they share similar context words (e.g., "I" and "learning").

Co-occurrence Matrix X:

	I	like	deep	learning	enjoy	NLP
I	0	2	1	1	1	2
like	2	0	1	1	0	1
deep	1	1	0	1	0	0
learning	1	1	1	0	1	1
enjoy	1	0	0	1	0	1
NLP	2	1	0	1	1	0

Notes: Values represent how many times the words co-occur within a context window.

Further Reading: Illustrated Word2Vec

Co-occurrence Matrix X:

	word1	word2	...
word1	X_{11}	X_{12}	...
word2	X_{21}	X_{22}	...
...	...	...	...

Word2Vec vs. GloVe

- Word2Vec learns embeddings by predicting context.
- GloVe derives embeddings from word co-occurrence statistics.

- Skip-Gram can better capture rare word semantics.
- CBOW is faster and better for frequent words.

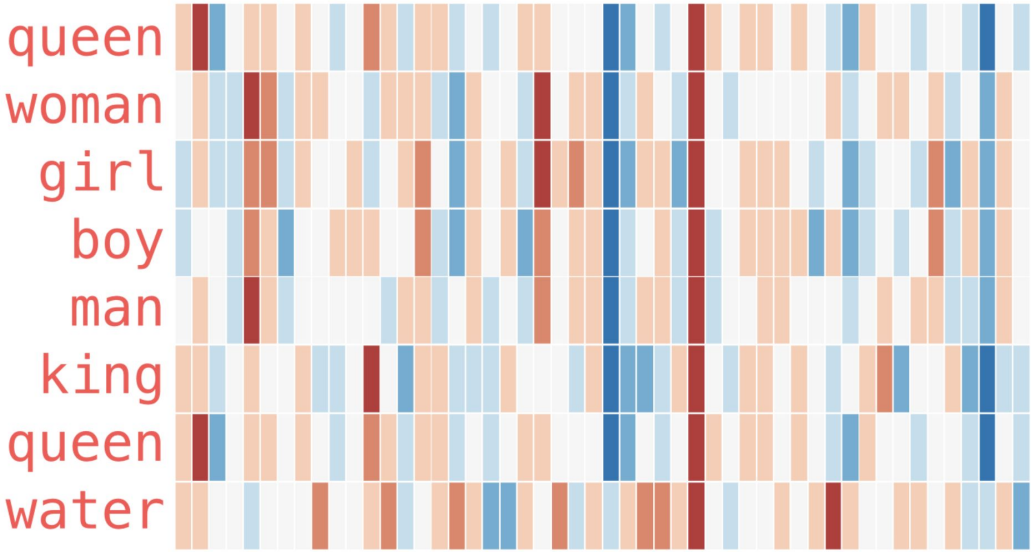
Illustrate Word2Vec

This is a word embedding for the word “king” (GloVe vector trained on Wikipedia):

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 ,
-0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961
, -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 ,
-0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 ,
-1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

https://jalammar.github.io/illustrated-word2vec/

Illustrate Word2Vec

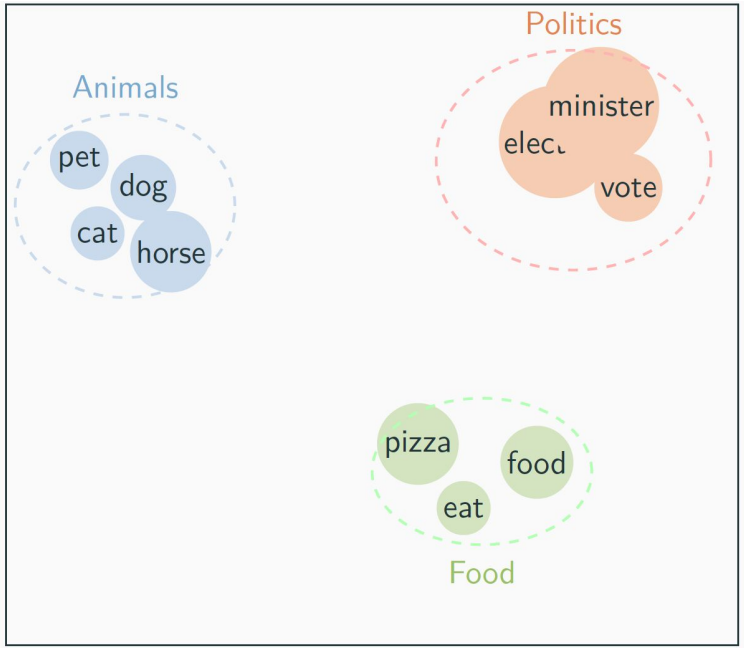


https://jalammar.github.io/illustrated-word2vec/

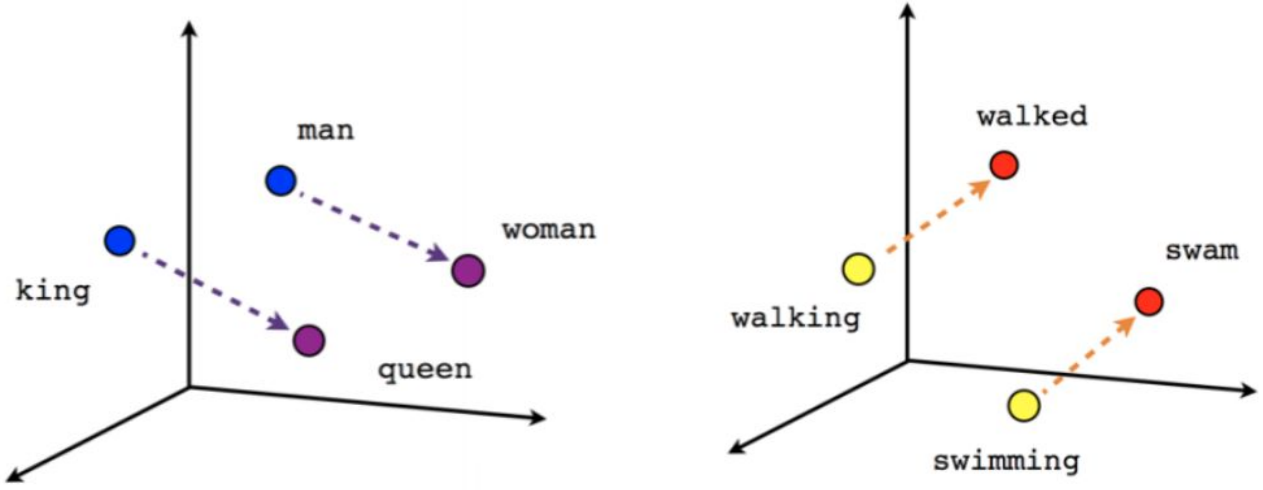
Visualizing Embeddings: 2D

How to visualize a 300-dimensional space?

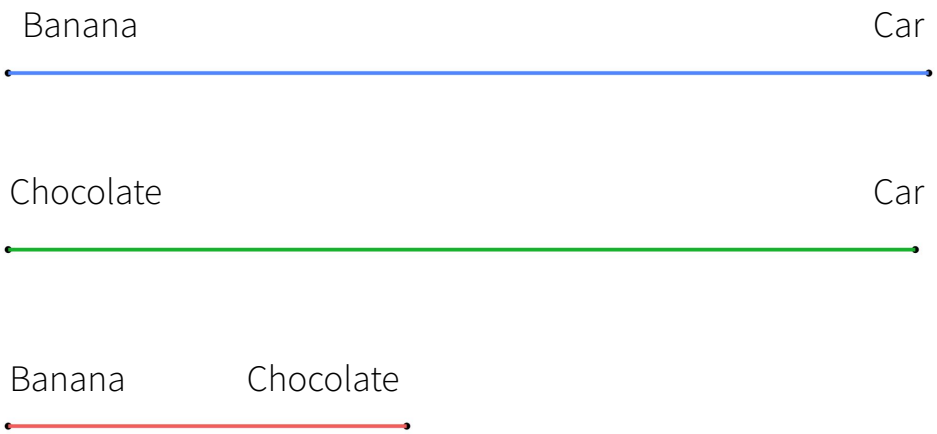
Dimensionality reduction techniques like t-SNE or UMAP project these vectors into 2D



Visualizing Embeddings: 3D



Visualizing Embeddings: 1D



Using Embeddings representation

On its own:

- Helps understand and explore the corpus
- Synonym detection, word analogies, bias detection

Combined with other methods:

- Produces a dense embedding that feeds into:
  - Document classification
  - Topic modeling
  - Similarity analysis

Application: Bias Detection

Extreme *she* occupations		
1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

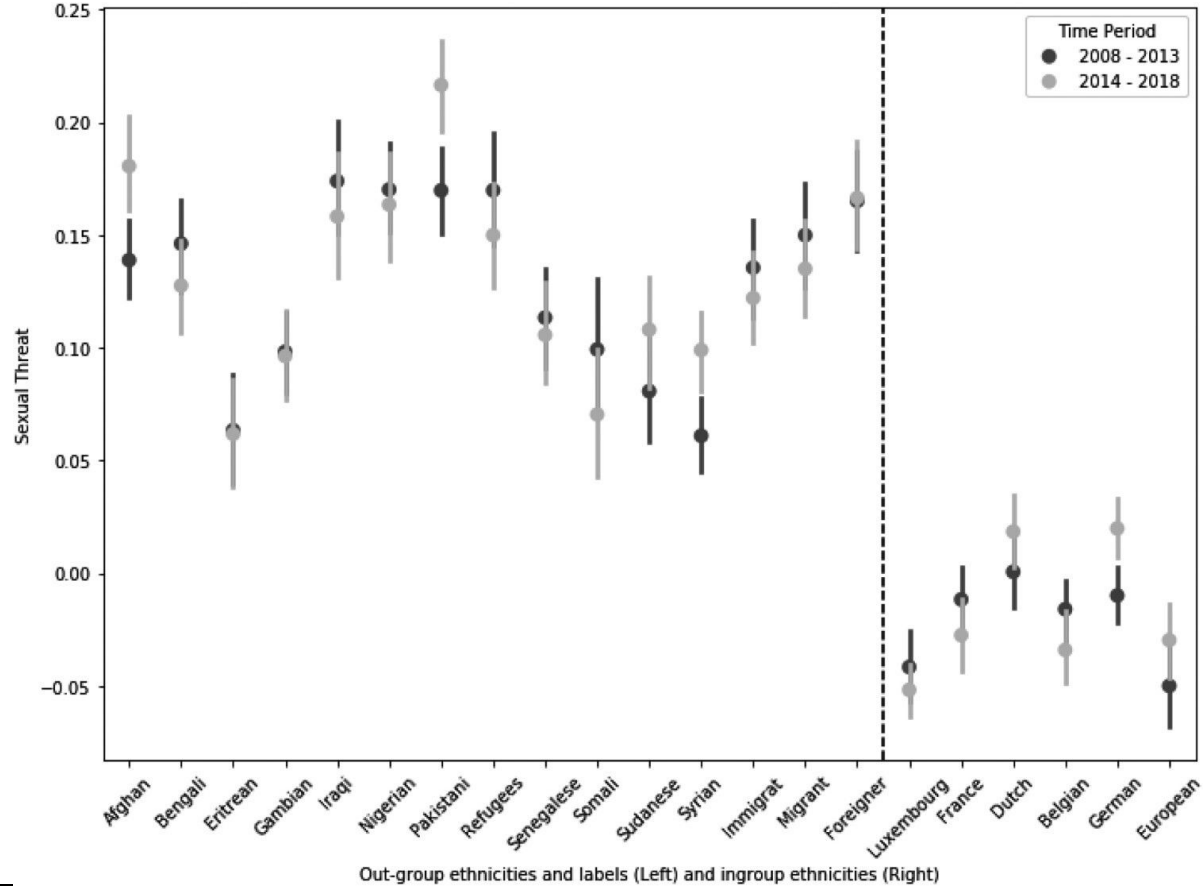
Extreme *he* occupations		
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

Figure 1: The most extreme occupations as projected on to the *she*−*he* gender direction on g2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*..

Application: Bias Detection

- 2 million Dutch news articles
- Ethnic outgroups are associated more strongly with sexual threat than ethnic ingroups.



Kroon, A. C., & van der Meer, T. G. L. A. (2023). Guilty by Association: Using Word Embeddings to Measure Ethnic Stereotypes in News Coverage. Journalism Practice, 17, 319–335.

Application: Sentiment Analysis

Embeddings were used as part of one of the sentiment approaches compared: CNN

				Positive			Neutral			Negative		
section	**name**	**Acc.**	**alpha**	**Pr.**	**Re.**	**F1**	**Pr.**	**Re.**	**F1**	**Pr.**	**Re.**	**F1**
Manual Coding	Single Coder	0.82	0.82	0.88	0.86	0.87	0.76	0.81	0.78	0.84	0.80	0.82
Manual Coding	Vote (3 Coders)	0.88	0.90	0.97	0.91	0.94	0.82	0.88	0.85	0.87	0.84	0.86
Crowd-Coding	Single Coder	0.72	0.75	0.69	0.84	0.76	0.69	0.58	0.63	0.78	0.78	0.78
Crowd-Coding	Vote (3 Coders)	0.77	0.81	0.73	0.89	0.80	0.74	0.65	0.69	0.83	0.81	0.82
Crowd-Coding	Vote (5 Coders)	0.77	0.81	0.73	0.90	0.81	0.73	0.65	0.69	0.84	0.80	0.82
Machine Learning	CNN	0.63	0.50	0.68	0.49	0.56	0.58	0.78	0.66	0.72	0.57	0.63
Machine Learning	NB	0.58	0.39	0.74	0.34	0.47	0.52	0.83	0.64	0.65	0.47	0.55
Machine Learning	SVM	0.57	0.41	0.69	0.37	0.48	0.52	0.79	0.62	0.64	0.48	0.55
Dictionaries	DANEW	0.42	0.10	0.75	0.08	0.15	0.40	0.97	0.57	0.80	0.04	0.08
Dictionaries	DamstraBoukes	0.41	0.05	0.83	0.07	0.13	0.40	0.99	0.57	0.00	0.00	0.00
Dictionaries	Muddiman	0.49	0.31	0.53	0.38	0.44	0.46	0.64	0.53	0.53	0.39	0.45
Dictionaries	NRC	0.47	0.32	0.39	0.53	0.45	0.46	0.44	0.45	0.59	0.46	0.52

Van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, *15*(2), 121-140.

Outlook

Limitations of Static Embeddings (Word2Vec, GloVe)

- Out-of-Vocabulary: Cannot handle new words like "metaverse" or misspellings like "enviroment".
- Context Independence: "light" has the same vector in “light reading” and “light weight”.

Contextual Embeddings (BERT, Transformers): generate different embeddings for the same word in different contexts

Summary: Word embeddings

- A dense vector capturing semantics
- Word2Vec, GloVe
- Visualization
- Applications

Coding session:

Jointly:

Word2vec and GloVe embeddings.RmD

Get script:

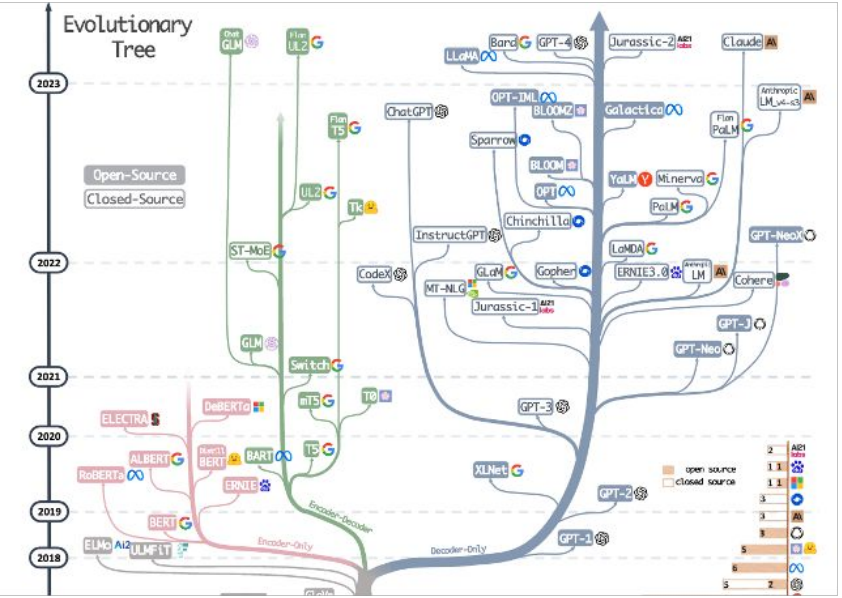
https://github.com/fabiennelind/text-as-data-in-R

Wrap up

Central decision criteria

- Approach to measurement (discovery or finding pre-defined concept)?
- Availability of methods (dictionaries, labeled data, pretrained models)
- Time (and Importance for a specific RQ)
- Budget (Modeling, Validation)
- Skills (R, Python, Patience)
- Ethical and environmental considerations
- …

Conclusion



Yang et al., 2023, p. 3

- The ways that we conduct text analysis is changing drastically, but questions that we ask about validation do not
- Social scientists with computational training can help to ask the interesting questions, curate datasets, can ensure valid implementation of methods, and interpret the results

Thank you for joining