

# Advanced quantitative text analysis (2023W)

Petro Tolochko, Fabienne Lind





# Introduction

---

# About us



## Petro Tolochko

- Post-Doc, CCL, University of Vienna
- Research focus: Text complexity, social networks
- [petro.tolochko@univie.ac.at](mailto:petro.tolochko@univie.ac.at)

## Fabienne Lind

- Post-Doc, CCL, University of Vienna
- Research focus: Multilingual automated text analysis
- [fabienne.lind@univie.ac.at](mailto:fabienne.lind@univie.ac.at)



## Your turn :)

- Name
  - Affiliation? Background?
  - Experience with content analysis, R & Python
  - What are the expectations and wishes for the workshop and the workshop leaders?
-

# Course objectives

- Getting to know (basic) procedures of automated text analysis
- Insight into practical challenges
- Critical reflection on the method and its results
- Inspiration for your own projects
- Make an informed decision about a suitable method for a given application scenario

# Contents (smaller changes possible)

Day	Session 1	Session 2
1	Text as data, Text representation	Feature Engineering
2	Concepts & Data	Dictionaries
3	Supervised machine learning	Unsupervised machine learning
4	Neural network models, transformers	Using LLMs
5	Multilingual text analysis, Wrap-up	Project talks

# Course philosophie

Topics are covered with

- Lecture style input
- Guided coded session
- Coding challenge

Interrupt us, ask all kinds of questions

Work on your own data, code, and projects

---

# Course assessment

Participation in class (20%)

Final paper: application of one or several automated text analysis methods on a topic related to the PhD thesis or a topic of free choice (80%)

- Contents: short motivation, analysis (commented code), description and interpretation of results (about 10 pages)
- Format: R Markdown
- Deadline: February 15th, 2024

## Receive feedback on your projects

Very informal opportunity to talk about your text analysis use case and (initial) design (plan)

- Research question, Data, Methods, Current struggles

And to receive some feedback (no grades, points, etc. just free brainstorming opportunity)

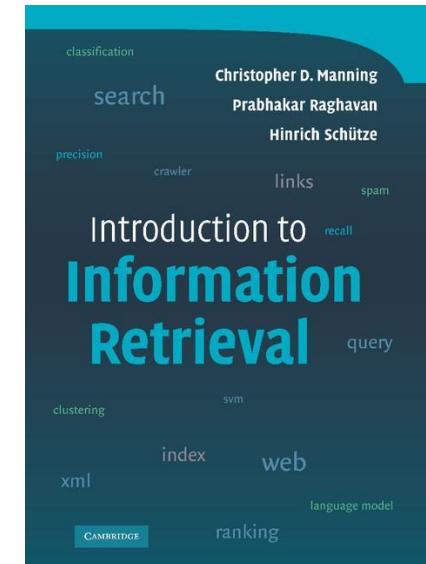
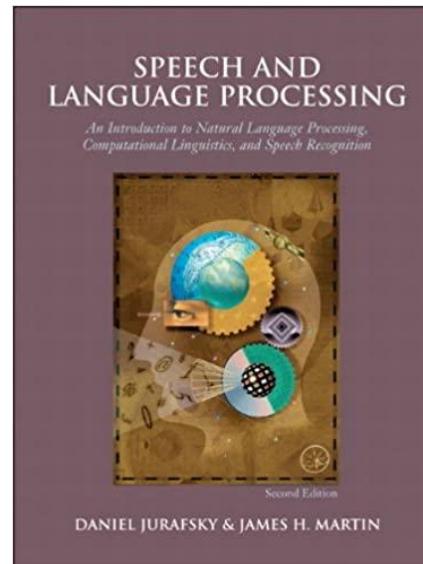
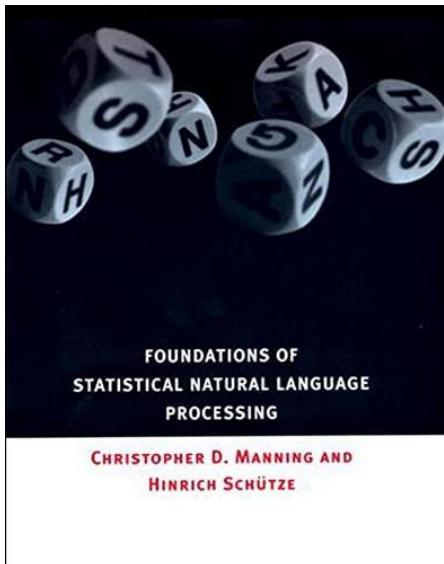
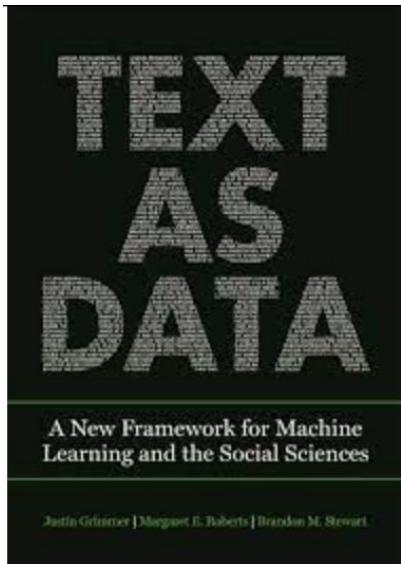
Can also be useful to get initial feedback for your final paper in this course.

**When:** Friday

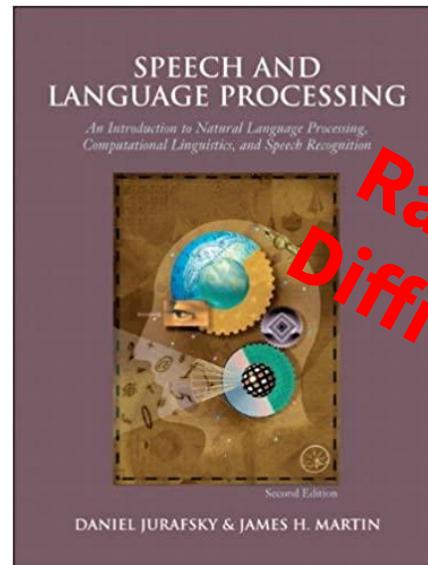
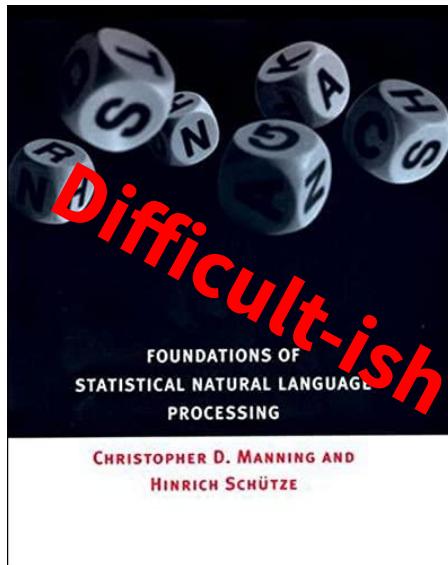
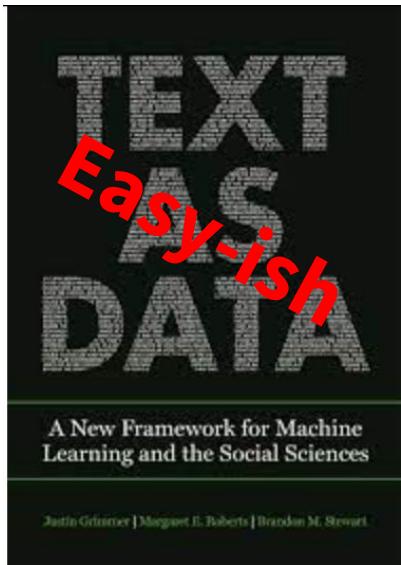
# Course repository: GitHub

<https://github.com/PeterTolochko/text-as-data-2024>

# Resources



# Resources



## R and Python

- Both are free and open source
- Most things can be done in both languages
- Great online support resources

## Motivations to use R and RStudio



- Great for visualization and statistical analysis
- Popular in the social science community
- A full programming language and thus a good start to pick up other programming languages

# Motivations to use Python

- Speed, simple syntax
- General-purpose programming language, not limited to the data science community
- Certain text analysis approaches (e.g., deep learning) are more accessible in Python

## Some R Ressources

- R for Data Science <https://r4ds.had.co.nz/>
- Tidy Modeling with R. <https://www.tmwr.org/>
- Big book of R. Database of R resources. <https://www.bigbookofr.com/>

Twitter accounts: @WeAreRLadies

---

## Some Python Ressources

- Learn Python for Everybody:  
<https://www.youtube.com/watch?v=8DvywoWv6fl>
- Data Analysis with Python:  
<https://www.youtube.com/watch?v=GPVsHOIRBBI>
- VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools  
<https://github.com/jakevdp/PythonDataScienceHandbook>

# R and Python Ressources for Text Analysis

- Atteveldt, W., Trilling, D., & Arcila, C. (2021). Computational Analysis of Communication. Wiley-Blackwell.
  - URL: <http://cssbook.net>
  - Code examples for R and Python
- NLP Course:  
<https://huggingface.co/learn/nlp-course/chapter0/1>

# Today

9:45-10:40	Orga & Intro & Coding
10:40-10:50	Coffee Break
10:50-11:45	Input & Coding
11:45-12:45	Lunch
12:45-14:14	Input & Coding

# Text as Data

Day 1 Session 1

---

## Motivations to analyse text

- Huge volumes of digital available information
- Measure contents of texts
- Traces of social behaviour
- Explain, understand, predict feelings, attitudes and behaviour of individuals, groups, societies

# Field experiments

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES | 

## Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer , Jamie E. Guillory, and Jeffrey T. Hancock [Authors Info & Affiliations](#)

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

June 2, 2014 | 111 (24) 8788-8790 | <https://doi.org/10.1073/pnas.1320040111>

THIS ARTICLE HAS BEEN CORRECTED +

 1,819,531  1,356

### Significance

We show, via a massive ( $N = 689,003$ ) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.

# Effects of social media communication

New Media & Society  
OnlineFirst  
© The Author(s) 2022. Article Reuse Guidelines  
<https://doi.org.uaccess.univie.ac.at/10.1177/14614448221088970>



## Article

### The relationship between humanitarian NGO communication and user engagement on Twitter

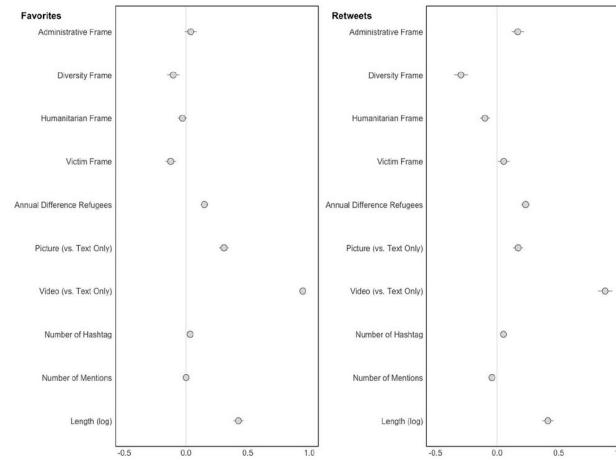
Daniela Dimitrova <sup>1</sup>, Tobias Heidenreich <sup>2</sup>, and Teodor Antonio Georgiev<sup>3</sup>

#### Abstract

*One of the few actors whose mission is to provide support and advocacy for refugee communities with limited access to information and services are humanitarian non-governmental organizations (NGOs). This study examines the narratives produced by the leading humanitarian NGOs on one of the most popular social media platforms today—namely, Twitter. The study investigates which narratives are most popular among global NGOs and whether the way they frame the refugee issue is related to Twitter engagement. The findings contribute to scholarship on online communication and user engagement, and also inform humanitarian NGO practices and policy discussions regarding media and migration.*

#### Keywords

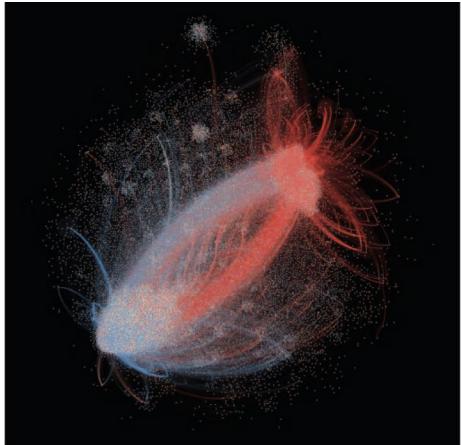
Message framing, migration, NGO communication, refugees, Twitter, user engagement



**Figure 2.** Estimated posterior fixed-effects parameters.

*Note.* Estimated posterior fixed-effects parameters for the models considering the number of favorites (left) and retweets (right). Thin lines represent 95% confidence intervals (CIs).

# Polarization



**2012 Presidential Election**



**2013 Super Bowl**

Dark red = conservative  
dark blue = liberal

Barberá 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.

# Policy diffusion

AJPS AMERICAN JOURNAL  
of POLITICAL SCIENCE

## Policy Diffusion: The Issue-Definition Stage

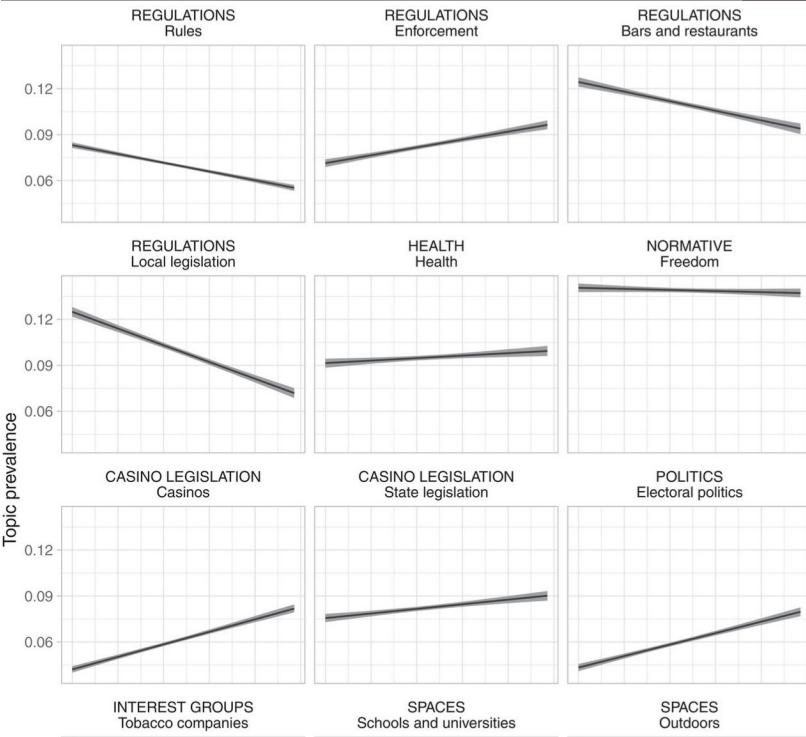


**Fabrizio Gilardi** University of Zurich  
**Charles R. Shipan** University of Michigan  
**Bruno Wüest** Forschungsstelle sotomo

**Abstract:** We put forward a new approach to studying issue definition within the context of policy diffusion. Most studies of policy diffusion—which is the process by which policymaking in one government affects policymaking in other governments—have focused on policy adoptions. We shift the focus to an important but neglected aspect of this process: the issue-definition stage. We use topic models to estimate how policies are framed during this stage and how these frames are predicted by prior policy adoptions. Focusing on smoking restriction in U.S. states, our analysis draws upon an original data set of over 52,000 paragraphs from newspapers covering 49 states between 1996 and 2013. We find that frames regarding the policy's concrete implications are predicted by prior adoptions in other states, whereas frames regarding its normative justifications are not. Our approach and findings open the way for a new perspective to studying policy diffusion in many different areas.

**Verification Materials:** The data and materials required to verify the computational reproducibility of the results, procedures, and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at <https://doi.org/10.7910/DVN/QEMNP1>.

Gilardi, F., Shipan, C. R., & Wüest, B. (2021). Policy diffusion: The issue-definition stage. *American Journal of Political Science*, 65(1), 21-35.



# Motivations to analyse text automatically

- Impossible to read all texts for large corpora



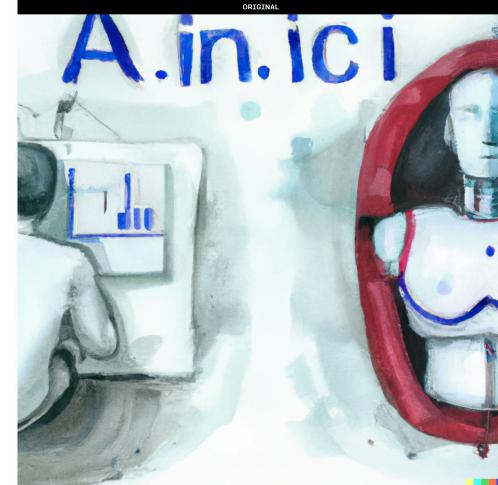
<https://www.mittelbayerische.de/junge-leser/klasse-informiert-nachrichten/warum-wir-heute-zeitung-lese-n-koennen-24440-art1260807.html>

# Purpose of obtaining measures for a large number of documents

- Filter options
  - If measurements are available for a large number of data, other filtering options are possible. E.g.; linkage studies combination with media usage data
- Evidence-based policy making
  - Making the opinion of populations visible, holding politicians accountable

# 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



# Some major challenges when working with large corpora

Big data, big bias?

The end of theory?

Generalizing from online to offline behavior

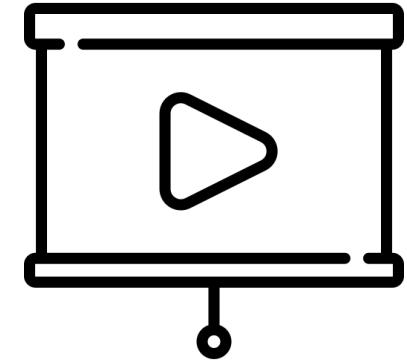
Ethical concerns

## Potential applications scenarios

- Measure contents of large naturally occurring text collections
  - As standalone interest (descriptive interests)
  - As input for another interest (inference interests)
- Explore the texts that survey participants type into open answer questions
- Etc.

## Exercise: First text analysis project idea (15min)

- Let's form groups
- Jointly discuss about initial text analysis projects ideas
  - You can briefly talk about project ideas that you already have or about areas in your discipline where large text quantities play a role



by Freepik - Flaticon

# Text Representation

# What is text?

- Data
- Unstructured
- Multidimensional (Highly)

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

---

# What is text?

- So, how do we make computers process our language?

# 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”
  - Different research questions
  - Different ways to think about what text is

# Two approaches to text representation

- Discrete text representation
- Distributed text representation

## 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 ice-cream”

Document 2: “John loves oranges”

Document 3: “Marry hates ice-cream”

**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

$$N = 3$$

$$K = 6$$

$$X = 3 \times 6 \text{ matrix}$$

# 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)

---

# Bag-of-Words representation

Representation of text as a bag of words

- Collection of words
- Ordering is unimportant
- Each text is represented as a **count** of words contained in it

## 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

# Multinomial Model of Language

You can think of a text as a draw from a Multinomial distribution

# Binomial

$$\binom{n}{k} p^k (1-p)^{n-k}$$

# Binomial

n = number of events  
k = number of successes

$$\binom{n}{k} p^k (1 - p)^{n-k}$$

# Multinomial

$$\frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

# Multinomial

$$\frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

n = text length

k = size of vocabulary

p = probability of a word

## 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"

---

## Example Texts

Text\_1 = "banana banana banana banana chocolate"

Text\_3 = "chocolate chocolate chocolate banana fudge"

John

Text\_2 = "banana banana"

Text\_4 = "icecream icecream fudge ice-cream"

Text\_5 = "fudge fudge fudge"

Text\_6 = "ice-cream ice-cream fudge fudge"

## Example Texts

Text\_1 = "banana banana banana banana chocolate"

Text\_3 = "chocolate chocolate chocolate banana fudge"

Text\_2 = "banana banana"

John

Text\_4 = "icecream icecream fudge ice-cream"

Text\_5 = "fudge fudge fudge"

Marry

Text\_6 = "ice-cream ice-cream fudge fudge"

## 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

## Document-Term Matrix

banana	chocolate	fudge	icecream
7	4	1	0

John rates

banana	chocolate	fudge	icecream
1	3	7	5

Marry rates

# John Language Model

## Document-Term Matrix

banana	chocolate	fudge	icecream
7	4	1	0

John rates

banana	chocolate	fudge	icecream
1	3	7	5

Marry rates

# Marry Language Model

## New Texts

```
new_text_1 = "ice-cream fudge fudge"
```

```
new_text_2 = "chocolate chocolate banana banana"
```

## New Texts

```
new_text_1 = "ice-cream fudge fudge"
```

```
new_text_2 = "chocolate chocolate banana banana"
```

What's the probability that they have been generated by John or Marry?

---

## New Texts

banana	chocolate	fudge	icecream
0	0	2	1
2	2	0	0

## Probability Spoken by John

new\_text\_1 = "ice-cream fudge fudge"

$$Pr(b = 0, ch = 0, f = 2, i = 1) = \frac{3!}{0!0!2!1!} 0.583^0 \times 0.333^0 \times 0.08^2 \times 0^1 = 0$$

## Probability Spoken by John

new\_text\_2 = "chocolate chocolate banana banana"

$$Pr(b = 2, ch = 2, f = 0, i = 0) = \frac{4!}{2!2!0!0!} 0.583^2 \times 0.333^2 \times 0.08^0 \times 0^0 = 0.23$$

## Probability Spoken by Marry

new\_text\_1 = "ice-cream fudge fudge"

P = 0.18

new\_text\_2 = "chocolate chocolate banana banana"

P = 0.0008

---

# Vector Space Representation

Representation of texts as **vectors** in a multidimensional **space**

# Multidimensional?

- Position on a map:
  - **X** = Longitude **Y** = Latitude
- Position in a real world:
  - **X** = Longitude **Y** = Latitude **Z** = Height
- Point in time and space
  - **X** = Longitude **Y** = Latitude **Z** = Height **T** = Time

# Multidimensional?

Number of dimensions is the ***number of data points*** needed to describe an object in space

- ***Coordinates*** in the context of ***geographical position***
- ***Words*** in the context of position of text within a ***linguistic space***

## Two-dimensional space

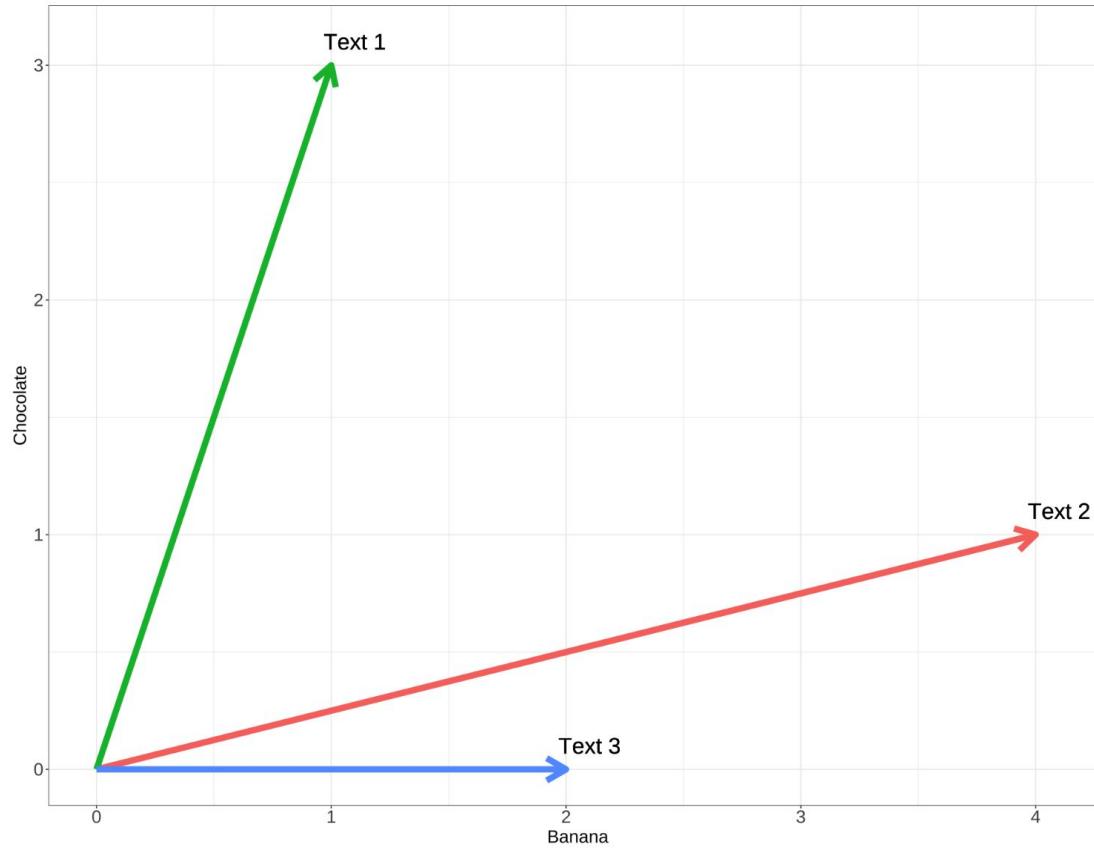
vocabulary = {"banana", "chocolate"}

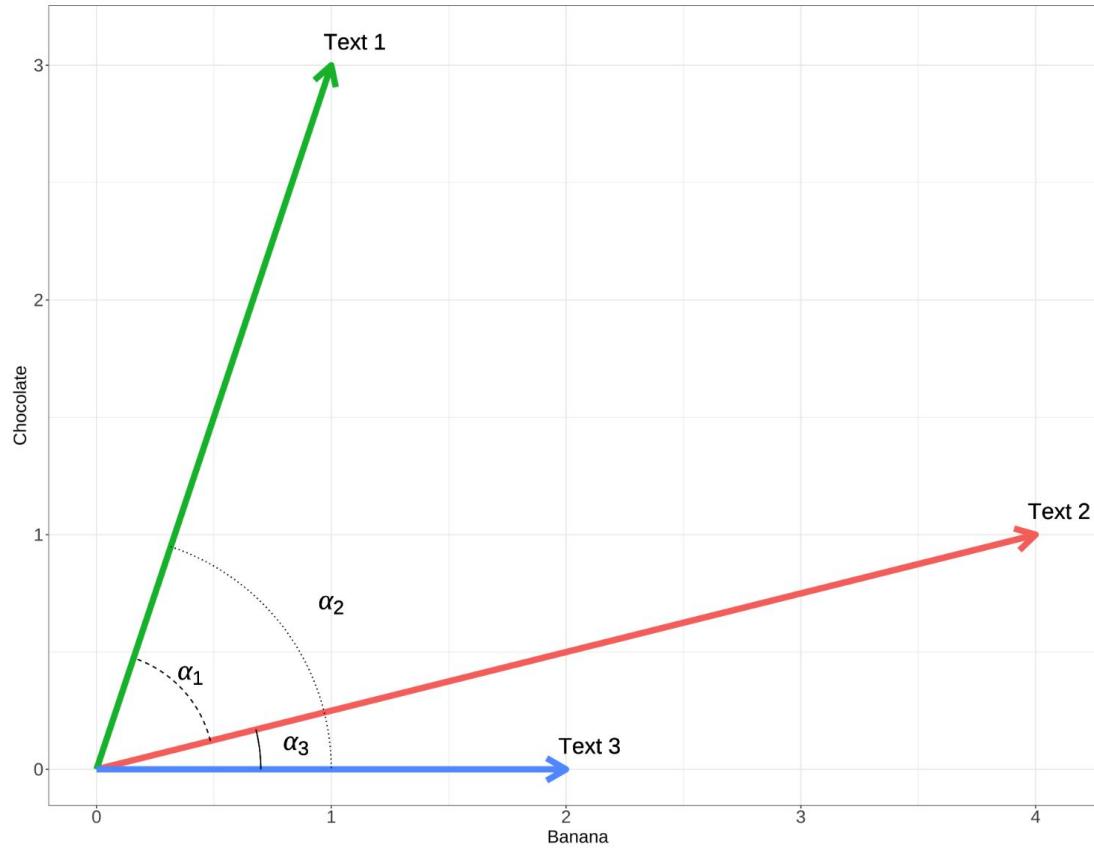
Text 1 = "chocolate, chocolate, chocolate, banana"

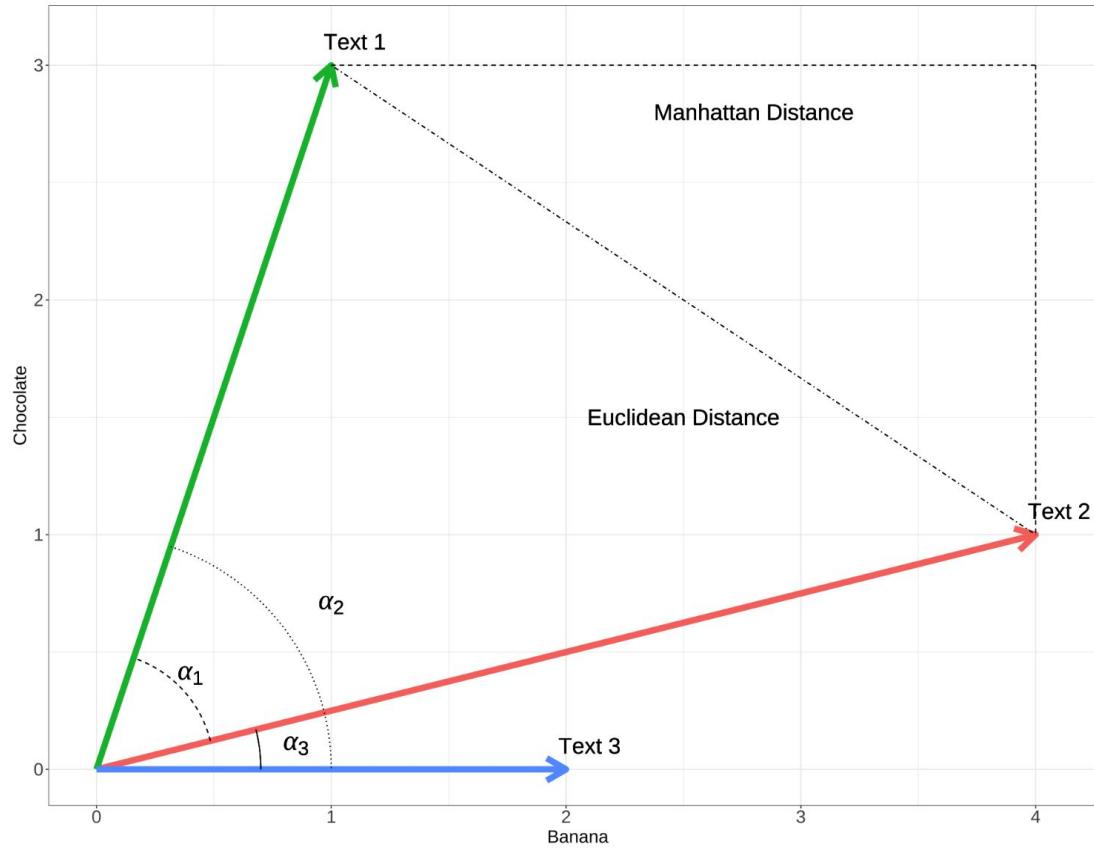
Text 2 = "banana, banana, banana, banana, chocolate"

Text 3 = "banana, banana"

---







# Multidimensional?

Estimated ~1,022,000 words in English language

**~ 1,022,000-dimensional space**



[adult swim]

Same math, though

# Discrete representation

Summary: Each word is considered unique & independent from each other

Main advantages:

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

Main disadvantage:

- For corpora with large corpora, memory constraints
  - Positional information of the word is not captured
  - Semantics is not captured
-

## Distributed representation

- aims to capture the semantic meaning and contextual information of words or sentences
- the information about a word is distributed along the vector it is represented as
- The fundamental concept that forms the basis for distributed representation = **the distributional hypothesis**

# Distributional Hypothesis / Word Embeddings

***The*** new thing kids do...

# Distributional Hypothesis / Word Embeddings

**The** new thing kids do... at least since '57...

## Distributional Hypothesis

- difference of meaning correlates with difference of distribution (Harris, 1954)
- In most cases, the meaning of a word is its use (Wittgenstein, 1953)
- A word is characterized by the company it keeps (Firth, 1957)
- Words which are similar in meaning occur in similar contexts (Rubenstein & Goodenough, 1965)
- Words with similar meanings will occur with similar neighbors if enough text material is available (Schütze & Pedersen, 1995)
- Or: words that are used in similar ways, surrounded by similar words, are likely to have similar semantic meanings.

## Distributional Hypothesis / Word Embeddings

Each word is *embedded* in an n-dimensional vector space (typically, 300 dimensions)

Neural network that “learns” the context of a word by looking at the words *around* it

**Word2Vec** (Mikolov, 2013) – one of the most famous methodologies, developed by Google engineer

---

[	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	,	0.9166	,	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	,	-0.20527	,	-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	,	
-1.8835	,	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	,	-1.1385e+00	,	1.393e+00	,		
2.3373e+00	,	2.3882e+00	,	-1.5618e+00	,	-7.5315e-01	,	-5.9527e-01	,		
-1.9020e+00	,	2.2420e-01	,	9.4108e-01	,	4.8844e-01	,	1.4787e-01	,		
-2.2083e+00	,	1.0549e+00	,	2.0587e+00	,	6.3069e-01	,	3.4058e+00	,		
2.0758e+00	,	-1.0663e+00	,	-8.4464e-01	,	-5.2534e-01	,	-7.9447e-01	,		
3.0140e+00	,	-8.9454e-01	,	2.1576e+00	,	-3.0407e+00	,	1.3439e+00	,		
-2.1920e+00	,	-2.8486e-01	,	1.1748e+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+00	,	4.8440e+00	,	8.9114e-01	,	-2.5486e+00	,	-6.9544e-01	,		
6.2643e+00	,	2.9279e-01	,	3.1008e+00	,	-3.2464e+00	,	1.5747e+00	,		
-1.1939e+00	,	3.0120e+00	,	-0.1923e+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	,	-2.3737e+00	,	4.8981e-01	,		
5.6333e-01	,	-1.3467e+00	,	9.6163e-01	,	-6.1717e-01	,	-1.3454e-01	,		
-1.3337e+00	,	2.9608e+00	,	-3.7193e+00	,	-1.1941e+00	,	-1.0349e+00	,		
-2.5133e+00	,	1.7521e+00	,	-1.5778e+00	,	5.4771e-01	,	4.6839e-01	,		
-1.9399e+00	,	-1.3847e-01	,	3.9830e+00	,	4.9884e+00	,	4.8193e-01	,		
-1.7010e+00	,	6.2994e-01	,	2.9822e+00	,	-2.4728e-02	,	2.6371e+00	,		
1.0801e+00	,	-2.2338e+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+00	,	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-01	,	-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	,	9.1374e-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	,	-1.2935e+00	,	2.3924e+00	,		
1.4003e+00	,	-1.5616e+00	,	7.1990e-01	,	-1.2839e-01	,	1.5071e+00	,		
-2.3197e+00	,	-5.9906e-01	,	3.1609e-01	,	4.8745e+00	,	1.7453e+00	,		
4.0927e+00	,	-5.4239e-04	,	-6.3825e-04	,	3.4356e+00	,	1.4135e+00	,		
2.8070e-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	,	-2.1909e+00	,		
4.1397e+00	,	4.2648e-01	,	4.6854e+00	,	1.0355e+00	,	1.4013e+00	,		
-1.0843e-01	,	-5.9694e-01	,	-4.020e-01	,	4.2305e+00	,	-5.2332e-01	,		
4.1959e+00	,	-2.2805e-02	,	-6.3232e-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+00	,	3.1652e-01	,	-5.7089e-01	,	-1.2883e+00	,		
9.8610e-01	,	1.0568e+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	,		
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	],

[ 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 , -0.6964 , 0.9548 , 0.62282 , -1.8315 , -2.8587 ,  
0.48424 , -0.20527 , 0.67801 , -2.3472 , 1.0354 , -1.8257 ,  
-0.3892 , -0.24943 , 0.6767 , -1.5195 , 1.2166 , -2.698 ,  
-0.96698 , 2.2175 , -0.16889 , 0.9677 , -0.19646 , 1.3284 ,  
4.0824 , 1.3919 , 0.80669 , 0.303 , -0.28056 , -1.8632 ,  
0.47716 , -0.53628 , 1.3853 , -2.105 , 0.2354 , 2.4933 ,  
-0.87255 , 1.4493 , -0.10778 , -0.4411 , 3.462 , 4.4211 ,  
1.8385 , 0.3985 , 0.47637 , -0.60074 , 0.3583 , -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 , -1.385e+00 , 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+00 , -2.6846e-02 , 1.1748e+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 ,  
-2.2626e+00 , 4.8440e+00 , 8.9114e-01 , -2.5486e+00 , -6.9544e-01 ,  
6.2643e+00 , 2.9279e-01 , 3.1008e+00 , -3.2464e+00 , 1.5747e+00 ,  
-1.1939e+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 , -6.1671e-01 , -3.1454e-01 ,  
-1.3337e+00 , 9.6086e+00 , -3.7193e+00 , -1.1941e+00 , -1.0349e+00 ,  
-2.5313e+00 , -1.7110e+00 , -1.5778e+00 , 5.4771e-02 , 4.6839e-01 ,  
-1.9399e+00 , -1.1710e+01 , -3.9830e+00 , 4.9884e+00 , -4.8193e-01 ,  
-1.7010e+00 , 6.9911e+00 , 2.9822e+00 , -2.4728e-02 , -6.3717e+00 ,  
1.8801e+00 , -2.2748e+00 , 5.2203e+00 , 1.5099e+00 , -1.8284e+00 ,  
-3.9196e-01 , 1.7775e+00 , 6.5188e-01 , -1.0951e+00 , -1.5319e+00 ,  
2.8311e+00 , 2.9736e+00 , 5.1989e+00 , -1.0706e+00 , 5.9841e-01 ,  
-5.2798e+00 , -5.5499e-01 , -5.7471e+00 , 3.1155e+00 , -1.0095e+00 ,  
3.3329e+00 , 6.5440e+00 , -1.3949e+00 , -1.4996e+00 , -2.4218e+00 ,  
6.9150e+00 , -1.4240e+00 , 3.3080e-01 , -1.1709e+00 , 3.7678e+00 ,  
-1.6502e+00 , -1.6829e+00 , 2.3409e-01 , -2.0174e+00 , 2.0174e+00 ,  
-2.6225e+00 , -3.7696e-01 , -2.1272e+00 , 3.4266e+00 , 56119e+00 ,  
-2.1298e+00 , 9.7029e-01 , 5.1133e-02 , 8.2768e-01 , 3.564e+00 ,  
5.9028e-02 , -5.1805e-01 , -1.0821e+00 , -1.7695e+00 , -1.9489e-02 ,  
-2.5980e+00 , -4.9045e-02 , 6.9158e-01 , 9.1374e-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 , 1.9750e+00 , -1.2935e+00 ,  
1.4003e+00 , -1.5616e+00 , 7.1990e-01 , -1.2839e-01 , 1.5071e+00 ,  
-2.3197e+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 , -2.1909e+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 , -6.3232e-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 , -1.0584e+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 ,  
-1.1681 , -2.6261 , 1.9156 , -1.5651 , 1.8315 , -1.4257 ,  
-1.354 , -0.51953 , 1.7635 , -0.50722 , 1.388 , -1.1012 ],

Banana  
Chocolate  
Car

Banana



Car

Chocolate



Car

Banana

Chocolate



## Summary: text representation

Text representations are powerful algorithms

- Just by themselves: Useful for understanding & exploring the corpus (e.g., word frequencies)
- Even more so in combination with other procedures such as document classification, topic modeling, named entity recognition, etc.

**More on text representation later in the week**

# Text Processing / Feature Engineering

Day 1 Session 2

---

# Text Preprocessing

- Texts are **highly** dimensional
- When possible, it is nice to reduce this dimensionality
- Ideally, without losing too much information

# Danny & Spirling, 2018

- Punctuation
  - Numbers
  - Lowercasing
  - Stemming
  - Stop-words
  - N-grams
  - Removal of words by frequency
-

# Punctuation / Numbers / Lowercasing

- Fairly straightforward
- Often we don't care about punctuation and/or numbers – so, might be better to remove them
- We probably do care about the letter case
  - But to what extent?
  - Reduction in dimensions might be worth the reduction in accuracy
  - Can you think of examples when we do /don't care about the case?

# Stemming / Lemmatization

- A **stem** is the part of the word responsible for lexical meaning
  - A stem is invariable part of the word under inflection
  - “wait” is a stem of:
    - “waiting”
    - “waited”
    - “waits”
  - A **lemma** is the base / “original” part of the word
  - Both are useful for dimension reduction and often produce similar results
-

# Stop Words

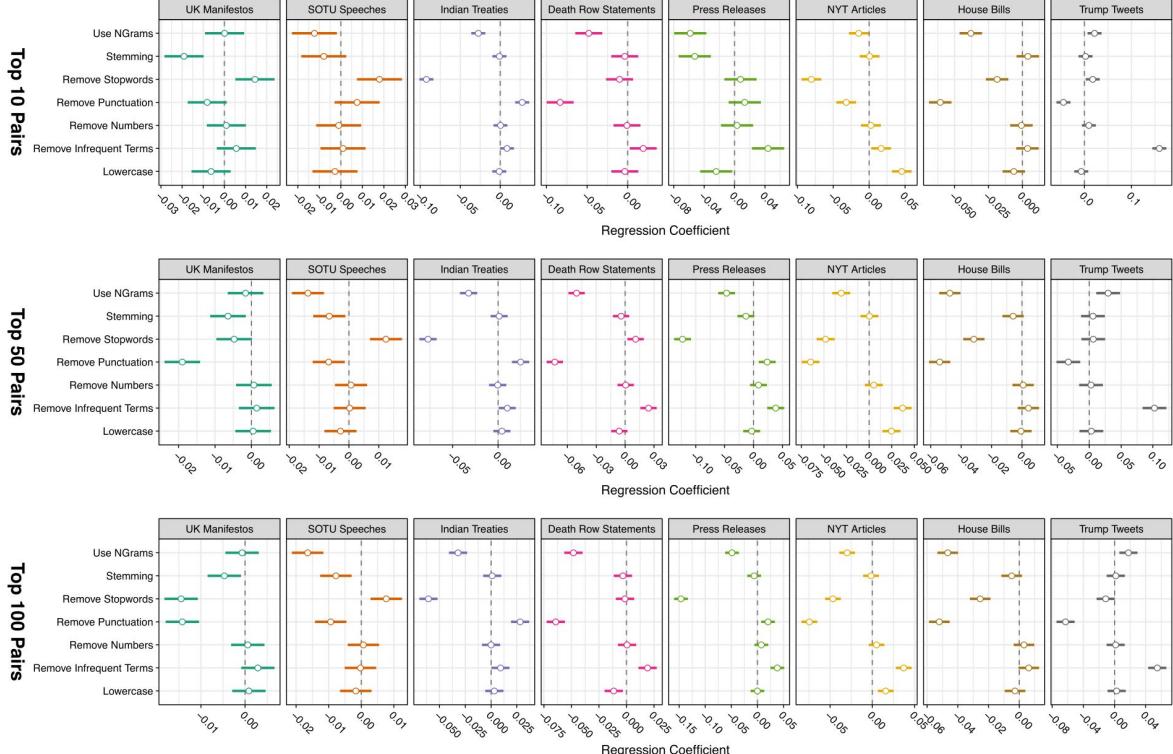
- Words that are filtered out before the analysis begins
- Could be any type of words that you do not want in the analysis
- Usually, function words are used as stop words (*FORESHADOWING...*)
  - “The”
  - “Is”
  - “I”
  - “That”
  - etc.
- Domain-specific words are also often excluded from the analysis
- E.g., “Global Warming” in the corpus of texts about Global Warming

## N-grams

- So far, we've only looked at "unigrams" – individual words
  - Texts can be broken down into any n-gram sequences
  - "I love ice-cream and bananas"
    - "I" "love" "ice-cream" "and" "bananas"
    - "I love" "love ice-cream" "ice-cream and" "and bananas"
    - 3-grams?
-

## Removal of terms by frequency

- Further removal of dimensionality can be achieved by removing either very frequent or very infrequent terms
- If they are very frequent, they probably don't carry much discriminating information for our analysis (think stopwords)
- If they are very infrequent, they probably carry a lot of discriminating information, but very low statistical power



Danny & Spirling, 2018

## Tf-idf

We can do more than just **count** words

We can transform these counts

Use some sort of a weight in order to transform

***Term frequency inverse document frequency*** in one form of weighting

---

# Term Frequency

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

## Term Frequency Frequency

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

## Inverse Document

$$idf = \log \frac{N}{n_j}$$

## Term Frequency

$$tf(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

## Inverse Document Frequency

$$idf = \log \frac{N}{n_j}$$

Number of Documents

Number of Documents where the term  $j$  appears

**tfidf**

$$W_{ij} \times \log \frac{N}{n_j}$$

# What?

# What?

- ***Exactly!***
  - Introduced in 1972 by a computer scientist (Spärck Jones, 1972)
  - No theoretical justification
  - Apart from “it seems to work...”
-

# What?

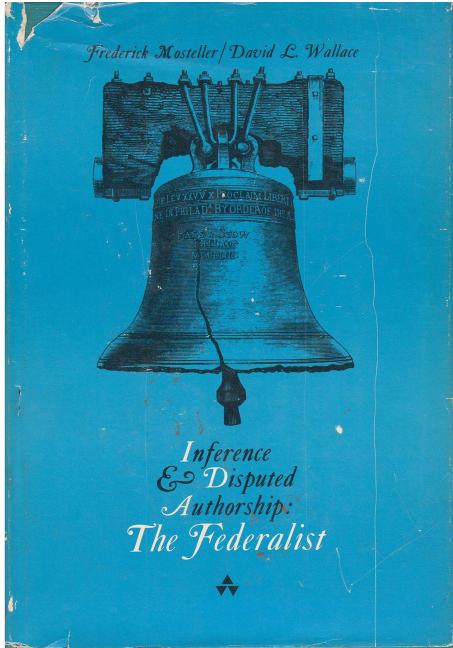
- ***Exactly!***
- Introduced in 1972 by a computer scientist (Spärck Jones, 1972)
- No theoretical justification
- Apart from “it seems to work...”
  
- And *sometimes* it does

## Log Odds / Log Odds Ratio

$$\log O_w^i = \log \frac{f_w^i}{1 - f_w^i}$$

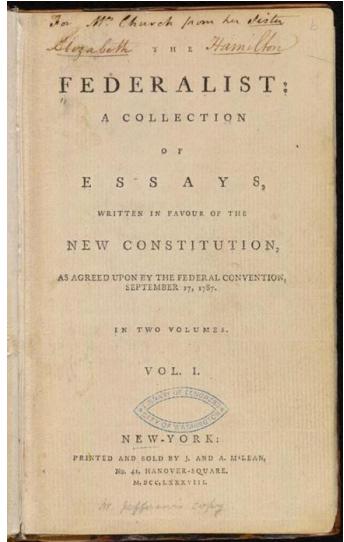
$$\log \frac{O_w^i}{O_w^j} = \log \frac{f_w^i}{1 - f_w^i} / \frac{f_w^j}{1 - f_w^j} = \log \frac{f_w^i}{1 - f_w^i} - \log \frac{f_w^j}{1 - f_w^j}$$

# Inference and Disputed Authorship: The Federalist

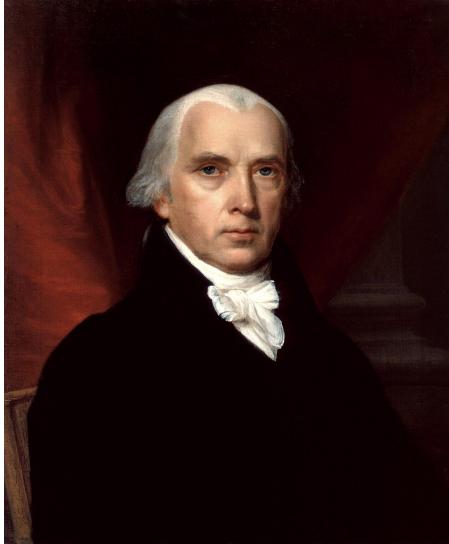
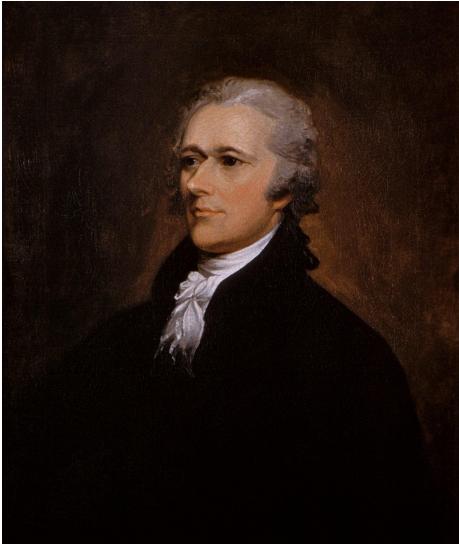
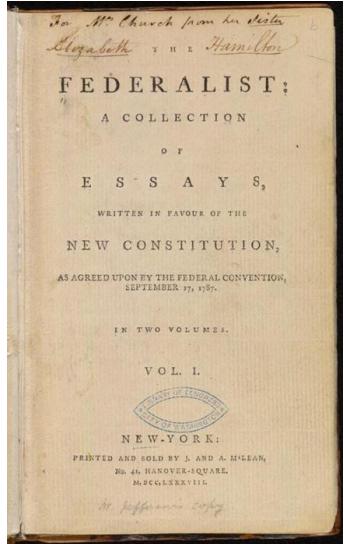


Mosteller & Wallace, 1963

# One of the first (if not the first) text-as-data study

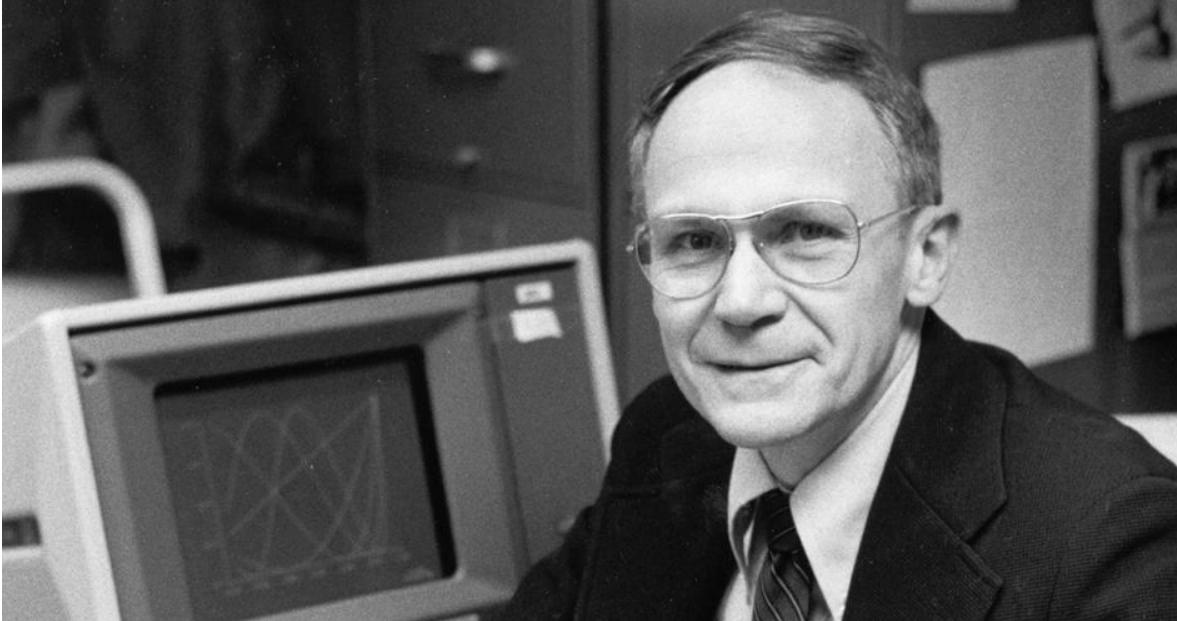


# One of the first (if not the first) text analysis study



## Who wrote them?

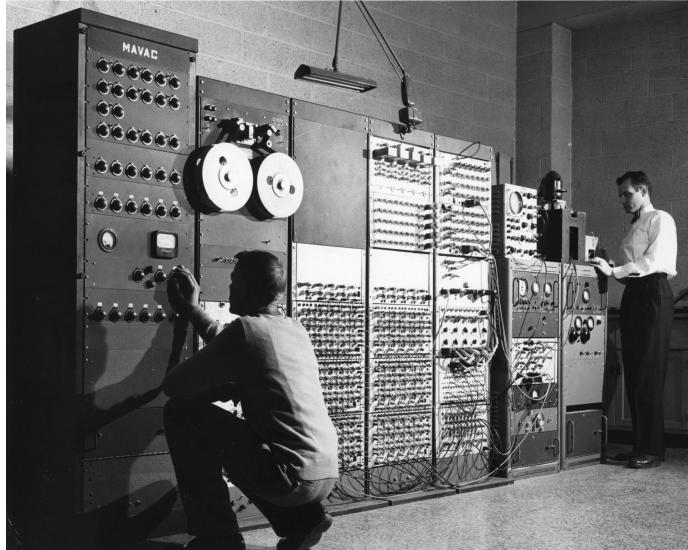
- 71 of the essays have a fairly certain authorship
- 12 are disputed
- Big historical debate as to how to ascribe authorship



---

# Computer-assisted text analysis!

# Computer-assisted text analysis...?



# Dimension Reduction

Remove all the stop-words!

# Dimension Reduction

Remove all the stop-words!

Still, too many words...

# Dimension Reduction

Remove all the stop-words!

Still, too many words...

Remove **all** words, **but** the stop-words

---

# Dimension Reduction

Remove all the stop-words!

Still, too many words...

Remove **all** words, **but** the stop-words

Maybe there is information in them?

---

## Simplified example from Grimmer et al., 2022

- Focus on:
  - “Man”
  - “By”
  - “Upon”
- The rates with which the authors use these words may indicate authorship

# Word Rates

	man	by	upon
Hamilton	102	859	374
Madison	17	474	7
Jay	0	82	1

# Word Proportions

	man	by	upon
Hamilton	.076	.643	.28
Madison	.034	.952	.014
Jay	0	.988	.012

# Multinomial Language Models

## Word Proportions



	man	by	upon
Hamilton	.076	.643	.28
Madison	.034	.952	.014
Jay	0	.988	.012

# Disputed Paper

	man	by	upon
Disputed	2	15	0

## Disputed Paper

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0$$

Total Words

Disputed Paper

Raw Rates from text

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0$$

Hamilton Rates

# Calculate Jay and Madison yourself

---

$$p(D|H) = \frac{17!}{2!15!0!} (.076)^2 \times (.643)^{15} \times (.28)^0 = .001$$

$$p(D|J) = \frac{17!}{2!15!0!} (0)^2 \times (.988)^{15} \times (.012)^0 = 0$$

$$p(D|M) = \frac{17!}{2!15!0!} (.034)^2 \times (.952)^{15} \times (.014)^0 = .076$$

---

# Federalist Vector Space Model

In the markdown file...

# Coding challenge

Today:

R\_basics.RmD

basics\_text\_analysis.RmD

feature\_engineering.Rmd