Quantitative Text Analysis

Day 1: Text Representation

Petro Tolochko

Contents of the Course

- Introduction to Text-as-Data, Text Representation
- Preprocessing (tokenization, stemming, etc.) and Feature Engineering (BoW, TF-IDF, n-grams)
- Regular Expressions and Dictionary Methods
- Text Annotation and Supervised Labeling
- Machine Learning for Text Classification (Naive Bayes, SVM, etc.)
- Unsupervised Learning: Topic Modeling and Clustering
- Word Vectors and Semantic Similarity (word2vec, GloVe)
- Transformers and Large Language Models (BERT, GPT)
- Prompt Engineering and Zero-shot Learning
- Model Validation and Human Evaluation
- Networks approaches to Text-as-Data
- Ethics, Bias, and Critical Reflection in Text Analysis

- Agenda subject to slight change

Course Structure

- Theoretical Input
- Practical Sessions

Introduction

- Post-Doc, Computational Communication Science Lab, Univ. of Vienna
- Research focus: Text complexity, social networks, statistical modeling
- petro.tolochko@univie.ac.at

Introduction: Students

- Name?
- Background?
- Experience with Text Analysis?
- Programming?
- Expectations / wishes for the course?

Objectives

- Get to know methods of automated text analysis
- Practical challenges
- Critical reflection
- Utility for your own projects

Course Assessment

- Participation + Engagement
- Homeworks
- Take-home Assignment

Personal Feedback

- Informal feedback on your projects
- Research question, Data, Methods, Current struggles

Course Resources

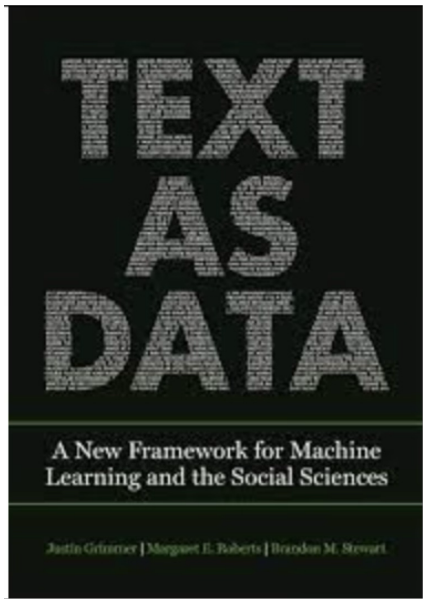
- https://github.com/PeterTolochko/CEU_text_as_data_python

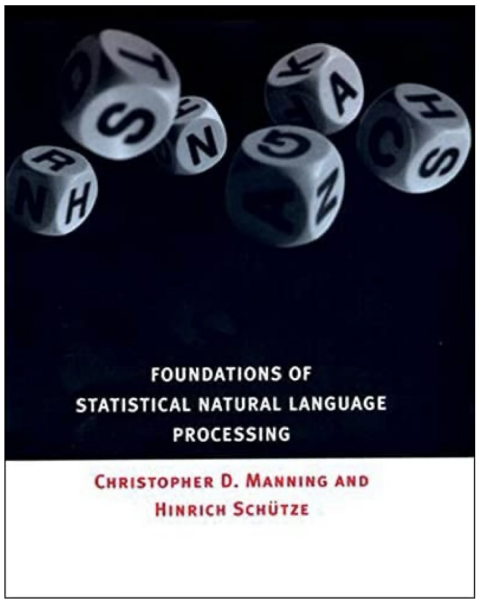
Today

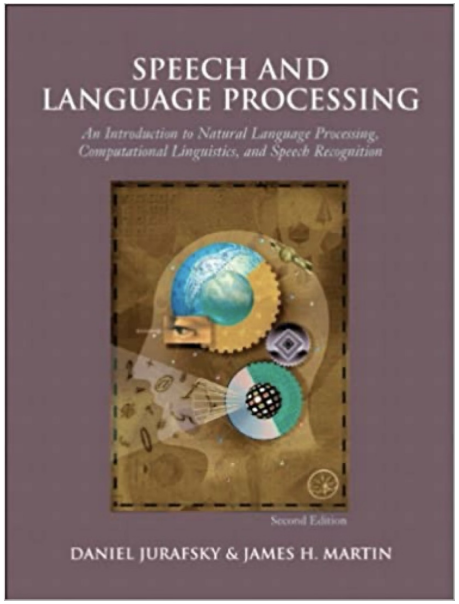
- Text Representation

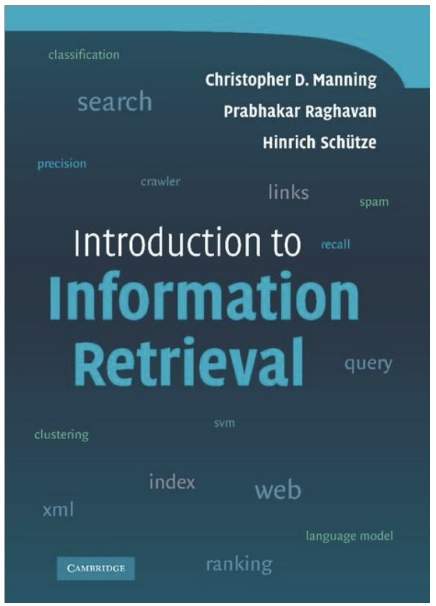
Questions?

Resources

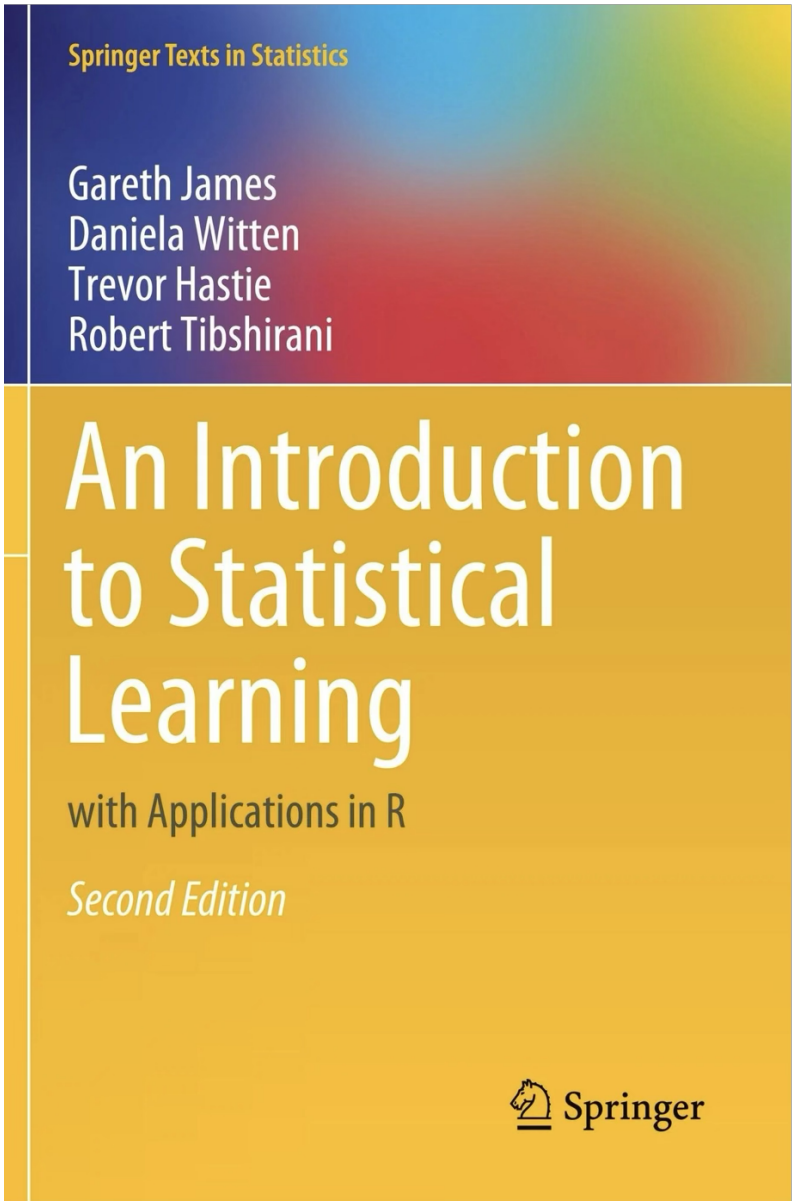

TEXT AS DATA
A New Framework for Machine Learning and the Social Sciences
Justin Grimmer | Margaret E. Roberts | Brandon M. Stewart

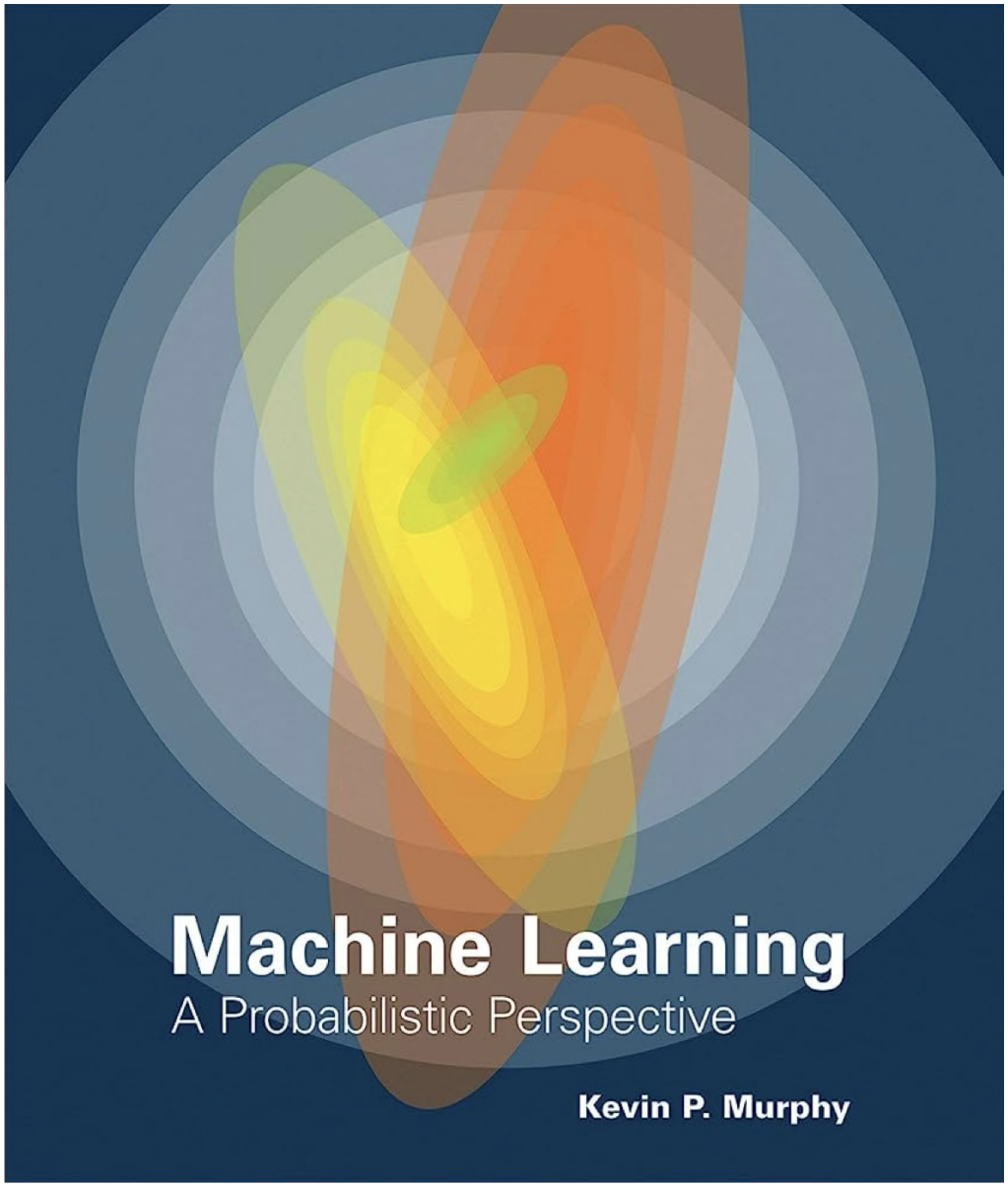

FOUNDATIONS OF STATISTICAL NATURAL LANGUAGE PROCESSING
CHRISTOPHER D. MANNING AND HINRICH SCHÜTZE


SPEECH AND LANGUAGE PROCESSING
An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition
Second Edition
DANIEL JURAFSKY & JAMES H. MARTIN


classification
search
Christopher D. Manning
Prabhakar Raghavan
Hinrich Schütze
precision
crawler
links
spam
Introduction to
recall
Information Retrieval
query
clustering
svm
index
web
xml
language model
CAMBRIDGE
ranking

Resources


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer


Machine Learning
A Probabilistic Perspective
Kevin P. Murphy

Resources

- *Advanced Data Analysis from an Elementary Point of View*, Cosma Rohilla Shalizi (free online)
- `https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/`

Why Analyse Texts?

- Vast quantities of data available
- Texts carry meaning
- Texts capture social behaviour

Quantitative Analysis of Culture Using Millions of Digitized Books

Article *in* Science · January 2011

DOI: 10.1126/science.1199644 · Source: PubMed

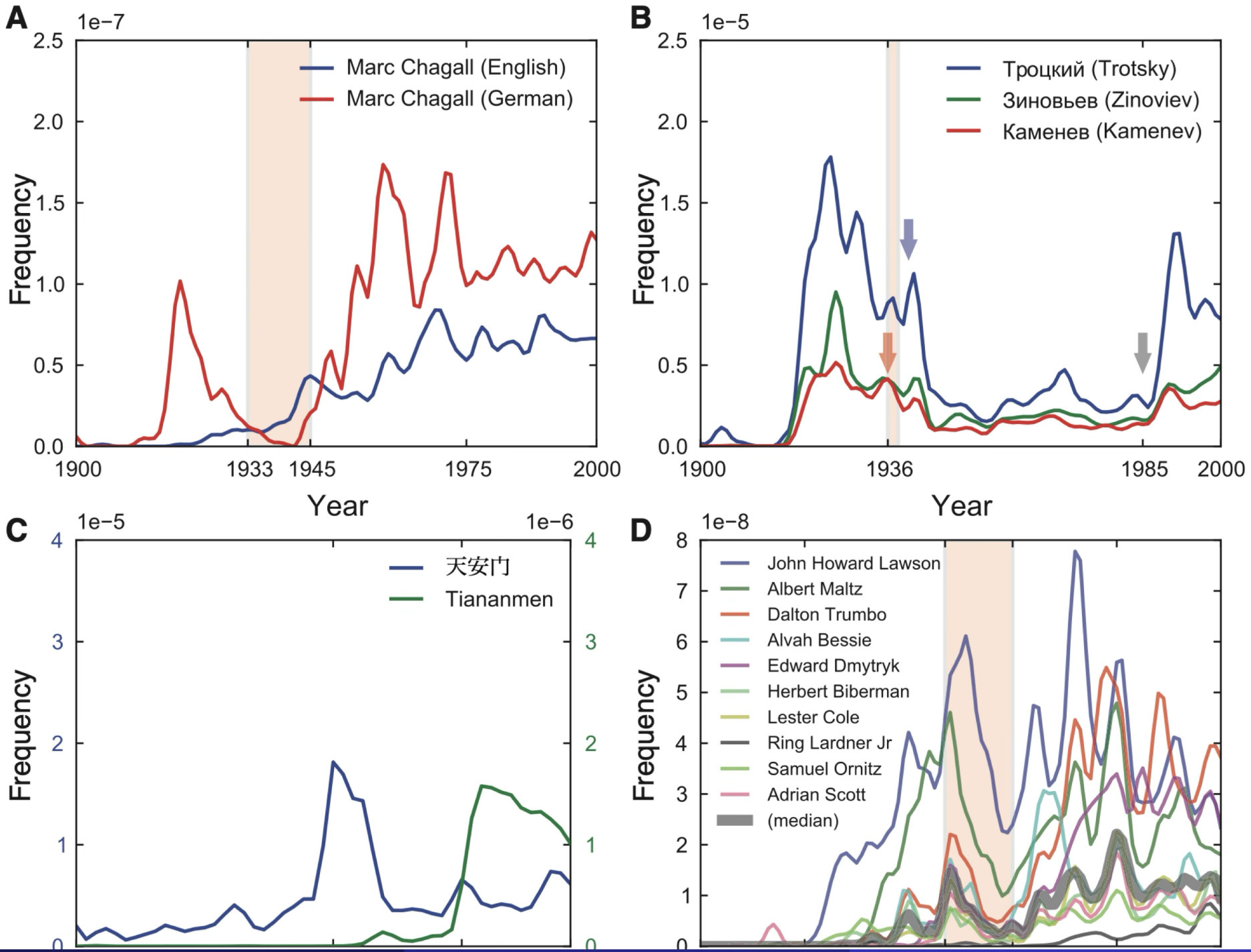
13 authors, including:

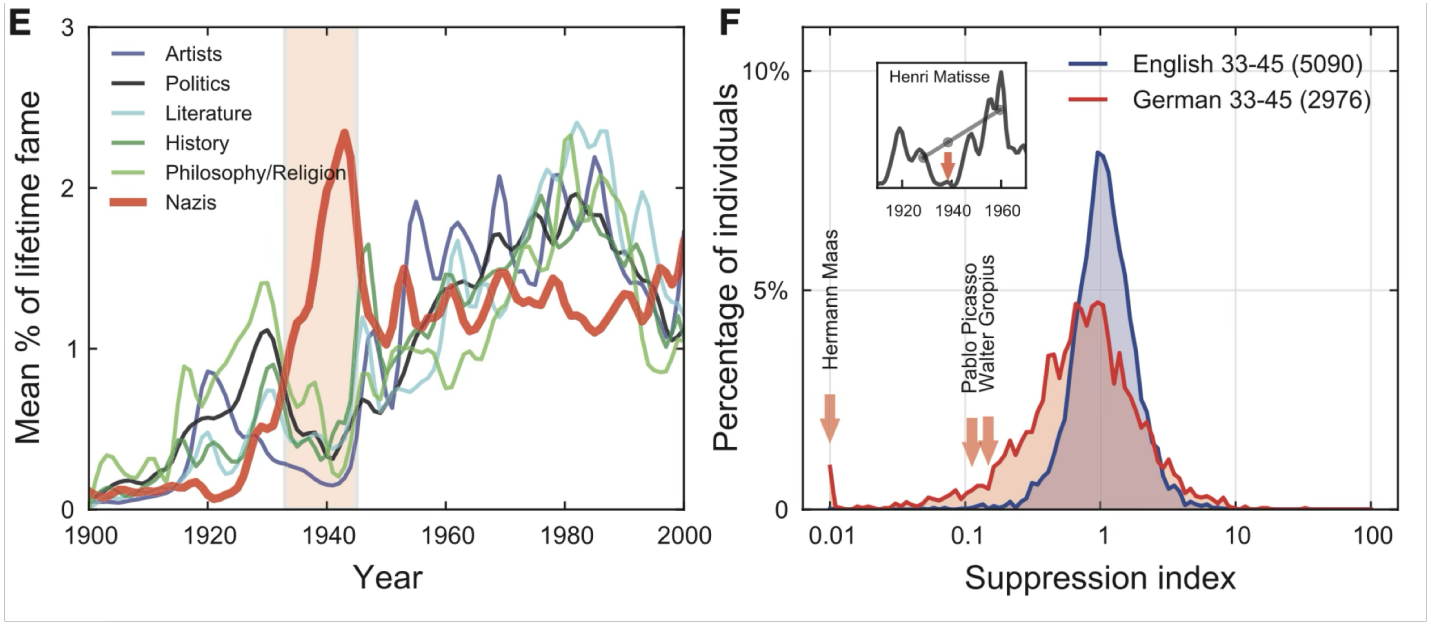
Yuan Kui Shen
Massachusetts Institute of Technology
11 PUBLICATIONS **2,326** CITATIONS
SEE PROFILE

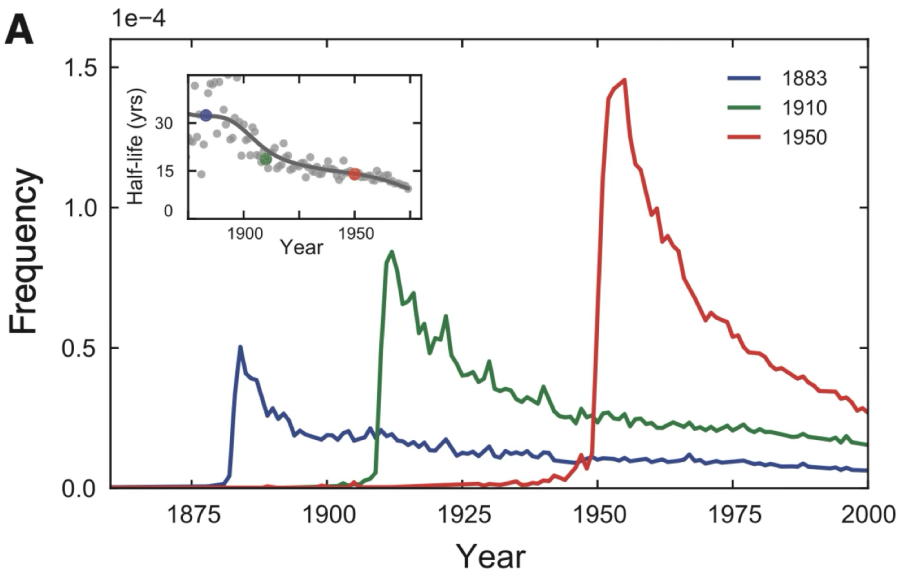
Aviva Aiden
Baylor College of Medicine
3 PUBLICATIONS **2,656** CITATIONS
SEE PROFILE

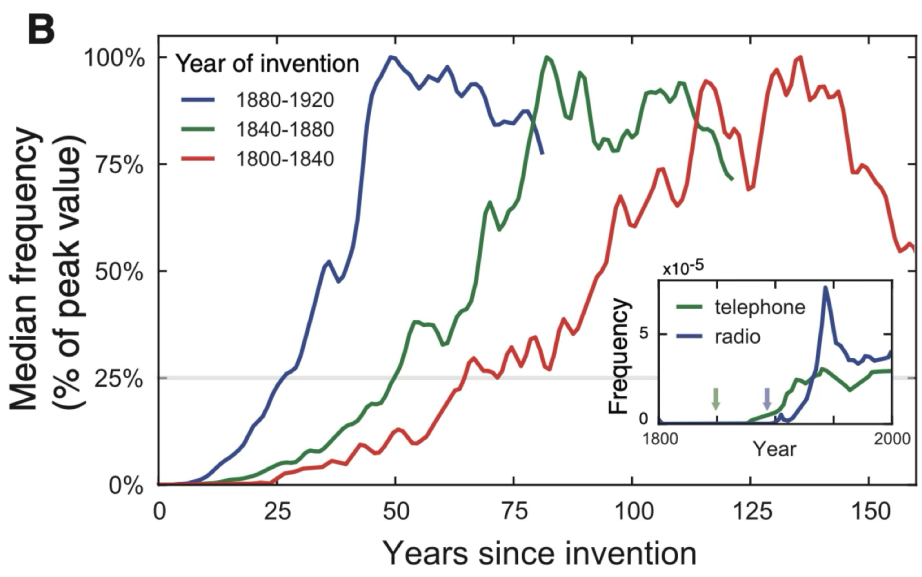
Adrian Veres
Harvard University
18 PUBLICATIONS **8,790** CITATIONS
SEE PROFILE

Peter Norvig
Google Inc.
134 PUBLICATIONS **48,502** CITATIONS
SEE PROFILE


A
1e−7
Marc Chagall (English)
Marc Chagall (German)
Frequency
Year
1900
1933
1945
1975
2000
B
1e−5
Троцкий (Trotsky)
Зиновьев (Zinoviev)
Каменев (Kamenev)
Frequency
Year
1900
1936
1985
2000
C
1e−5
1e−6
天安门
Tiananmen
Frequency
D
1e−8
John Howard Lawson
Albert Maltz
Dalton Trumbo
Alvah Bessie
Edward Dmytryk
Herbert Biberman
Lester Cole
Ring Lardner Jr
Samuel Ornitz
Adrian Scott
(median)
Frequency


E
Mean % of lifetime fame
Year
Artists
Politics
Literature
History
Philosophy/Religion
Nazis
F
Percentage of individuals
Suppression index
English 33-45 (5090)
German 33-45 (2976)
Henri Matisse
Hermann Maas
Pablo Picasso
Walter Gropius


A
1e−4
Frequency
Year
Half-life (yrs)
1883
1910
1950


B
Median frequency
(% of peak value)
Years since invention
Year of invention
1880-1920
1840-1880
1800-1840
telephone
radio
Frequency
Year

AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Predicting and Interpolating State-Level Polls Using Twitter Textual Data

Nicholas Beauchamp Northeastern University

Abstract: *Spatially or temporally dense polling remains both difficult and expensive using existing survey methods. In response, there have been increasing efforts to approximate various survey measures using social media, but most of these approaches remain methodologically flawed. To remedy these flaws, this article combines 1,200 state-level polls during the 2012 presidential campaign with over 100 million state-located political tweets; models the polls as a function of the Twitter text using a new linear regularization feature-selection method; and shows via out-of-sample testing that when properly modeled, the Twitter-based measures track and to some degree predict opinion polls, and can be extended to unpolled states and potentially substate regions and subday timescales. An examination of the most predictive textual features reveals the topics and events associated with opinion shifts, sheds light on more general theories of partisan difference in attention and information processing, and may be of use for real-time campaign strategy.*

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.

General Article

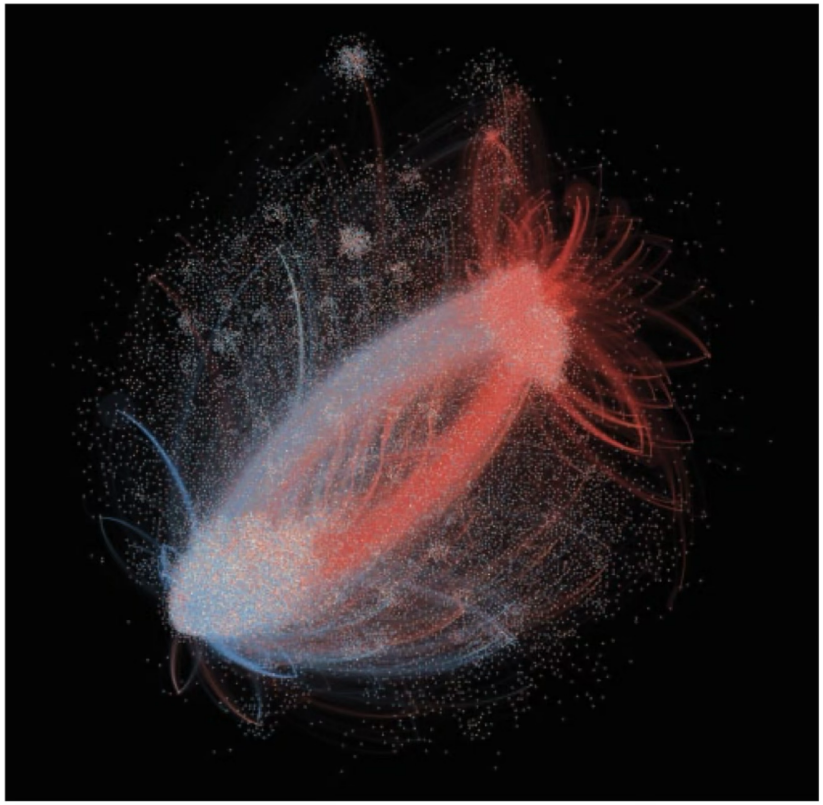
Psychological Science
2015, Vol. 26(10) 1531–1542

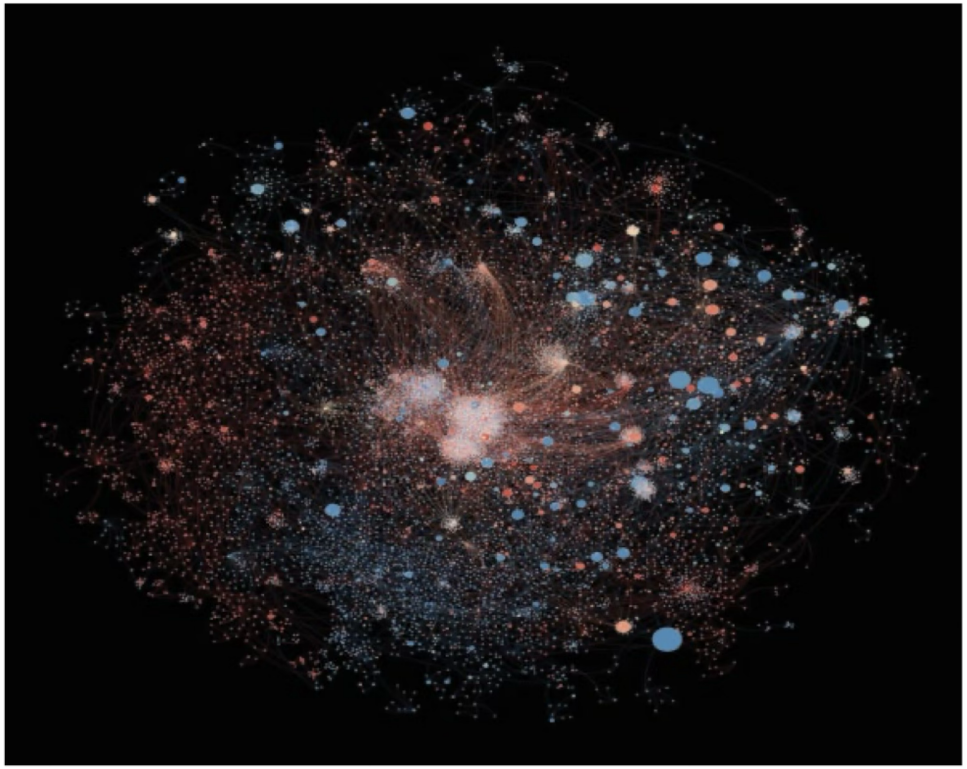
DOI: 10.1177/0956797615594620
pss.sagepub.com

Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?

Pablo Barberá[1], John T. Jost[1,2,3], Jonathan Nagler[3], Joshua A. Tucker[3], and Richard Bonneau[4]

[1]Center for Data Science, [2]Department of Psychology, [3]Department of Politics, and [4]Center for Genomics and Systems Biology, New York University





What is Text?

- Data
- Unstructured
- (Highly) Multidimensional
- Difficult to work with (if you're not human)

From Text to Structure

- We need to structure the text before analyses
- Different ways to represent text so computers "understand"
- Different ways to model text so that both (we and computer) "understand"
- Different research questions $\rightarrow$ different representations

Document-Term Matrix

$$X = N \times K$$

- N = number of documents
- K = number of terms/features

Document-Term Matrix

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

$$N = 3 \quad K = 6$$

Example Corpus

- Doc 1: “John loves ice-cream”
- Doc 2: “John loves oranges”
- Doc 3: “Mary hates ice-cream”

$$N =?, \quad K =? \quad X =?$$

Document-Term Matrix with Vocabulary

	John	loves	ice-cream	oranges	Mary	hates
D_1	1	1	1	0	0	0
D_2	1	1	0	1	0	0
D_3	0	0	1	0	1	1

Document-Term Matrix with Vocabulary

	John	loves	ice-cream	oranges	Mary	hates
D_1	1	1	1	0	0	0
D_2	1	1	0	1	0	0
D_3	0	0	1	0	1	1

$$N = 3, \quad K = 6 \quad X = 18$$

Types & Tokens

- Types: unique words
- Tokens: all words

Example corpus: 6 types, 9 tokens

Bag-of-Words Representation

- Representation of text as a bag of words
- Order irrelevant
- Each text represented as counts of words

Multinomial Model of Language

- Text $\approx$ draw from a Multinomial distribution

Binomial Distribution

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

$n =$ number of trials

$k =$ number of successes

$p =$ success probability

Multinomial Distribution

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2! \cdots x_k!}\, p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where: $n =$ text length, $k =$ vocabulary size, $p_i =$ probability of word i

Binomial vs. Multinomial

- **Binomial:** one word (success/failure)

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

- **Multinomial:** multiple words/categories

$$P(x_1, \ldots, x_k) = \frac{n!}{x_1! \ldots x_k!} \prod_{i=1}^{k} p_i^{x_i}$$

Example Texts

- Text 1 = banana banana banana banana chocolate
- Text 2 = chocolate chocolate chocolate banana fudge
- Text 3 = banana banana
- Text 4 = ice-cream ice-cream fudge ice-cream
- Text 5 = fudge fudge fudge
- Text 6 = ice-cream ice-cream fudge fudge

Example Texts

John

Text 1 = banana banana banana banana chocolate

Text 2 = chocolate chocolate chocolate banana fudge

Text 3 = banana banana

Mary

Text 4 = ice-cream ice-cream fudge ice-cream

Text 5 = fudge fudge fudge

Text 6 = ice-cream ice-cream fudge fudge

Document-Term Matrix

	banana	chocolate	fudge	icecream
Doc 1	4	1	0	0
Doc 2	2	0	0	0
Doc 3	1	3	1	0
Doc 4	0	0	1	3
Doc 5	0	0	3	0
Doc 6	0	0	2	2

Document-Term Matrix

	banana	chocolate	fudge	icecream
John Rates	7	4	1	0
Mary Rates	1	3	7	5

Language Models

John Language Model

John Rates 7 4 1 0

Mary Language Model

Mary Rates 1 3 7 5

New Texts

- new_text_1 = ice-cream fudge fudge
- new_text_2 = chocolate chocolate banana banana

Question: Probability they were generated by John or Mary?

New Texts Document-Term Matrix

	banana	chocolate	fudge	icecream
John Rates	0	0	2	1
Mary Rates	2	2	0	0

Probability Spoken by John

- new_text_1 = "ice-cream fudge fudge"

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

Probability Spoken by John

- new_text_2 = "chocolate chocolate banana banana"

$$
\begin{aligned}
Pr(b = 2, ch = 2, f = 0, i = 0) = & \frac{4!}{2!\,2!\,0!\,0!} \\
& \times (0.583)^2 \times (0.333)^2 \times (0.08)^0 \times (0)^0 \\
& = 6 \times (0.583)^2 \times (0.333)^2 \\
& = 0.23
\end{aligned}
$$

Probability spoken by Mary

new_text_1 (Mary): $P = 0.18$
new_text_2 (Mary): $P = 0.0008$

Questions?

Vector Space Model

- Represent texts as vectors in multidimensional space

Multidimensional Spaces

Multidimensional?

- 2D (Position on a map):
  - X = longitude, Y = latitude
- 3D (Position in the real world):
  - X = longitude, Y = latitude, Z = height
- 4D:
  - X = longitude, Y = latitude, Z = height, T = time

Multidimensional Spaces

- 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

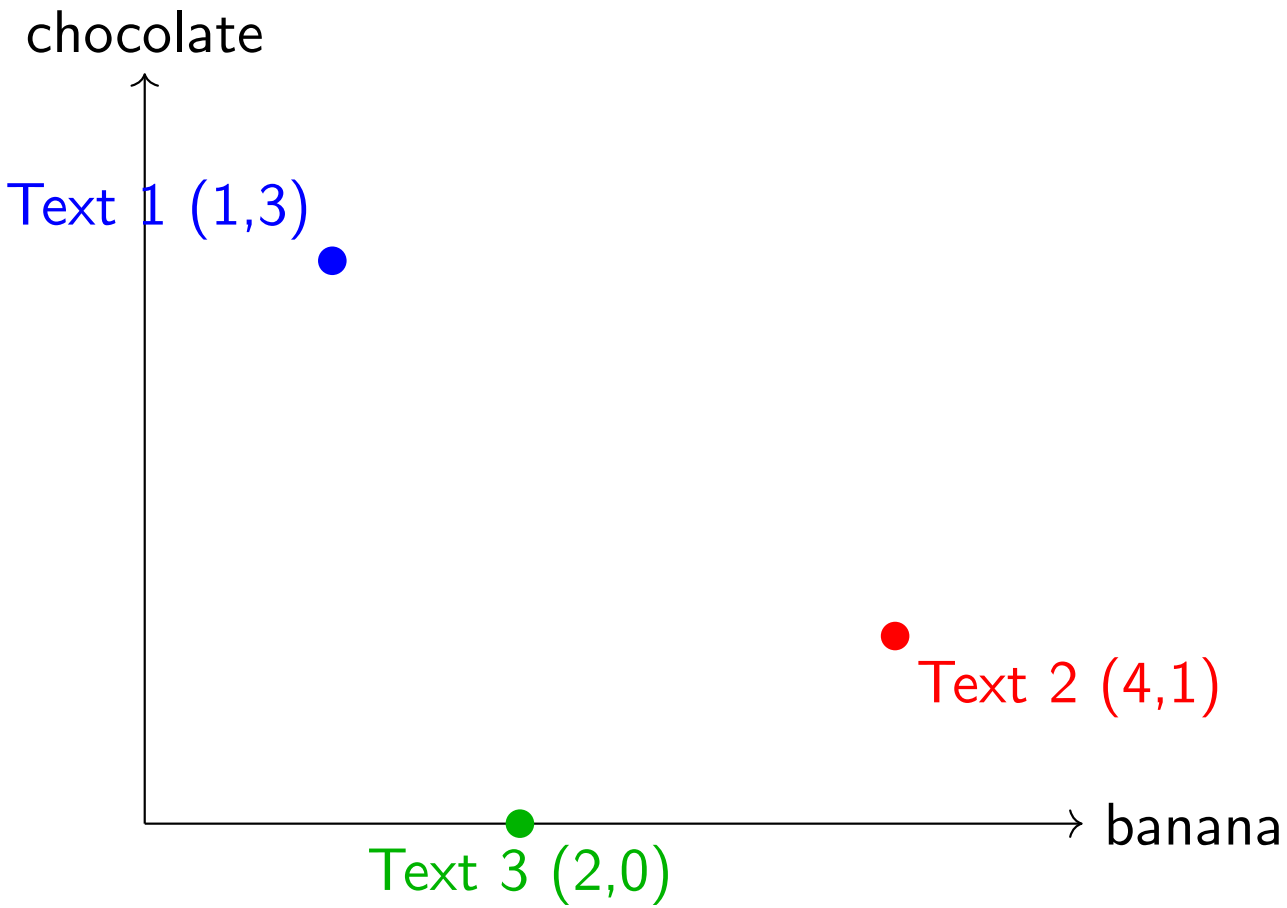
Two-dimensional space

- Vocabulary = {"banana", "chocolate"}
- Text 1 = "chocolate, chocolate, chocolate, banana"
- Text 2 = "banana, banana, banana, banana, chocolate"
- Text 3 = "banana, banana"

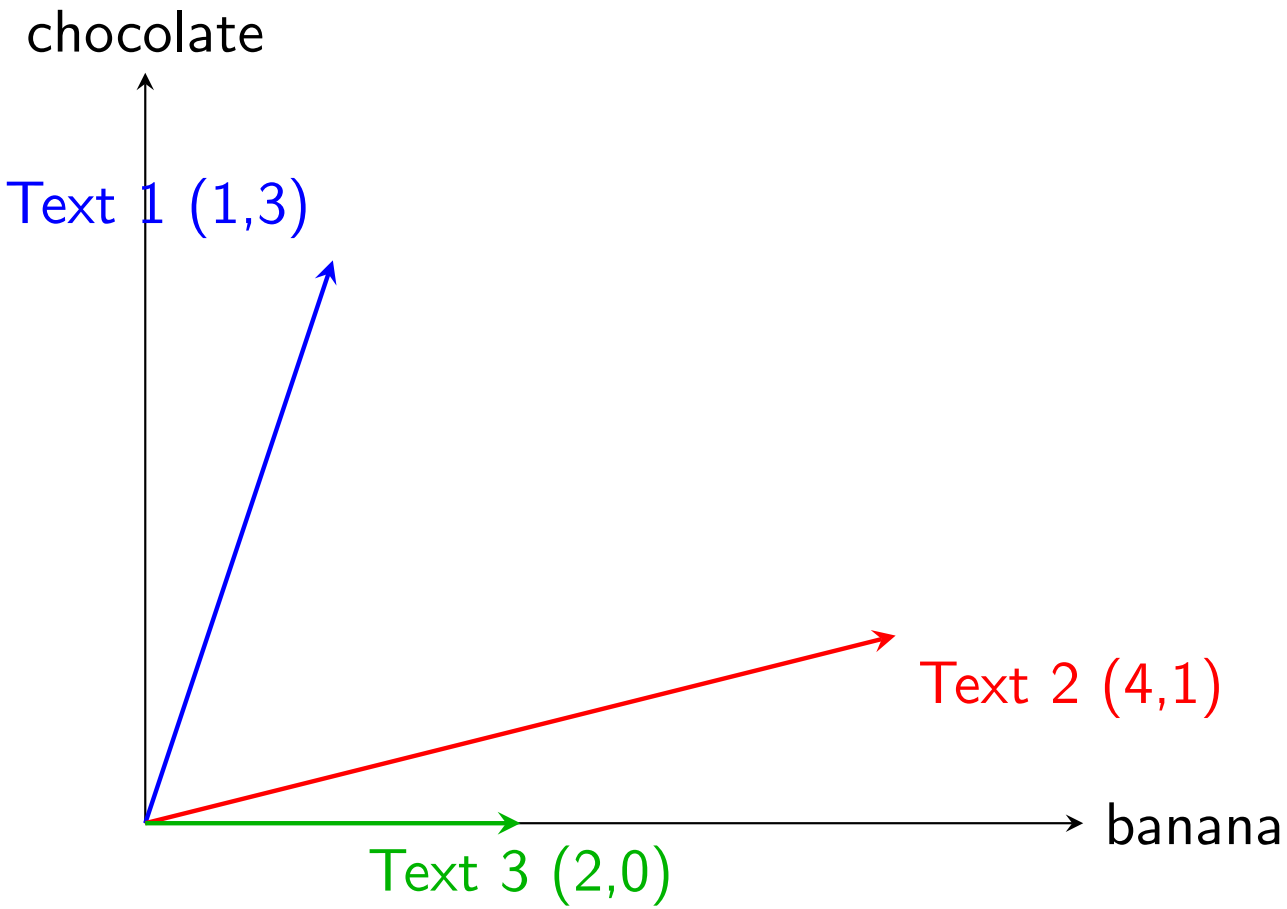
Two-dimensional space

- Vocabulary = {"banana", "chocolate"}
- Text 1 = "chocolate, chocolate, chocolate, banana" $\Rightarrow (1, 3)$
- Text 2 = "banana, banana, banana, banana, chocolate" $\Rightarrow (4, 1)$
- Text 3 = "banana, banana" $\Rightarrow (2, 0)$

Two-dimensional Vector Space



Two-dimensional Vector Space



Cosine Similarity

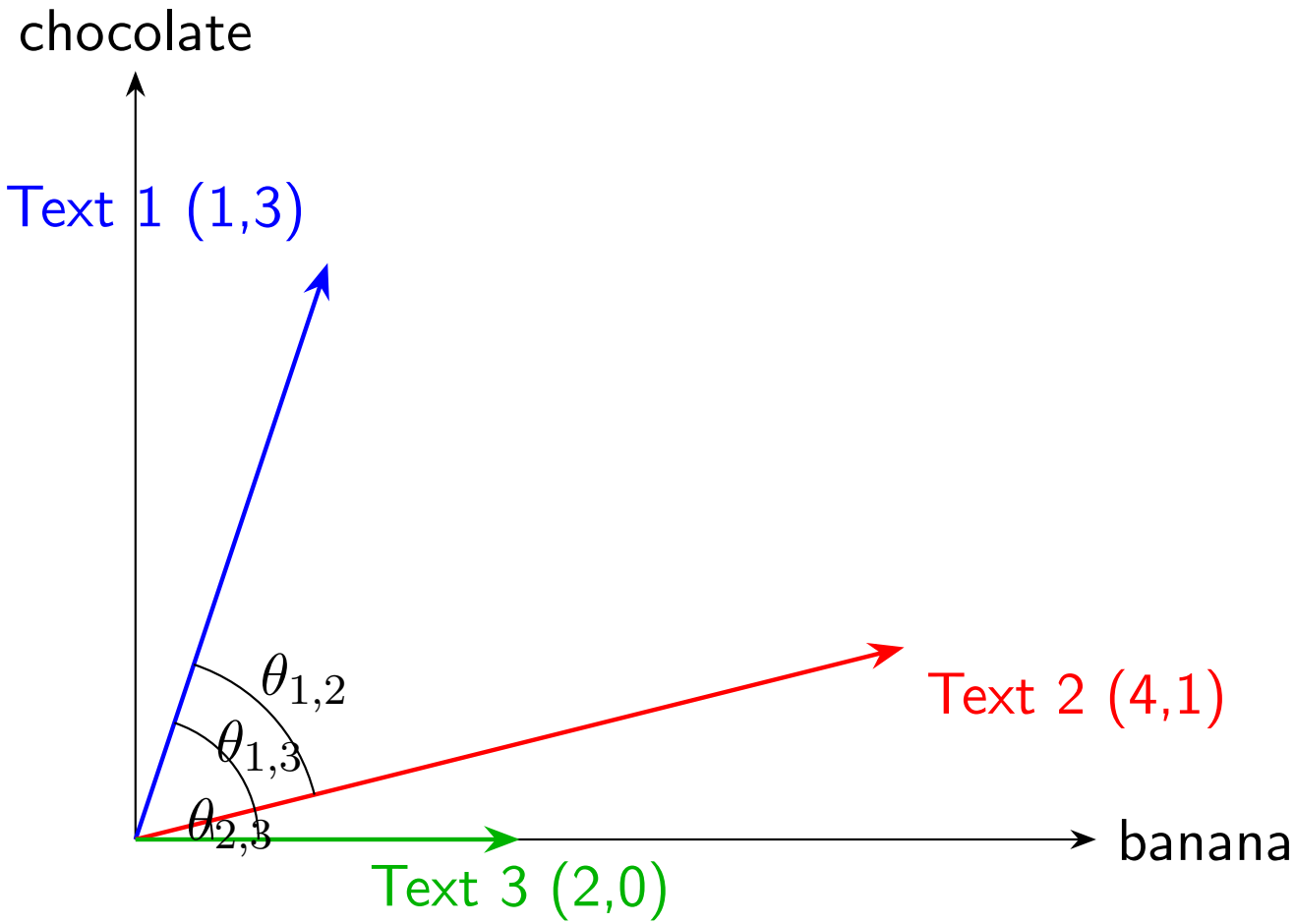
$$\cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|}$$

$$\cos(\theta_{1,2}) = \frac{1 \cdot 4 + 3 \cdot 1}{\sqrt{1^2 + 3^2}\sqrt{4^2 + 1^2}} = \frac{7}{\sqrt{10}\sqrt{17}} \approx 0.538$$

$$\cos(\theta_{1,3}) = \frac{1 \cdot 2 + 3 \cdot 0}{\sqrt{1^2 + 3^2}\sqrt{2^2 + 0^2}} = \frac{2}{\sqrt{10} \cdot 2} = 0.316$$

$$\cos(\theta_{2,3}) = \frac{4 \cdot 2 + 1 \cdot 0}{\sqrt{4^2 + 1^2}\sqrt{2^2 + 0^2}} = \frac{8}{\sqrt{17} \cdot 2} \approx 0.970$$

Two-dimensional Vector Space with Angles



Vector Distances

$$\text{Euclidean: } d_E(x,y) = \sqrt{\sum (x_i - y_i)^2},$$

$$\text{Manhattan: } d_M(x,y) = \sum |x_i - y_i|$$

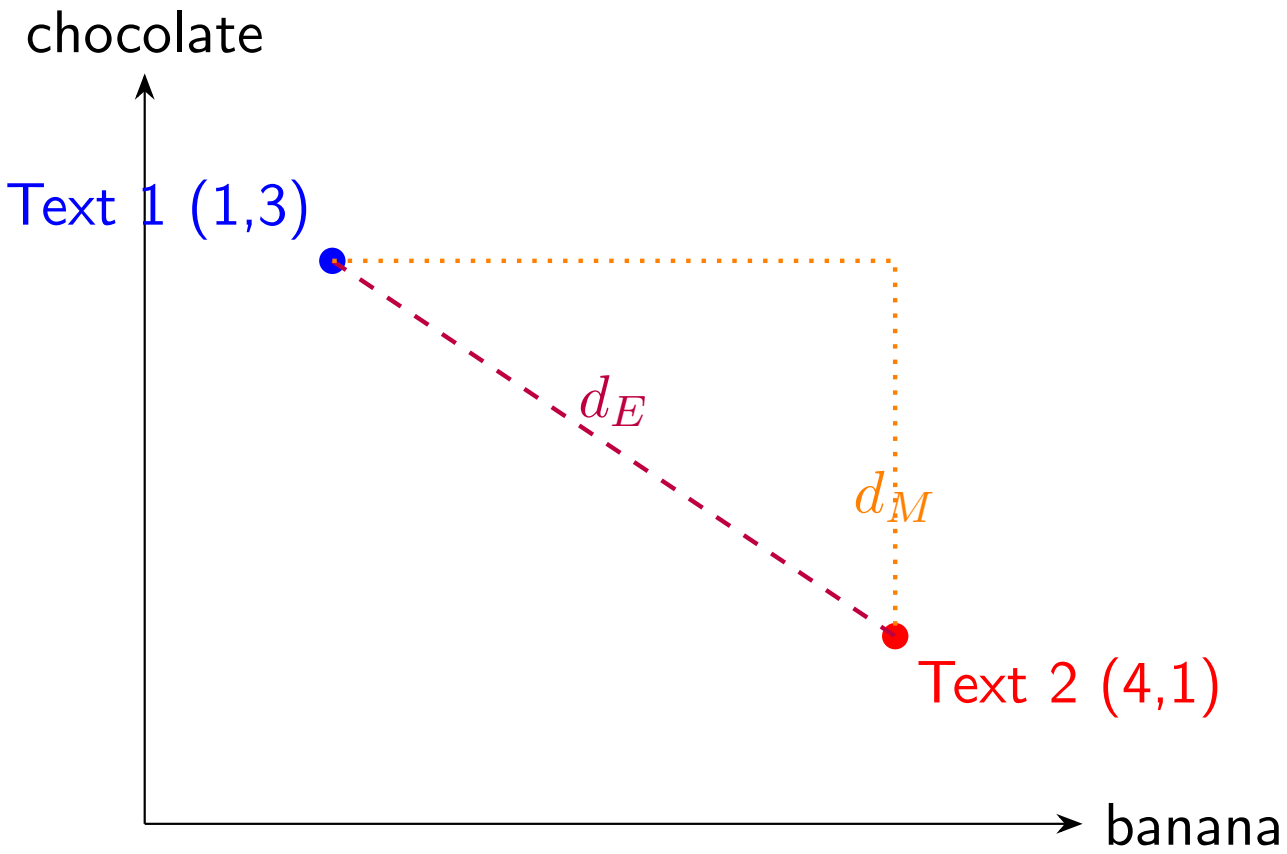
$$\begin{aligned} d_E(T1,T2) &= \sqrt{(1-4)^2 + (3-1)^2} \\ &= \sqrt{13} \approx 3.61 \end{aligned}$$

$$\begin{aligned} d_M(T1,T2) &= |1-4| + |3-1| \\ &= 5 \end{aligned}$$

$$\begin{aligned} d_E(T1,T3) &= \sqrt{(1-2)^2 + (3-0)^2} \\ &= \sqrt{10} \approx 3.16 \end{aligned}$$

$$\begin{aligned} d_M(T1,T3) &= |1-2| + |3-0| \\ &= 4 \end{aligned}$$

Euclidean vs. Manhattan Distance



Multidimensional Spaces

- $\approx$ 1,022,000 words in English language
- $\approx$ 1,022,000-dimensional space

Multidimensional Spaces

Same math, though

Questions?

Distributional Hypothesis / Word Embeddings

Distributional Hypothesis

- **Philosophy of language:** "The meaning of a word is its use" (Wittgenstein, 1953).
- **Linguistics:** "Difference in meaning correlates with difference in distribution" (Harris, 1954).
- **Key idea:** Words occurring in similar contexts tend to have similar meanings.
- (Firth, 1957): "You shall know a word by the company it keeps."

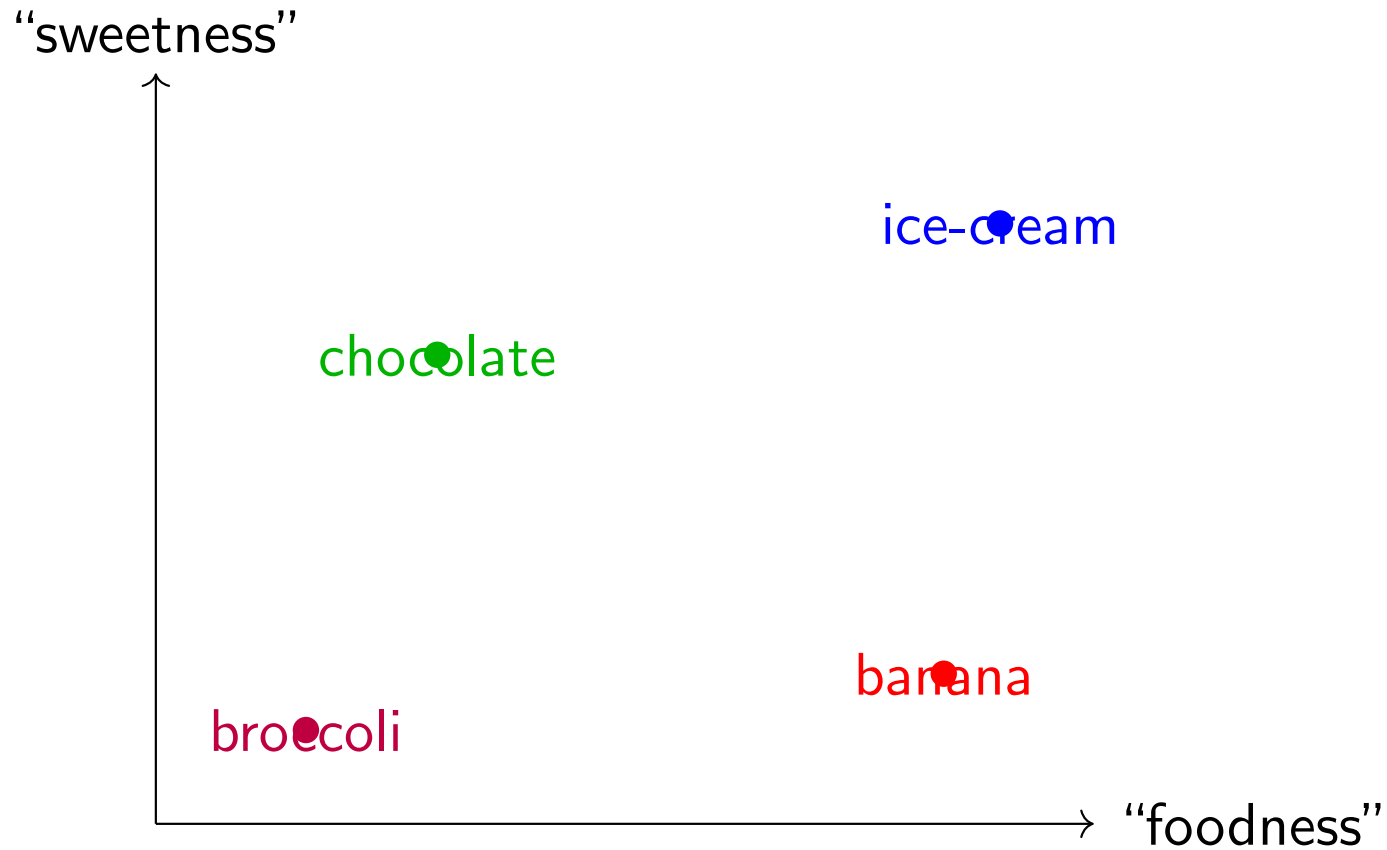
Word Vectors

- Represent each word as a vector in a high-dimensional space.
- Dimensions = features derived from context (co-occurrences).
- Similar words $\Rightarrow$ similar vectors (close in space).
- Early approaches: Count-based co-occurrence matrices + dimensionality reduction (e.g., SVD, LSA).

Neural Word Embeddings

- Predictive models (neural networks) learn vector representations.
- Word2Vec (Mikolov et al., 2013) — famous methodology:
  - Skip-gram: predict context words from a target word.
  - CBOW: predict target word from context words.
- Each word $\Rightarrow$ dense vector (e.g., 300 dimensions).
- Captures semantic relationships (e.g., `king - man + woman = queen`).

Word Vectors in 2D (Toy Example)



Measuring Similarity in Embeddings

- Word embeddings map words into a high-dimensional vector space.
- To compare meanings, we measure the **similarity between vectors**.
- Most common metric: **Cosine similarity**

$$\cos(\theta) = \frac{x \cdot y}{\|x\| \, \|y\|}$$

- Interpretation:
  - $\cos(\theta) \approx 1 \Rightarrow$ words are similar (small angle).
  - $\cos(\theta) \approx 0 \Rightarrow$ unrelated.
  - $\cos(\theta) < 0 \Rightarrow$ opposite contexts.
- Alternative metrics: Euclidean distance, Manhattan distance.

Example: Cosine Similarity of Word Vectors

- Word vectors (toy example):

$$\text{banana} = (0.7, 0.2, 0.1), \quad \text{chocolate} = (0.6, 0.3, 0.4)$$

$$\cos(\theta) = \frac{x \cdot y}{\|x\|\|y\|}$$

$$\begin{aligned} x \cdot y &= (0.7)(0.6) + (0.2)(0.3) + (0.1)(0.4) \\ &= 0.42 + 0.06 + 0.04 = 0.52 \end{aligned}$$

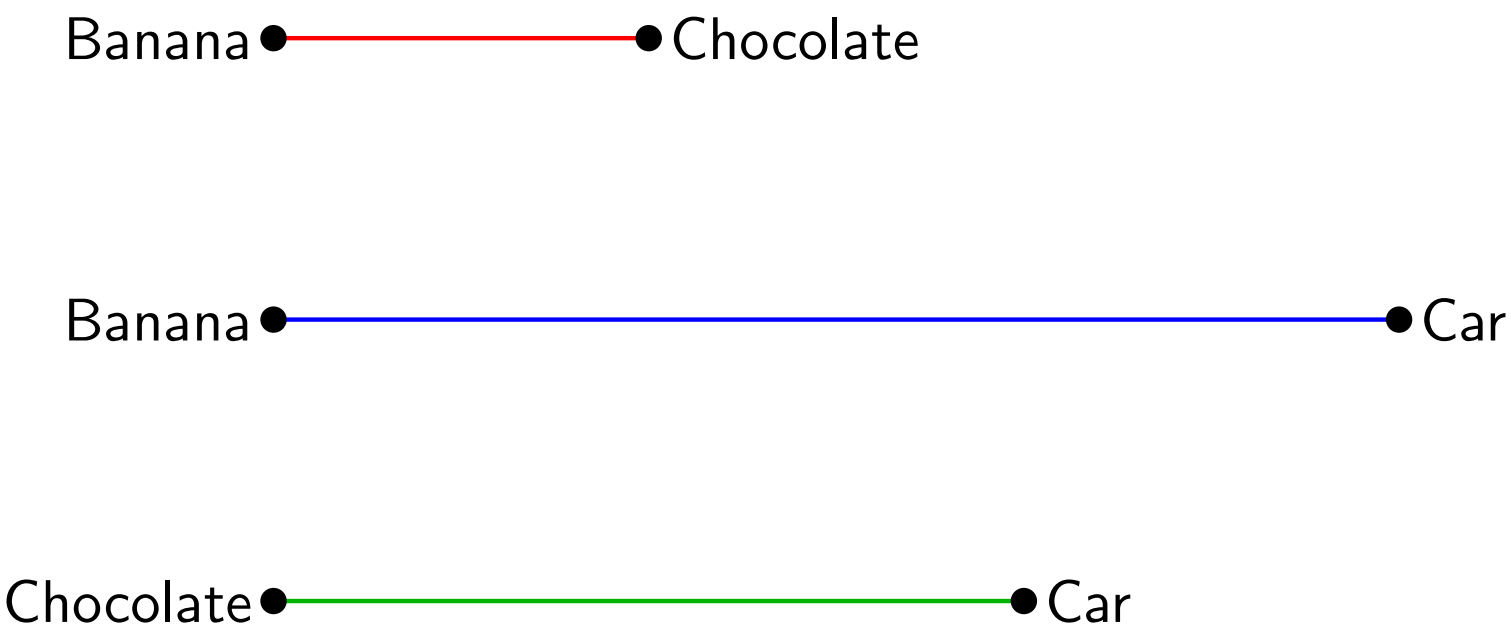
$$\|x\| = \sqrt{0.7^2 + 0.2^2 + 0.1^2} = \sqrt{0.54} \approx 0.735$$

$$\|y\| = \sqrt{0.6^2 + 0.3^2 + 0.4^2} = \sqrt{0.61} \approx 0.781$$

$$\cos(\theta) = \frac{0.52}{0.735 \times 0.781} \approx 0.90$$

- $\cos(\theta) \approx 0.90 \Rightarrow$ very similar meaning

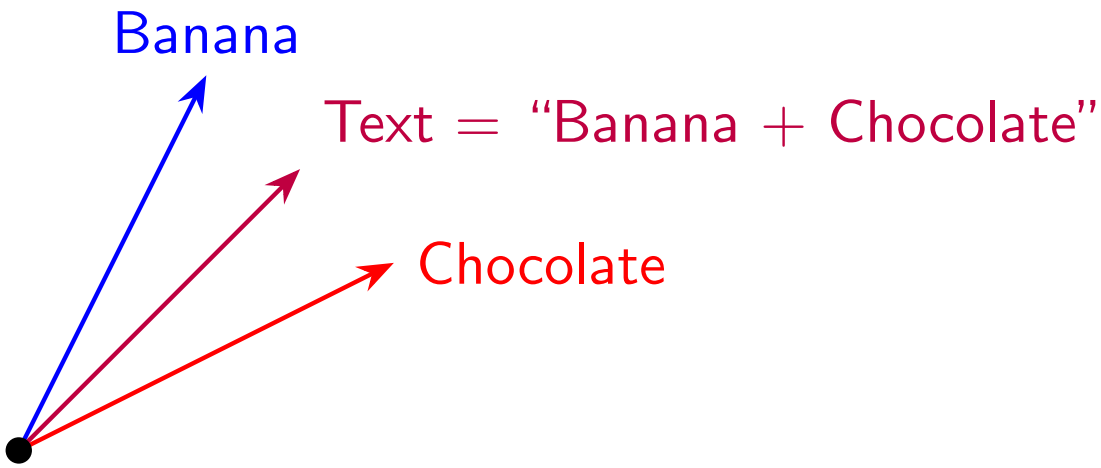
Visualizing Word Similarity in 1D



From Words to Texts

- Each word is represented as a vector in high-dimensional space.
- To represent a text (sentence, document):
  - Combine (e.g., average, sum, weighted average with TF–IDF) the word vectors.
- Result: a single **text vector** in the same embedding space.
- Then, we can compute similarity between texts just like between words.

Combining Word Vectors into Text Vectors



Similarity Between Texts

- Once we have text vectors, we can compute similarity:

$$\cos(\theta) = \frac{v_{\text{text A}} \cdot v_{\text{text B}}}{\|v_{\text{text A}}\| \, \|v_{\text{text B}}\|}$$

- Example:
  - Text A = "Banana Chocolate" $\Rightarrow v_A = (1.5, 1.5)$
  - Text B = "Banana Car" $\Rightarrow v_B = (3, 0.5)$
- Then:

$$\cos(\theta_{A,B}) = \frac{1.5 \cdot 3 + 1.5 \cdot 0.5}{\sqrt{1.5^2 + 1.5^2}\,\sqrt{3^2 + 0.5^2}} \approx 0.82$$

- $\Rightarrow$ Text A and Text B are fairly similar.