THE UNIVERSITY *of* EDINBURGH

Text Technologies for Data Science

INFR11145

Comparing Text Corpora

Instructor:

Björn Ross

08-Nov-2023

Pre-Lecture

- Today
  - Lecture: Comparing Text Corpora 1 & 2
- **No lecture** next week (15 November)!
- 22 November
  - Lecture: Text Classification 1 & 2
  - CW2: IR Eval, Comparing Corpora, Text Classification

Initial Text Analysis

- Scenario: you are given access to a new dataset
  - 2 corpora, each contains thousands of plain text files
  - You want to <u>understand</u> and <u>quantify</u>:
    - What is the *content* of these documents? What are they *about*?
    - How does the content of these corpora *differ*?

- What are some things you might try first?

Initial Text Analysis

- Scenario: you are given access to a new dataset
  - 2 corpora, each contains thousands of plain text files
  - You want to understand and quantify:
    - What is the *content* of these documents? What are they *about*?
    - How does the content of these corpora *differ*?

- What are some things you might try first?
  - Read some examples
  - Language identification
  - Compute basic statistics:
    - Number of words, most frequent words, avg. words per document, …
  - Build word clouds
  - …

Lecture Objectives

- Analyze text corpora
  - Content analysis background
  - Word-level differences
  - Dictionaries and Lexicons
  - Topic modeling
  - Annotation + classification

Content Analysis

- Goal: given some documents determine
  - What are the types of content present? (themes/topics)
  - Which documents contain which topics?
- Traditionally a manual process
  1. Read a subset of documents, define themes/topics
  2. Determine consistent coding* methodology
  3. Read all documents and label them according to codes
  4. Check agreement between human coders
  5. Settle disagreements via a third-party
  6. Analyze resulting annotations

Content Analysis

- Can this process be automated?
  - Yes, to an extent
- *Should* this process be automated?
  - Humans are better than machines at this task (for now?)
  - Computers are *much*, *much* faster
    - Avg. human reading speed: 250 wpm
    - Assume 1K words/document, 50K documents…
      - Average person needs > 4 months to read
      - This is a ***relatively small*** corpus for modern NLP
    - Modern computers can process millions of words/second

Automated Content Analysis

- Single corpus/class
  - Word frequency analysis
  - Dictionaries & Lexicons
  - Topic modelling
- Multiple corpora/classes
  - Word-level differences
  - Dominance Scores
  - Topic-level differences

Word Level Analysis

Word frequency analysis

- Very simple starting point

1. Preprocess as usual (lowercasing? stemming?...)
2. Count words
3. Normalize by document length
4. Average across all documents

THE UNIVERSITY of EDINBURGH

Word-level Differences

- Which words best characterize set of documents (such as a corpus or class)?
  - Need a reference corpus
- Some methods to do this:
  - Mutual information
  - Chi squared

- Can also be used for *feature selection*

Mutual Information

- I(X;Y)
  - How much can I learn about Y by observing X?
  - Is the same as *information gain*
  - Is **not** the same as *pointwise mutual information*
- We want to learn about important words in our class
- What should X and Y be?
  - X = U = document contains term t (Boolean)
  - Y = C = class is the target class (Boolean)

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U=e_t, C=e_c) \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t)P(C=e_c)}$$

Mutual Information

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

- Given a corpus and a term, how do we estimate the probability of this term appearing in a random document in the corpus?

Source: Manning, Raghavan, and Schütze, 2008

Mutual Information

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U=e_t, C=e_c) \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t)P(C=e_c)}$$

- Given count data for 2 classes, can be computed as:

$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}}$$

Source: Manning, Raghavan, and Schütze, 2008

Mutual Information

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

- Example:
  - What is I(U;C) given these values?

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{export}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{export}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

Example: Manning, Raghavan, and Schütze, 2008

Mutual Information

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{export}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{export}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

$$
\begin{aligned}
I(U;C) &= \frac{49}{801{,}948}\log_2\frac{801{,}948\cdot 49}{(49+27{,}652)(49+141)} \\
&+ \frac{141}{801{,}948}\log_2\frac{801{,}948\cdot 141}{(141+774{,}106)(49+141)} \\
&+ \frac{27{,}652}{801{,}948}\log_2\frac{801{,}948\cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)} \\
&+ \frac{774{,}106}{801{,}948}\log_2\frac{801{,}948\cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)} \\
&\approx 0.0001105
\end{aligned}
$$

Example: Manning, Raghavan, and Schütze, 2008

Mutual Information for News Data

UK

london	0.1925
uk	0.0755
british	0.0596
stg	0.0555
britain	0.0469
plc	0.0357
england	0.0238
pence	0.0212
pounds	0.0149
english	0.0126

China

china	0.0997
chinese	0.0523
beijing	0.0444
yuan	0.0344
shanghai	0.0292
hong	0.0198
kong	0.0195
xinhua	0.0155
province	0.0117
taiwan	0.0108

poultry

poultry	0.0013
meat	0.0008
chicken	0.0006
agriculture	0.0005
avian	0.0004
broiler	0.0003
veterinary	0.0003
birds	0.0003
inspection	0.0003
pathogenic	0.0003

coffee

coffee	0.0111
bags	0.0042
growers	0.0025
kg	0.0019
colombia	0.0018
brazil	0.0016
export	0.0014
exporters	0.0013
exports	0.0013
crop	0.0012

elections

election	0.0519
elections	0.0342
polls	0.0339
voters	0.0315
party	0.0303
vote	0.0299
poll	0.0225
candidate	0.0202
campaign	0.0202
democratic	0.0198

sports

soccer	0.0681
cup	0.0515
match	0.0441
matches	0.0408
played	0.0388
league	0.0386
beat	0.0301
game	0.0299
games	0.0284
team	0.0264

Example: Manning, Raghavan, and Schütze, 2008

Chi-squared

- Hypothesis testing approach
- H_0: Term appearance is independent from a document's class
  - i.e., $P(U = 1, C = 1) = P(U = 1)P(C = 1)$
- Compute:

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

- Or to directly plug in values like before:

$$X^2(\mathbb{D}, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}$$

Chi-squared

$$X^2(\mathbb{D}, t, c) = \frac{(N_{11} + N_{10} + N_{01} + N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11} + N_{01}) \times (N_{11} + N_{10}) \times (N_{10} + N_{00}) \times (N_{01} + N_{00})}$$

- Example
  - What is the value of X^2 given the example data?

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{export}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{export}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

Chi-squared

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{export}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{export}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

$$\frac{(49 + 27652 + 141 + 774106) \times (49 \cdot 774106 - 27652 \cdot 141)^2}{(49 + 141) \times (49 + 27652) \times (27652 + 774106) \times (141 + 774106)} \approx 284$$

Dictionaries and Lexicons

Dictionaries and Lexicons

- What if we know what we are looking for?
- Dictionaries (lexicons) are prebuilt mappings
  - Category -> word list
  - E.g., a tiny sentiment lexicon:
    - Positive: good, great, happy, amazing, wonderful, best, incredible
    - Negative: terrible, horrible, bad, awful, nasty, gross, worst, poor

- Domain can be important
  - "***unpredictable*** movie plot" ✓
  - "***unpredictable*** coffee pot" ✗

Dictionaries and Lexicons

- How to get a score per category?

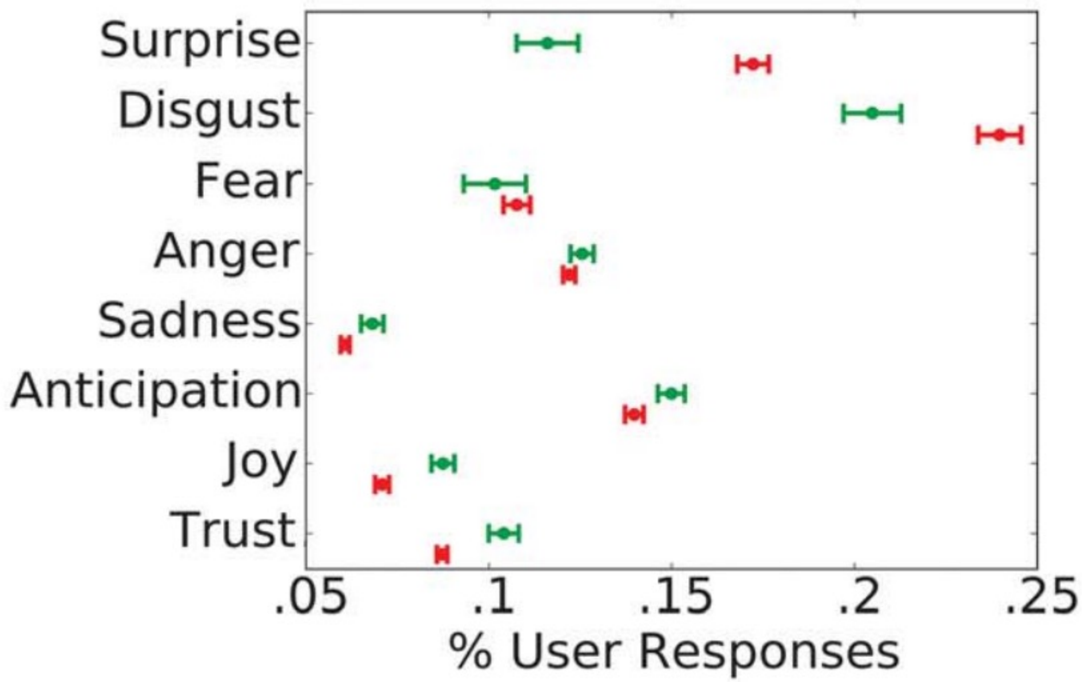
$$\frac{num\_dictionary\_words\_in\_document}{num\_total\_words\_in\_document}$$

- That's it!
- Can also be used as machine learning features
- A more advanced approaches to quantifying categories (optional reading)
  - https://www.ncbi.nlm.nih.gov/pubmed/28364281

Some Dictionaries

- LIWC (Pennebaker et al. 2015)
- General Inquirer (Stone 1997)
- Roget's Thesaurus Categories
- VADER (Hutto and Gilbert, 2014)
- Sentiwordnet (Esuli and Sebastiani 2006)
- Wordnet Domains (Magnini and Cavaglia, 2000)
- EmoLex (Mohammad and Turney, 2010)
- Empath (Fast et al., 2016)
- Personal Values Lexicon (Wilson et al., 2018)
- …

Reactions to Rumor Tweets with EmoLex



Red = reactions to false rumors
Green = reactions to true rumors

Vosoughi, Roy, and Aral, 2018

Dominance Scores

- The dominance score for a category w.r.t. a corpus:

$$\frac{category\_score\_in\_target\_corpus}{category\_score\_in\_background\_corpus}$$

  - From Mihalcea and Pulman, 2009

THE UNIVERSITY of EDINBURGH

LIWC category dominance scores

Truthful			
Interviews		Trials	
Class	Score	Class	Score
Metaphor	2.98	You	3.99
Money	2.74	Family	3.07
Inhibition	2.74	Home	2.45
Home	2.13	Humans	1.87
Humans	2.02	Posemo	1.81
Family	1.96	Insight	1.64

Deceptive			
Interviews		Trials	
Class	Score	Class	Score
Assent	4.81	Anger	2.61
Past	2.59	Anxiety	2.61
Sexual	2.00	Certain	2.28
Other	1.87	Death	1.96
Motion	1.68	Physical	1.77
Negemo	1.44	Negemo	1.52

Pérez-Rosas et al, 2015

Topic Level Analysis

Intro to Topic Modelling

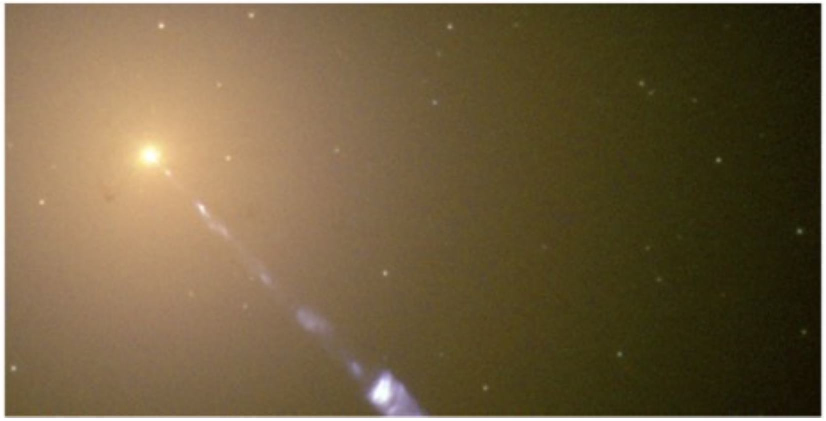
- Goals are similar to traditional content analysis:
  - What are the main themes/topics in this corpus?
  - Which documents contain which topics?

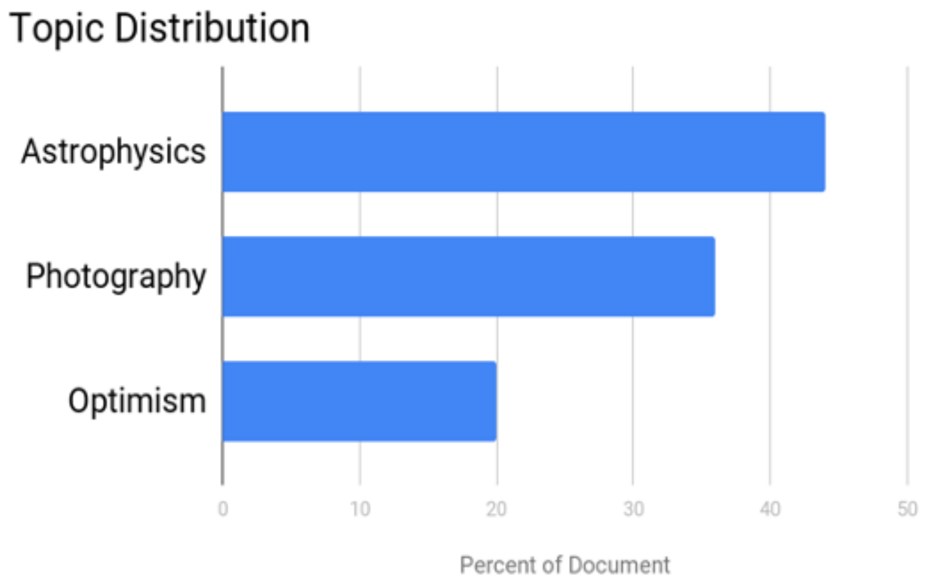
Topic Models

The New York Times

Expected Soon: First-Ever Photo of a Black Hole

Have astronomers finally recorded an image of a black hole? The world will know on Wednesday.

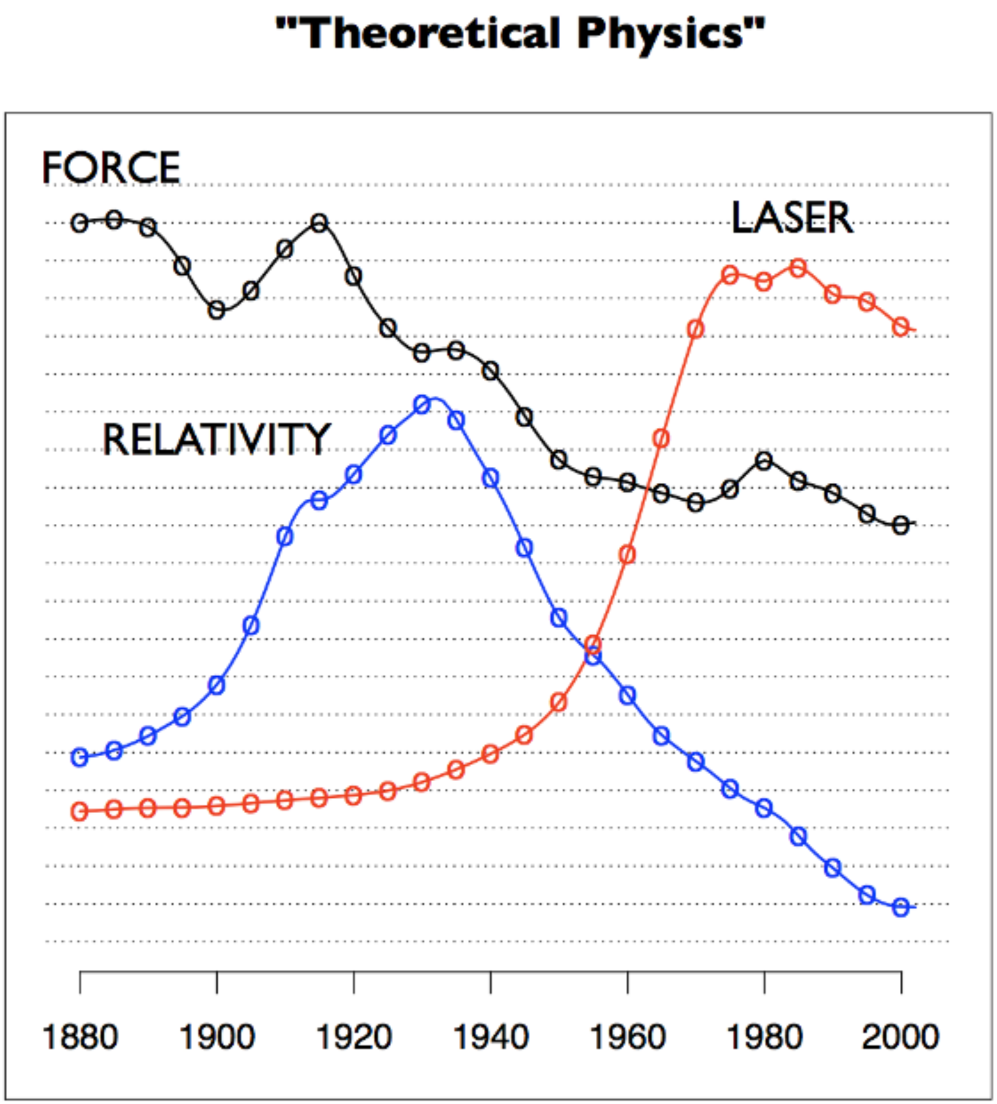


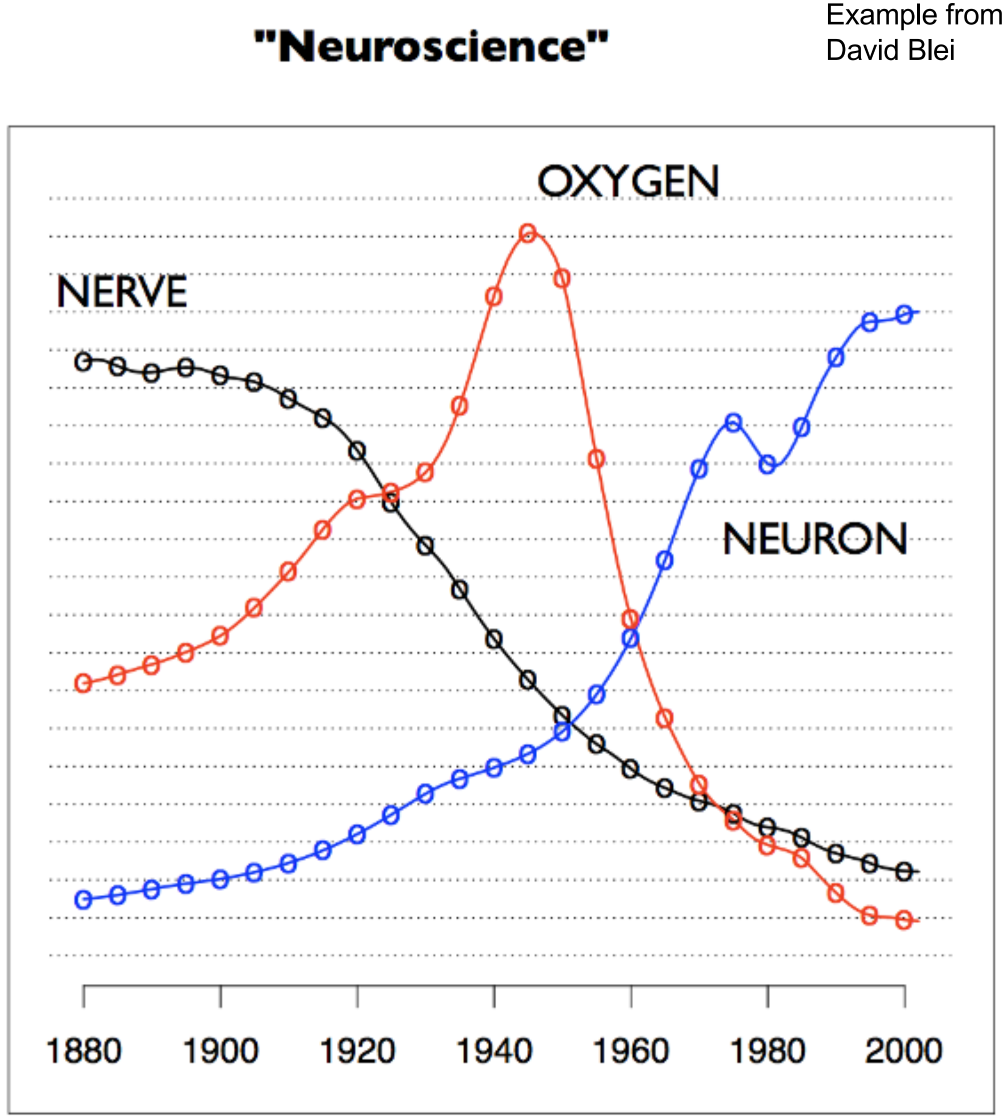


human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

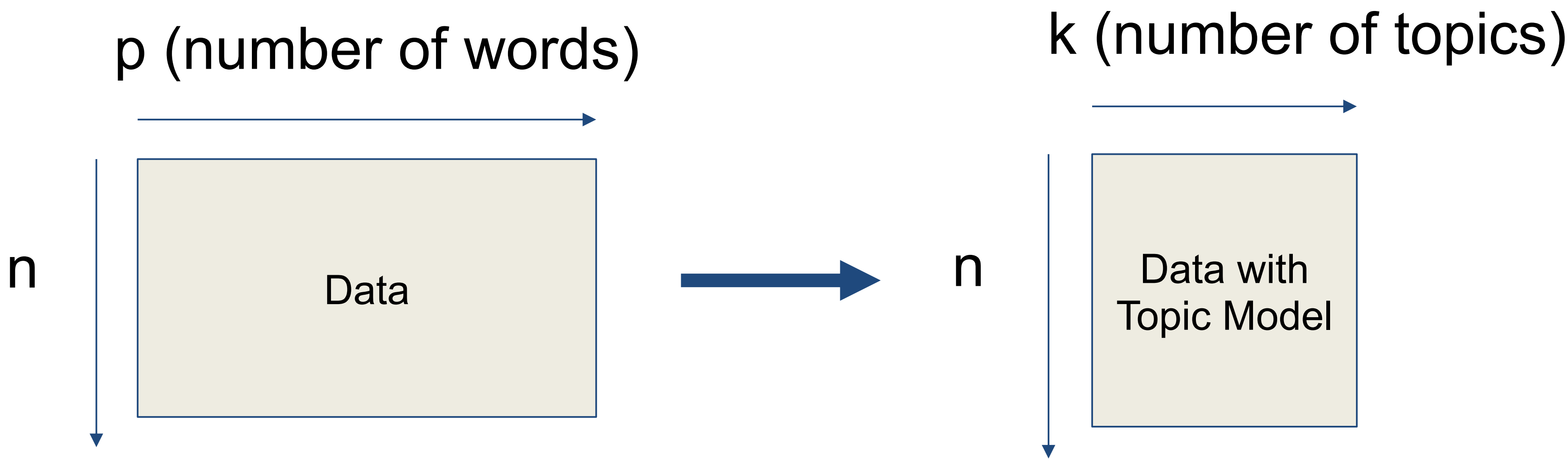
Example from David Blei

Example from
David Blei

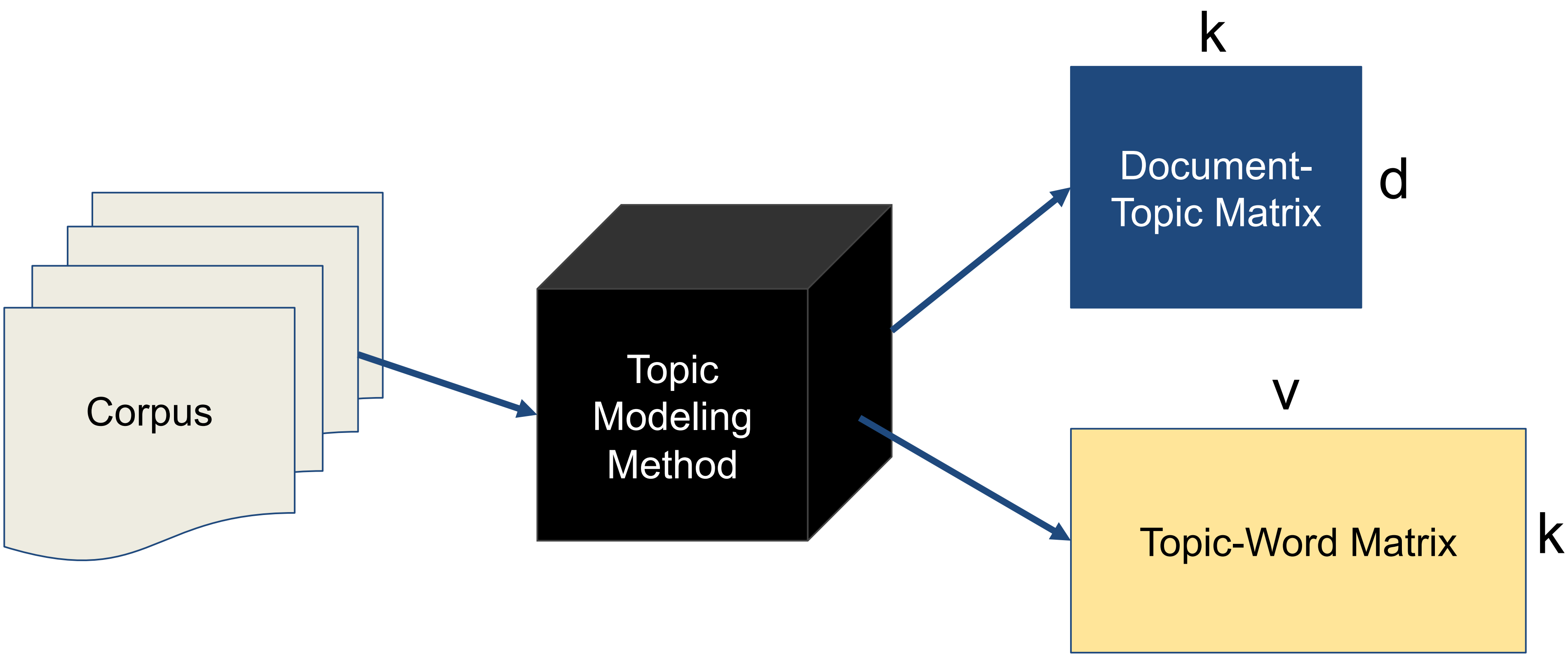


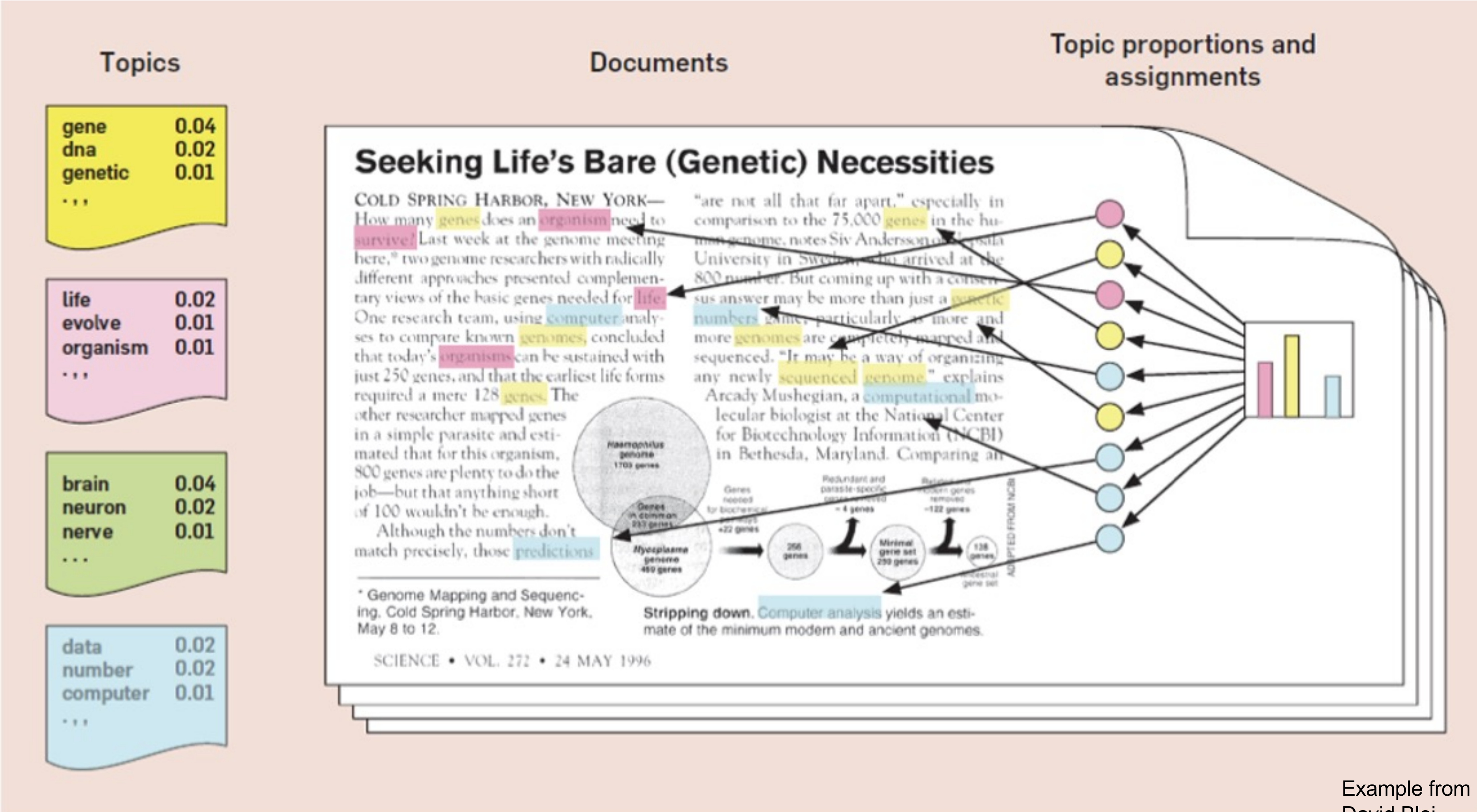


Dimensionality Reduction



Topic Modeling





Example from
David Blei

Topic Models

- Most often used for text data, but can also be applied in other settings:
  - Bioinformatics (Liu et al. 2016)
  - Computer code (McBurney et al. 2014)
  - Music (Hu and Saul 2009)
  - Network data (Cha and Cho 2014)

Topic Modeling Methods

- Most popular: Latent Dirichlet Allocation (LDA)
  - Introduced by David Blei, Andrew Ng, and Michael Jordan (2003)

- Other methods include
  - pLSI
  - PCA-based methods
  - Non-negative matrix factorization
  - Deep learning based topic modeling
  - ...

Topic Modeling Methods

- Most popular: Latent Dirichlet Allocation (LDA)
  - Introduced by David Blei, Andrew Ng, and Michael Jordan (2003)

- Other methods include
  - pLSI
  - PCA-based methods
  - Non-negative matrix factorization
  - Deep learning based topic modeling
  - ...

Latent Dirichlet Allocation (LDA)

- More details coming up in next lecture…