Lecture 16: Text Mining

EMAT31530/March 2018/Raul Santos-Rodriguez

Have a look at ...

... Christopher D. Manning and Hinrich Schtze (1999). Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA.

... David Blei (2012). Probabilistic topic models. Communications of the ACM 55(4): 77-84.

... Python: `http://scikit-learn.org/`

... NLTK: `http://www.nltk.org/`

Outline

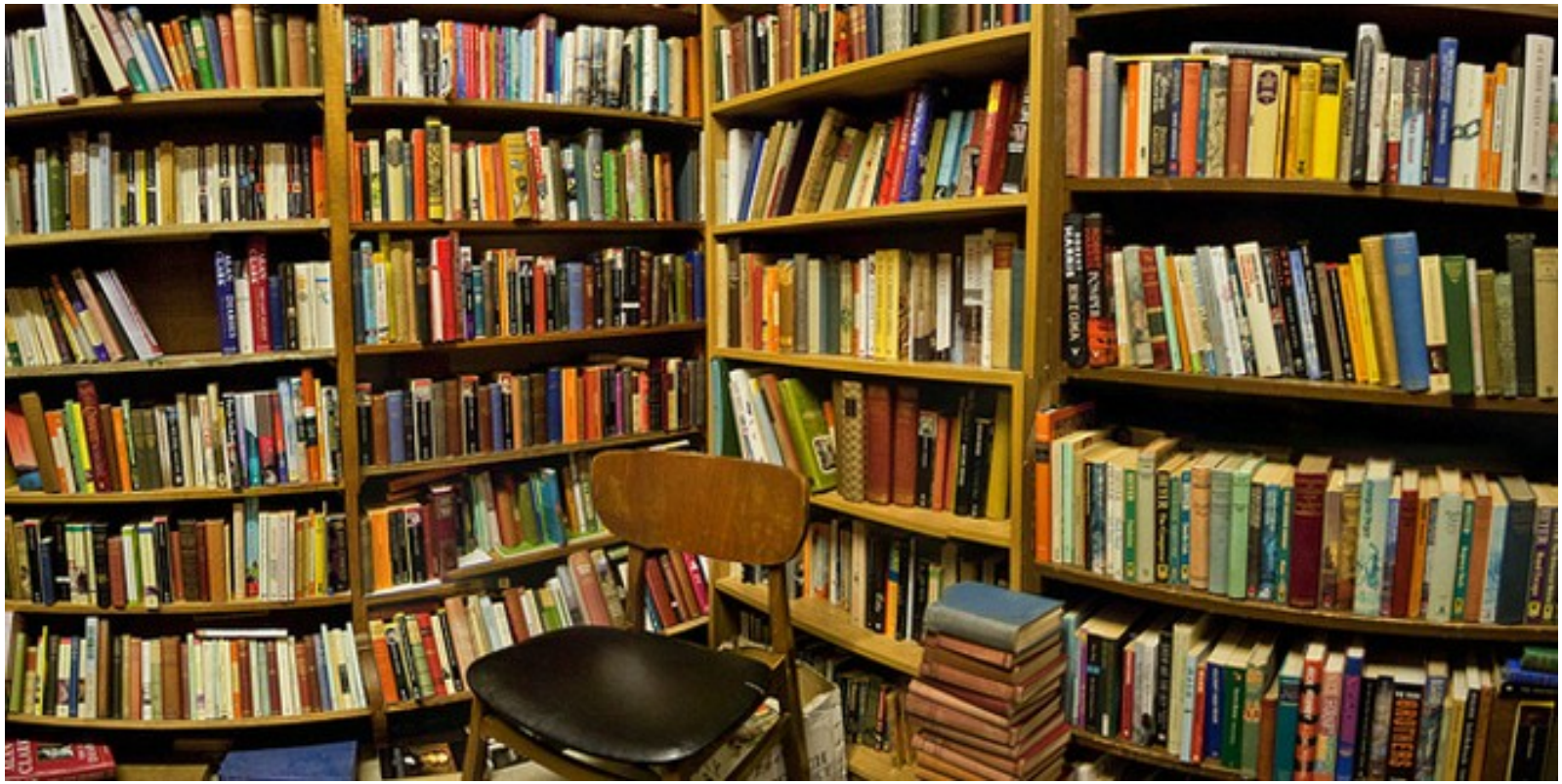
This lecture presents an introduction to text mining and natural language analysis. You will study:

- Main tasks in text mining.
- Text clustering.
- Topic models.

The material in this lecture is based on a talk given by Dr. Jefrey Lijffijt in Intro to AI 2015

Why text mining

The amount of text published on paper, on the web, and even within companies is inconceivably large!



Text mining

Automated methods to **find**, **extract**, and **link** information from documents

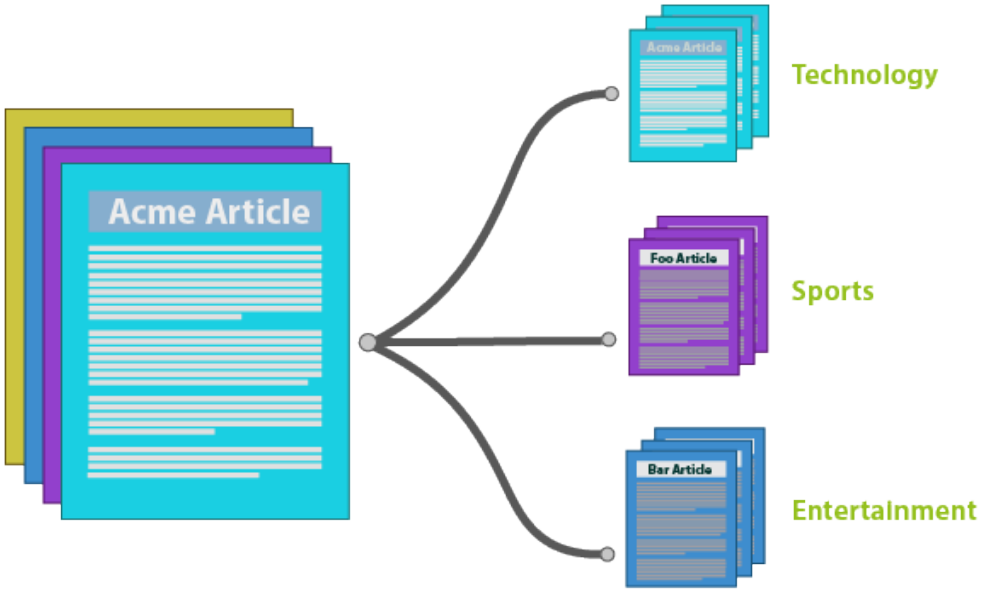
Main tasks

Classification
categorise texts into classes (given a set of classes)

Clustering
categorise texts into classes (not given any set of classes)

Sentiment analysis
determine the sentiment/attitude of texts

Keyword analysis
find the most important terms in texts



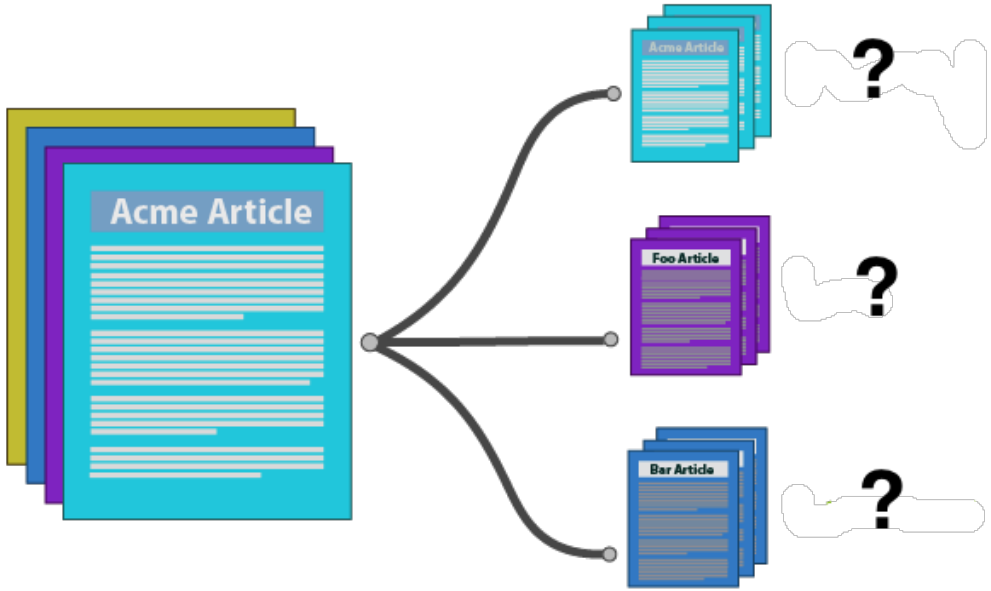
Main tasks

Classification
categorise texts into classes (given a set of classes)

Clustering
categorise texts into classes (not given any set of classes)

Sentiment analysis
determine the sentiment/attitude of texts

Keyword analysis
find the most important terms in texts



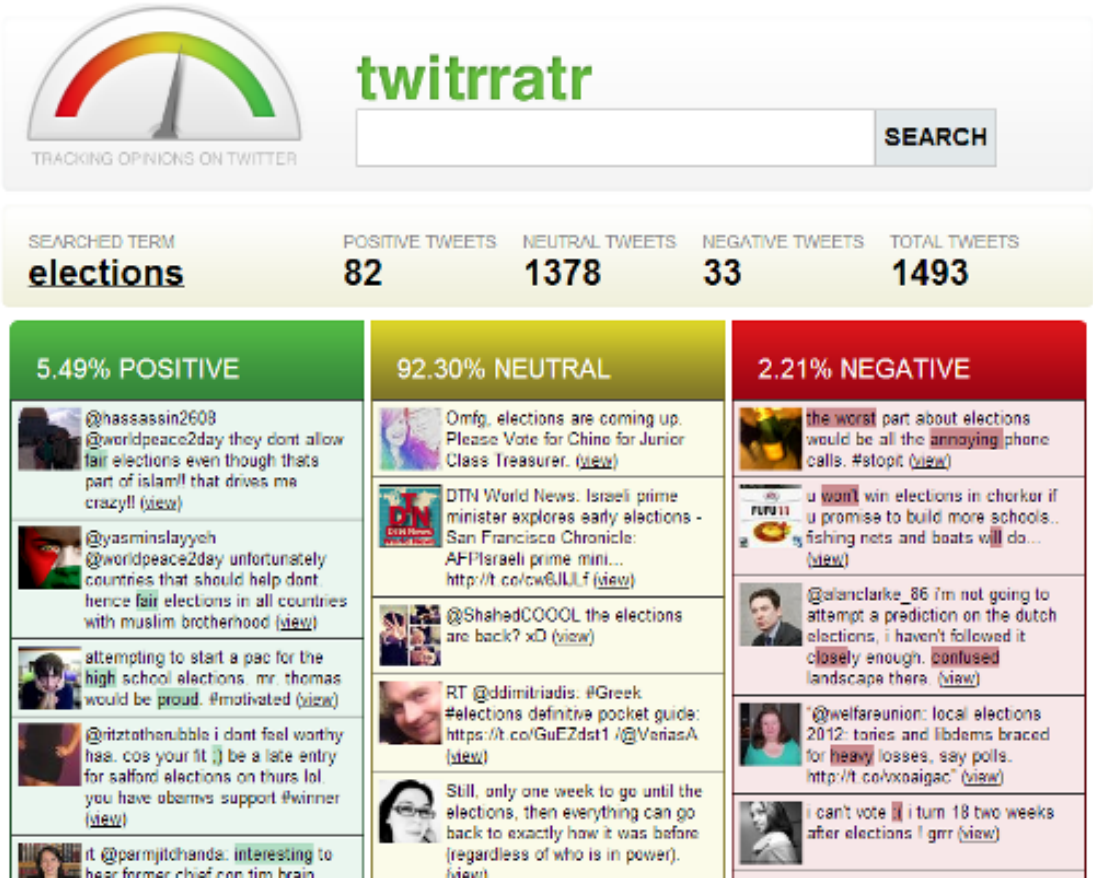
Main tasks

Classification
categorise texts into classes (given a set of classes)

Clustering
categorise texts into classes (not given any set of classes)

Sentiment analysis
determine the sentiment/attitude of texts

Keyword analysis
find the most important terms in texts



Main tasks

Classification
categorise texts into classes (given a set of classes)

Clustering
categorise texts into classes (not given any set of classes)

Sentiment analysis
determine the sentiment/attitude of texts

Keyword analysis
find the most important terms in texts

Main tasks

Summarisation
categorise texts into classes (give a brief summary of texts)

Retrieval
find the most relevant texts to a query

Question-answering
answer a given question

Language modelling
uncover structure and semantics of texts

https://www.youtube.com/watch?v=WFR3lOm_xhE

Main tasks

Summarisation
categorise texts into classes (give a brief summary of texts)

Retrieval
find the most relevant texts to a query

Question-answering
answer a given question

Language modelling
uncover structure and semantics of texts

https://www.youtube.com/watch?v=WFR3lOm_xhE

Main tasks

Summarisation
categorise texts into classes (give a brief summary of texts)

Retrieval
find the most relevant texts to a query

Question-answering
answer a given question

Language modelling
uncover structure and semantics of texts

https://www.youtube.com/watch?v=WFR3lOm_xhE

Main tasks

Summarisation
categorise texts into classes (give a brief summary of texts)

Retrieval
find the most relevant texts to a query

Question-answering
answer a given question

Language modelling
uncover structure and semantics of texts

`https://www.youtube.com/watch?v=WFR3lOm_xhE`

Related domains

Text mining
refers to the process of deriving high-quality information from text

Information retrieval (IR)
is the activity of obtaining information resources relevant to an information need from a collection of information resources

Natural language processing (NLP)
is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages

Computational linguistics
is an interdisciplinary field concerned with the statistical or rule-based modelling of natural language from a computational perspective

---

Wikipedia

Today

Text clustering and topic models

[Why] useful to categorise texts and to uncover structure in text corpora

[Problem] how to represent text? What are the relevant features?

Solution

Vector-space (bag-of-words) model

	Word 1	Word 2	Word 3	...
Text 1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$	
Text 2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$	...
Text 3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$	
...		...		

Simple text clustering

- Clustering with **k-means** algorithm and **cosine similarity**

- [Idea] two texts are similar if the frequencies at which words occur are similar

$$s(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1 \cdot \mathbf{w}_2}{||\mathbf{w}_1|| ||\mathbf{w}_2||}$$

$s \in [0, 1]$ (since $\mathbf{w}_{i,j} \geq 0$)

- Widely used in text mining

Demo

- Reuters-21578
- 8300 (categorised) newswire articles
- Clustering is a single command in Matlab!
- Data (original and processed .mat):

  `http://www.daviddlewis.com/resources/testcollections/reuters21578/`
  `http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html`

More advanced clustering

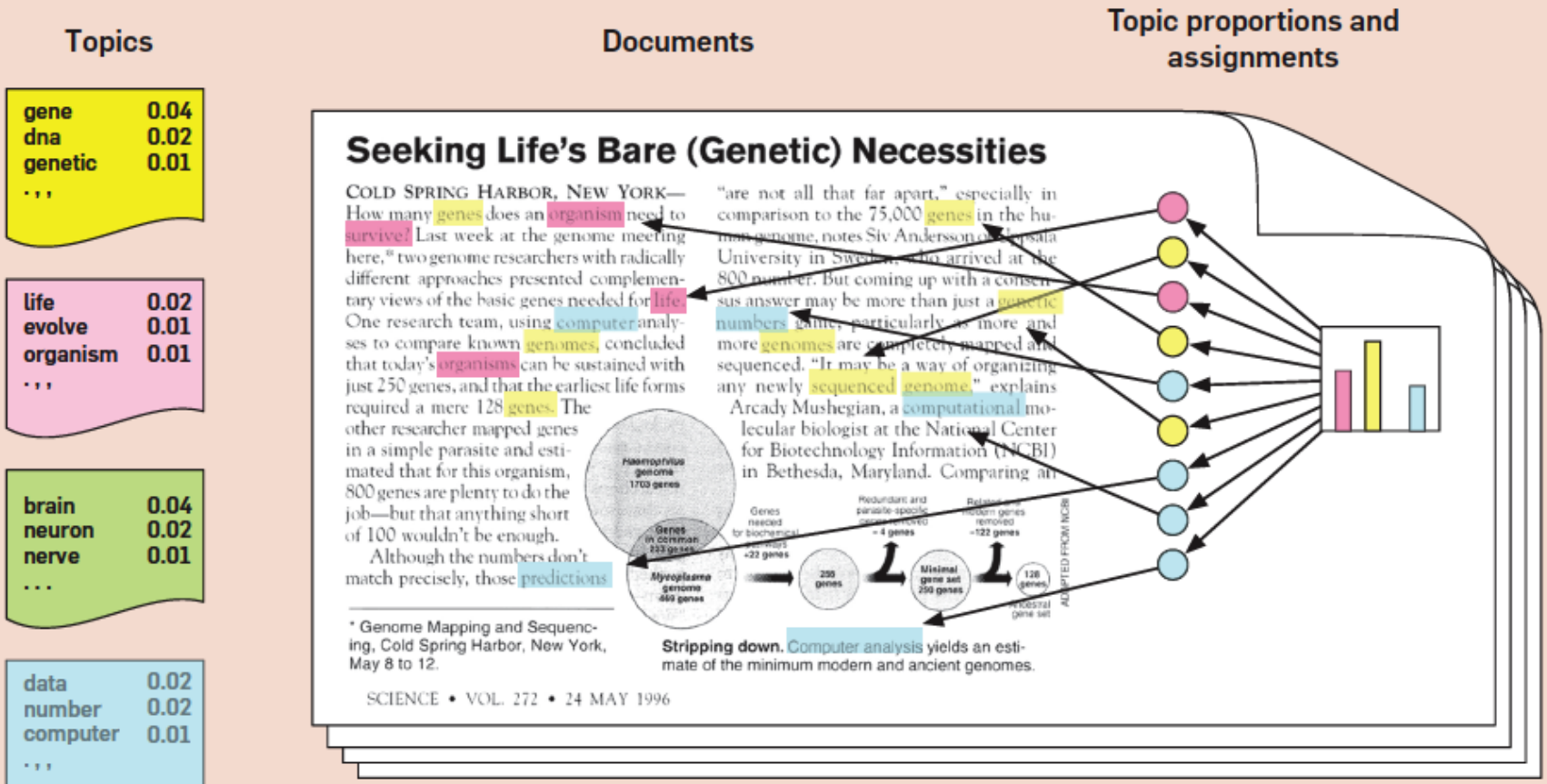
Latent Dirichlet Allocation (LDA), also known as topic modelling

[Idea] texts are a weighted mix of topics

David Blei (2012). Probabilistic topic models. CACM 55(4): 77-84.

More advanced clustering

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



David Blei (2012). Probabilistic topic models. CACM 55(4): 77-84.

More advanced clustering

Topics are probability distributions over words

The distribution defines how often each word occurs, given that the topic is discussed

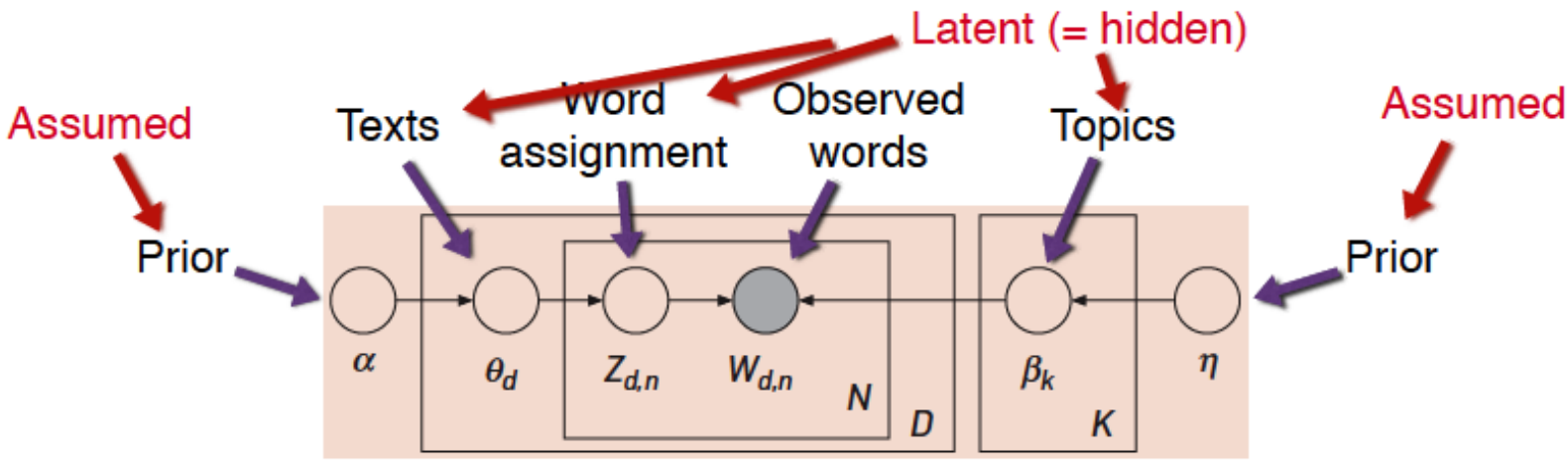
Texts are probability distributions over topics

The distribution defines how often a word is due to a topic

More advanced clustering

- For each word in a text, we can compute how probable it is that it belongs to a certain topic

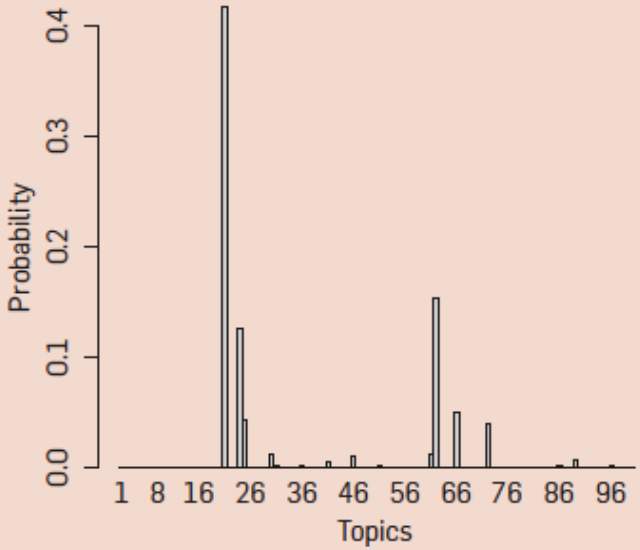
- Given the topic probability and the topics, we can compute the likelihood of the document



The **optimisation problem** is to find the posterior distributions for the topics and the texts (see article)

More advanced clustering

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.

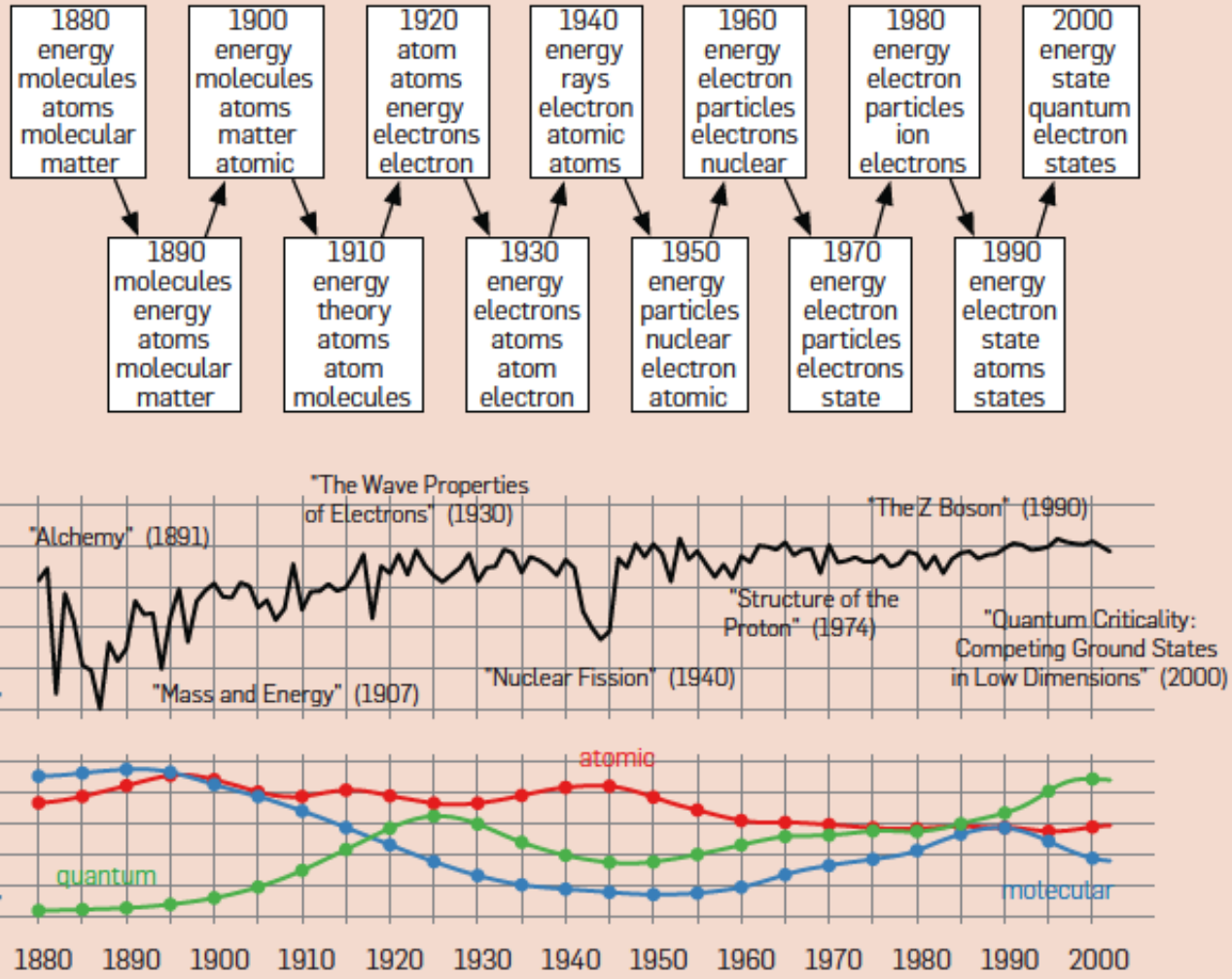


"Genetics"	"Evolution"	"Disease"	"Computers"
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	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

David Blei (2012). Probabilistic topic models. CACM 55(4): 77-84.

More advanced clustering

Figure 5. Two topics from a dynamic topic model. This model was fit to *Science* from 1880 to 2002. We have illustrated the top words at each decade.



Summary

Text mining is concerned with automated methods to find, extract, and link information from text

Text clustering and **topic models** help us

- **Organise** text corpora
- **Find** relevant documents
- **Uncover** relations between documents

Next lecture

We will discuss Markov Decision Processes!