Topic models for text analysis

Clint P. George

University of Florida Informatics Institute

March 18, 2017

A typical information retrieval problem.

PC: Chakansi-Jokic (2016)

Example

Suppose, we search for the keyword *computers* in a document collection

We may miss documents that do not have *computers* and contain *PC*, *laptop*, *desktop*, etc.

A typical information retrieval problem.

PC: Chakansi-Jokic (2016)

Example

Suppose, we search for the keyword *computers* in a document collection

We may miss documents that do not have *computers* and contain *PC*, *laptop*, *desktop*, etc.

- #1: Synonymy—words with similar meaning

A typical information retrieval problem.

PC: Chakansi-Jokic (2016)

Example

Suppose, we search for the keyword *chair* in a collection

We may get documents that contain "the *chair* of the board" and "the *chair* maker"

A typical information retrieval problem.

PC: Chakansi-Jokic (2016)

Example

Suppose, we search for the keyword *chair* in a collection

We may get documents that contain "the *chair* of the board" and "the *chair* maker"

- #2: Polysemy—words with multiple meanings

A typical information retrieval problem.

PC: Chakansi-Jokic (2016)

When we search for documents we usually look for concepts or topics, not keywords.

A typical information retrieval problem.

PC: Chakansi-Jokic (2016)

When we search for documents we usually look for concepts or topics, not keywords.

Question

How can we identify the underlying topics in a corpus?

How do we represent documents?

This is the first step in any information retrieval problem.

A typical approach:

- We consider documents as bags-of-words—ignores any word ordering in a document
- We take the appearance frequencies of unique words as document features or predictor variables

Example

Vector Space Models make these assumptions.

Implementing keyword-based search

An approach is via **Vector Space Modeling**

- Convert a corpus of D documents and V vocabulary terms into a **term-document** $(V \times D)$ matrix
- Translate both documents and user keywords into vectors in vector space
- Define similarity between these vectors,
  e.g., via cosine similarity—small angle $\equiv$ large cosine $\equiv$ similar

TF-IDF

Term-Frequency Inverse-Document-Frequency (Salton et al. 1975)

$$\text{tf-idf}_{dt} = \text{tf}_{dt} \times \log\left(\frac{D}{\text{df}_t}\right),\ d = 1, 2, \ldots, D; t = 1, 2, \ldots, V$$

where tf_{dt} = the frequency of term t in document d and df_t = the number of documents where term t appears

TF-IDF

Term-Frequency Inverse-Document-Frequency (Salton et al. 1975)

$$\text{tf-idf}_{dt} = \text{tf}_{dt} \times \log\left(\frac{D}{\text{df}_t}\right),\ d = 1, 2, \ldots, D; t = 1, 2, \ldots, V$$

where tf_{dt} = the frequency of term t in document d and df_t = the number of documents where term t appears

Advantages:

- Can represent the importance of a document word in the collection
- Can handle common terms in the corpus—via IDF term

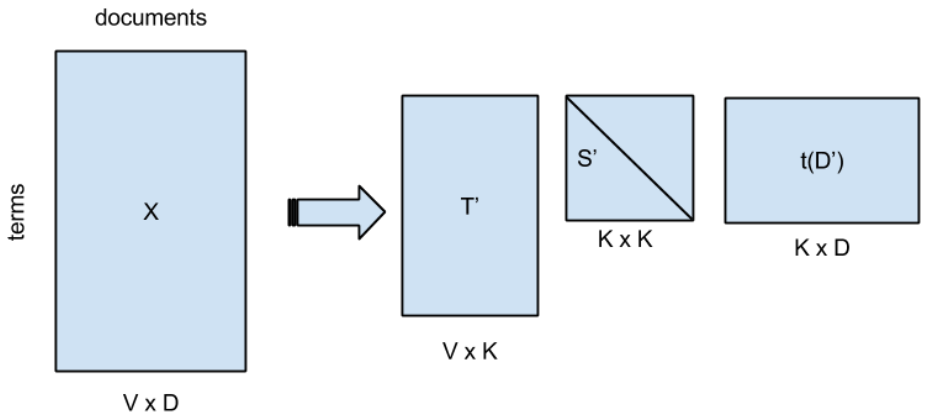
TF-IDF / VSM: limitations

- *Synonymy* can cause small cosine scores between documents that are related
- *Polysemy* can cause large cosine similarity between documents that are unrelated
- *Dimensionality*, D and V, can be very large

Latent Semantic Analysis (LSA)[1]

We approximate the document matrix using Singular Value Decomposition:

$$X \approx R = T \times S \times D^{\mathsf{T}}$$



- S contains K largest *singular values* of X
- T represents correlation between terms over documents
- D represents correlation between documents over terms

[1] Deerwester et al. (1990)

Latent Semantic Analysis (LSA)

Advantages:

- Identifies a *linear subspace* in the vocabulary space—significant compression in large collections
- Handles synonymy to some extent
- Efficient SVD implementations are available

Latent Semantic Analysis (LSA)

Advantages:

- Identifies a *linear subspace* in the vocabulary space—significant compression in large collections
- Handles synonymy to some extent
- Efficient SVD implementations are available

Limitations:

- A linear model—may not find nonlinear dependencies between words or documents
- Identified features are difficult to interpret

Probabilistic topic modeling

Latent Dirichlet Allocation (LDA, Blei et al. 2003) is a probabilistic, generative, topic model

LDA assumes

- Documents as bags of words
- A topic as a distribution over a fixed vocabulary
- Words are generated from document specific topic distributions[2]

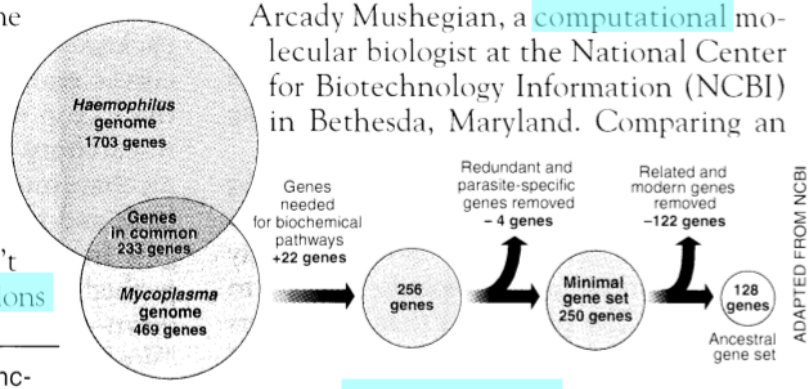
[2] Multinomial sampling experiment

Latent Dirichlet Allocation: intuition[3]

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

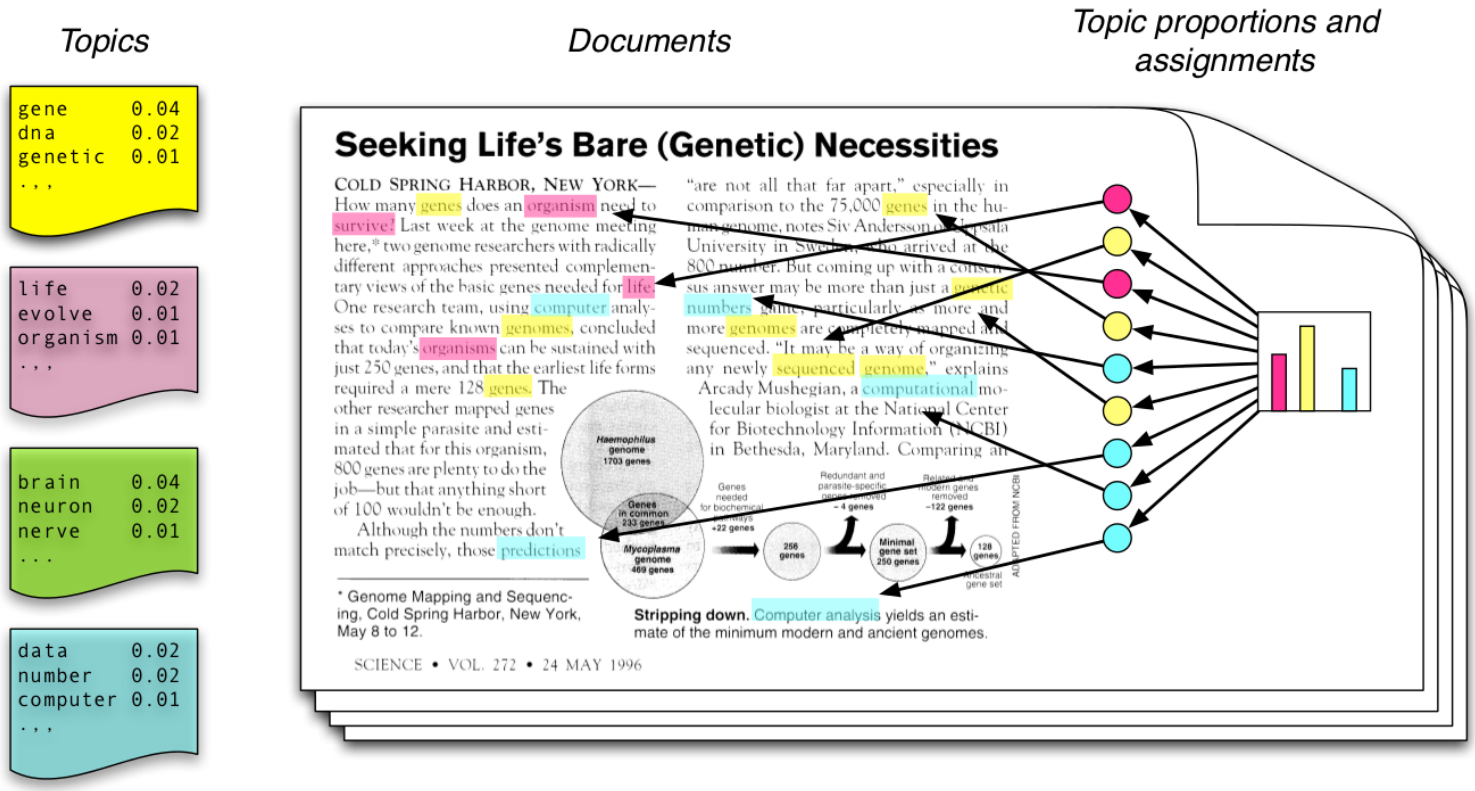
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Documents are formed from mixtures of topics

[3] Blei (2009, MLSS)

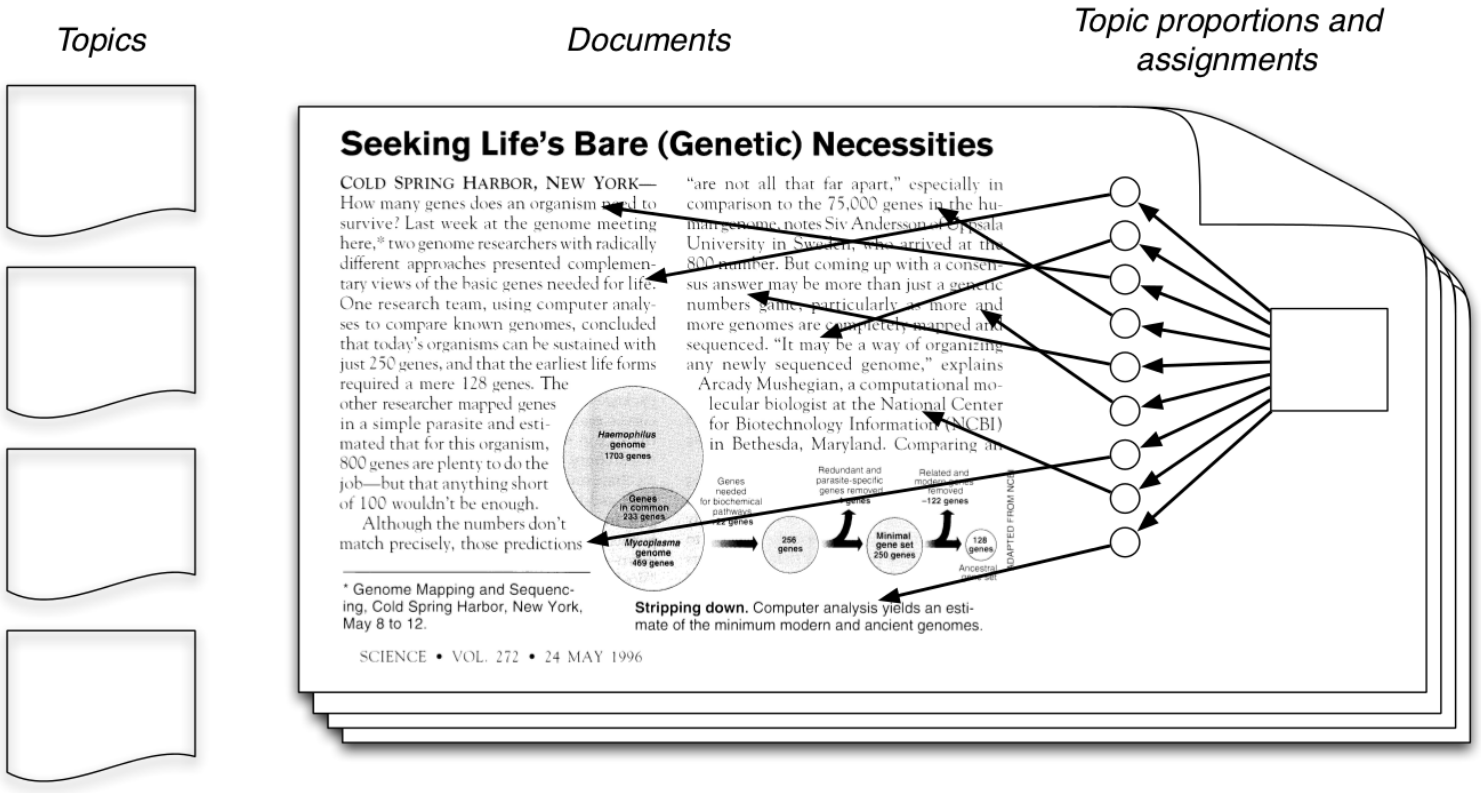
LDA: model visualization[4]



- Each document is a mixture (histogram) of topics (left)
- Each word is generated from one of those topics

[4] Blei (2009, MLSS)

LDA: inference visualization[5]



- In real life, we only observe documents and their words
- Our goal is to infer the underlying topic structure

[5] Blei (2009, MLSS)

What can we do with the LDA model?

Given the corpus, one can infer the hidden structures:

- Per-word topic assignment
- Per-document topic proportions—dimensionality reduction
- Per-corpus topic distributions—better representations

We can then use them for information retrieval, document clustering and exploration, etc.

An example inference with LDA

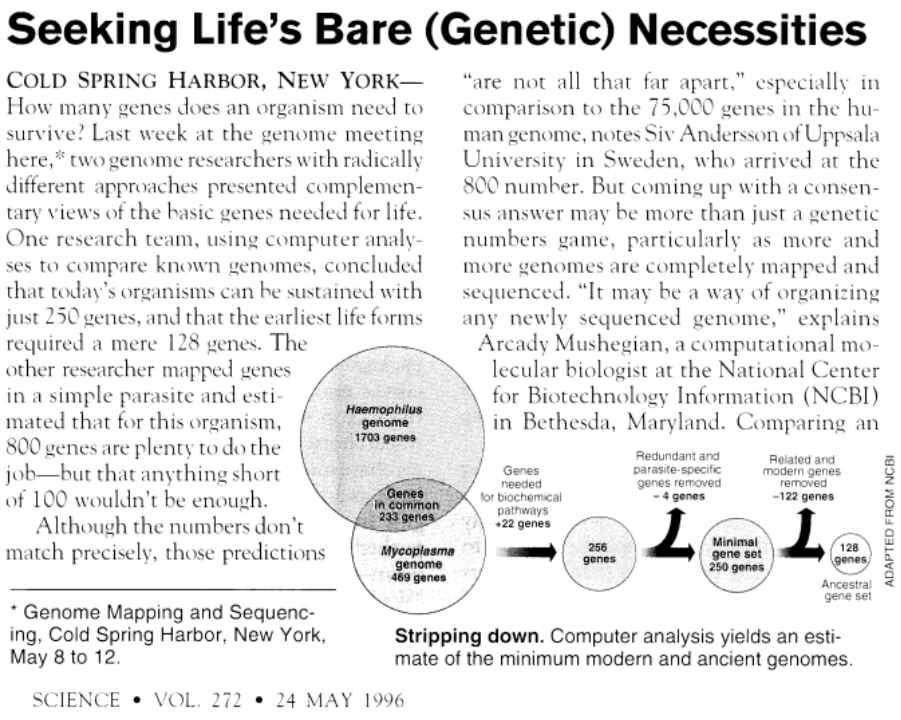
We used a collection of OCR'd Science magazine (1990-2000) articles

- $\sim 17{,}000$ documents
- $\sim 20{,}000$ unique words in the vocabulary

We built a 100-topic LDA model using the *variational inference* algorithm (Blei et al. 2003)[6]

[6] Blei (2011, KDD)

Inference with LDA: An example article[7]



Seeking Life's Bare (Genetic) Necessities

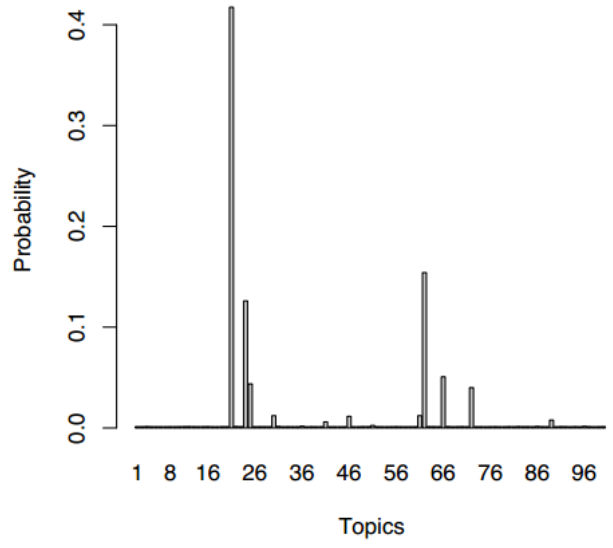
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

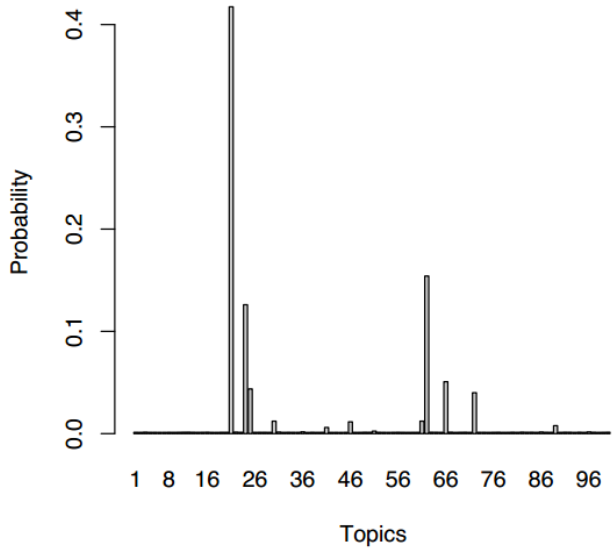


We can get

- Per-word topic assignment
- Per-document topic proportions—dimensionality reduction

[7] Blei (2011, KDD)

Inference with LDA: What are the topics?[8]



"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

Per-corpus topic distributions—15 most probable words from the most frequent topics found in the article is shown here

[8] Blei (2011, KDD)

Thank you! Questions?

clintpg@ufl.edu