Topic Model

Jialu Liu

2010-9-19

remenberl@gmail.com

Outline

- Introduction to topic models
- VSM (Vector Space Model)
- LSA (Latent Semantic Analysis)
- pLSA (probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

Outline

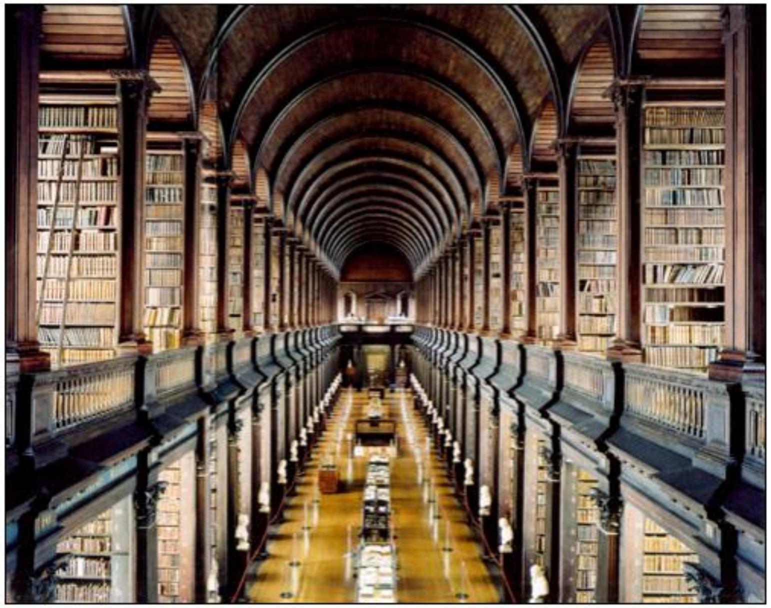
- Introduction to topic models
- VSM (Vector Space Model)
- LSA (Latent Semantic Analysis)
- pLSA (probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

Introduction

www.betaversion.org/~stefano/linotype/news/26/

- As more information becomes available, it becomes more difficult to access what we are looking for.
- We need new tools to help us organize, search, and understand these vast amounts of information.

Topic modeling



Topic modeling provides methods for automatically organizing, understanding, searching, and summarizing large electronic archives.

- Uncover the hidden topical patterns that pervade the collection.
- Annotate the documents according to those topics.
- Use the annotations to organize, summarize, and search the texts.

Bag-of-Words (BOW)

- Assumes order of words has no significance

  e.g., the term “home made” has the same probability as “made home”

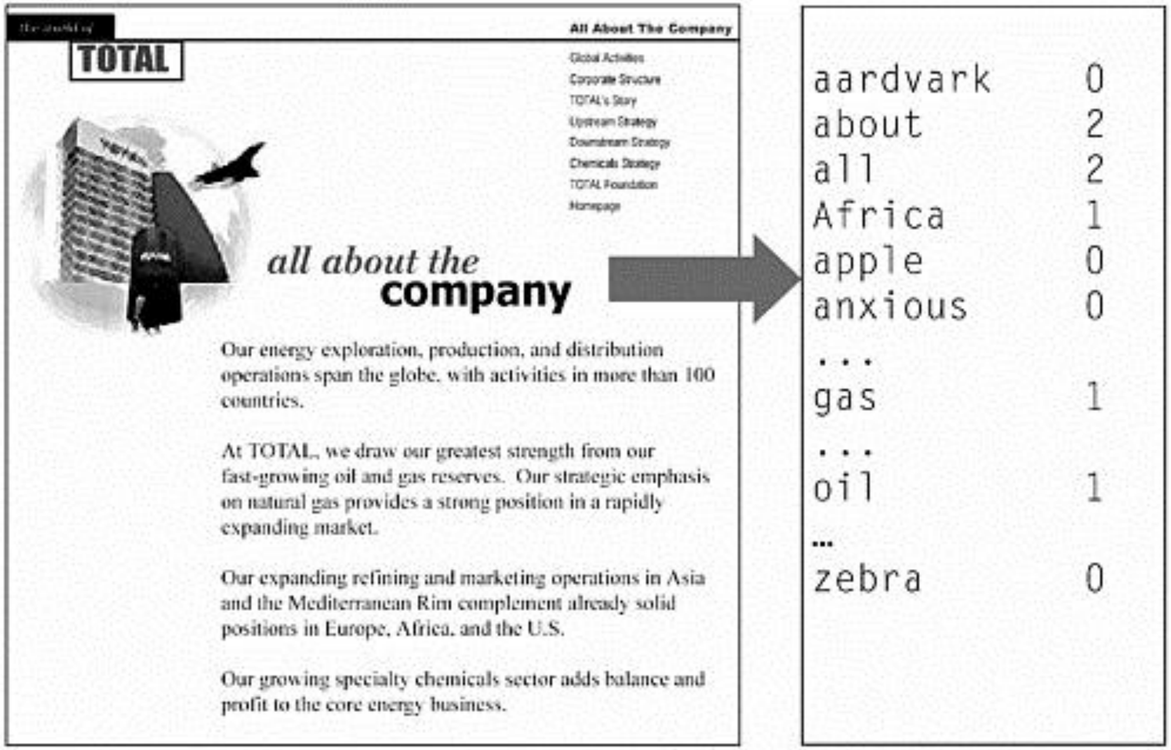
- It is a simplifying assumption used in natural language processing and information retrieval

Outline

- Introduction to topic models
- VSM (Vector Space Model)
- LSA (Latent Semantic Analysis)
- pLSA (probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

Salton’s Vector Space Model (Prior to 1988)

- Represent each document by a high-dimensional vector in the space of words



The world of

TOTAL

All About The Company

Global Activities
Corporate Structure
TOTAL's Story
Upstream Strategy
Downstream Strategy
Chemicals Strategy
TOTAL Foundation
Homepage

all about the **company**

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

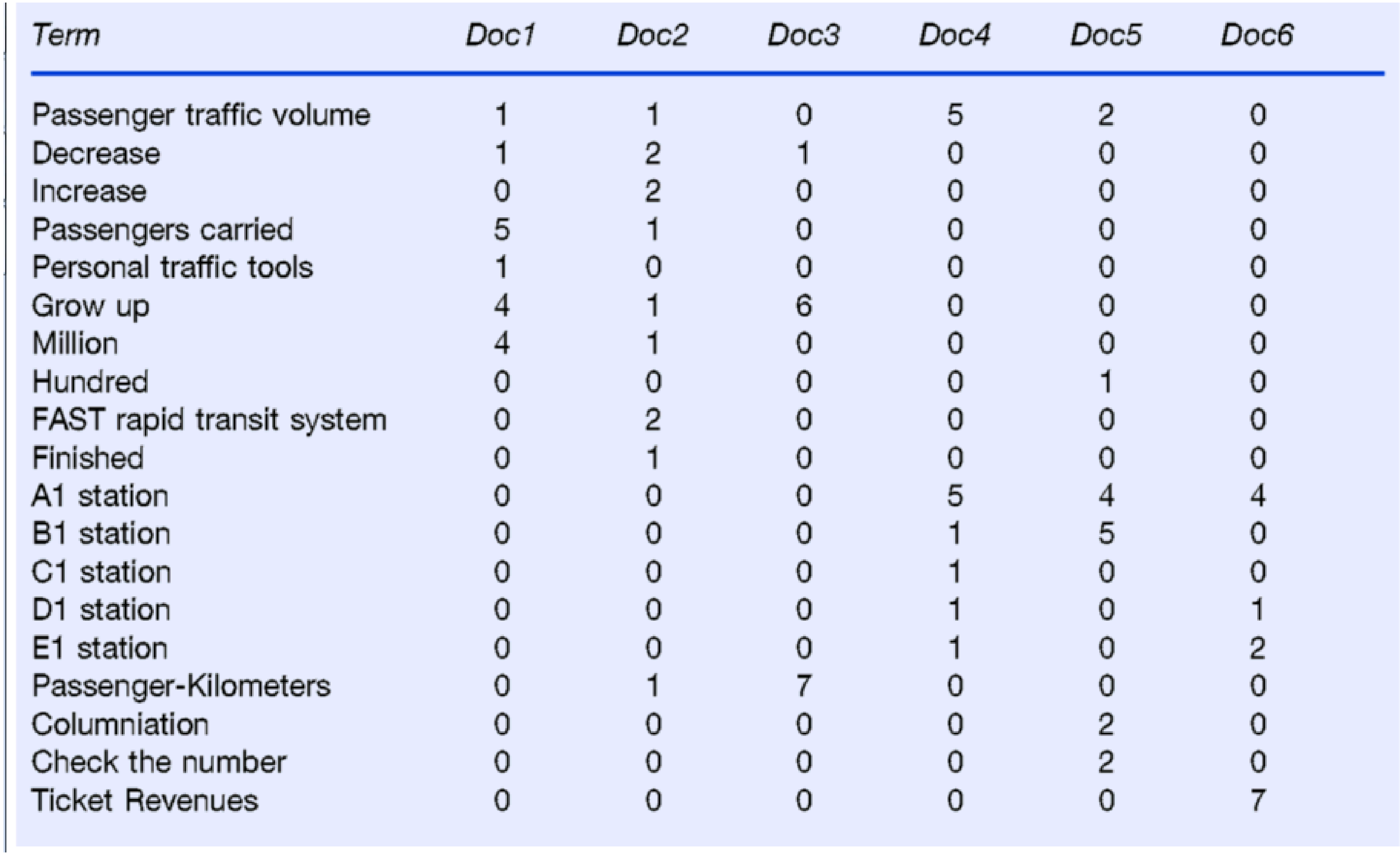
Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
zebra	0

- Represent the doc as a vector where each entry corresponds to a different word and the number at that entry corresponds to how many times that word was present in the document (or some function of it)
  - Number of words is huge
  - Select and use a smaller set of words that are of interest
  - E.g. uninteresting words: ‘and’, ‘the’ ‘at’, ‘is’, etc. These are called stop-words
  - Stemming: remove endings. E.g. ‘learn’, ‘learning’, ‘learnable’, ‘learned’ could be substituted by the single stem ‘learn’
  - Other simplifications can also be invented and used
  - The set of different remaining words is called dictionary or vocabulary. Fix an ordering of the terms in the dictionary so that you can operate them by their index.

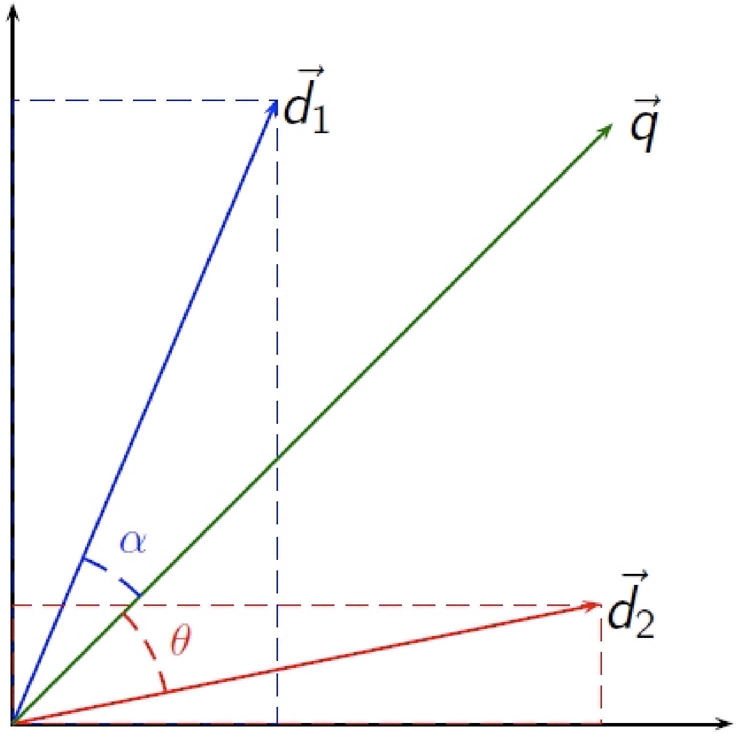
Term-document matrix



Term	*Doc1*	*Doc2*	*Doc3*	*Doc4*	*Doc5*	*Doc6*
Passenger traffic volume	1	1	0	5	2	0
Decrease	1	2	1	0	0	0
Increase	0	2	0	0	0	0
Passengers carried	5	1	0	0	0	0
Personal traffic tools	1	0	0	0	0	0
Grow up	4	1	6	0	0	0
Million	4	1	0	0	0	0
Hundred	0	0	0	0	1	0
FAST rapid transit system	0	2	0	0	0	0
Finished	0	1	0	0	0	0
A1 station	0	0	0	5	4	4
B1 station	0	0	0	1	5	0
C1 station	0	0	0	1	0	0
D1 station	0	0	0	1	0	1
E1 station	0	0	0	1	0	2
Passenger-Kilometers	0	1	7	0	0	0
Columniation	0	0	0	0	2	0
Check the number	0	0	0	0	2	0
Ticket Revenues	0	0	0	0	0	7

Query

- Compute the similarity between *queries(q)* and *documents(d)*



$$\cos(q, d) = \frac{q^T d}{\|q\| \|d\|}$$

Simple, intuitive

Fast to compute, because both they are sparse

Retrieval Methods

- Rank documents according to similarity with query
- Term weighting schemes, for example, TF-IDF

Limitations

- Dimensionality
  - Vector space representation is high-dimensional (several 10-100K)
  - Learning and estimation has to deal with curse of dimensionality
- Sparseness
  - Document vectors are typically very sparse
  - Cosine similarity can be noisy and inaccurate
- Semantics
  - The inner product can only match occurrences of exactly the same terms
  - The vector representation does not capture semantic relations between words
- Independence
  - Bag-of-Words Representation
  - Unable to capture phrases and semantic/syntactic regularities

The lost meaning of words

- Polysemy: words with multiple meanings
  - The vector space model is unable to discriminate between different meaning of the same word.

$$\mathrm{sim}(d,q) < \cos\left(\angle(\vec{d},\vec{q})\right)$$

- Synonymy: separate words that have the same meaning.
  - No associations between words are made in the vector space representation

$$\mathrm{sim}(d,q) > \cos\left(\angle(\vec{d},\vec{q})\right)$$

There is a disconnect between *topics* and *words*

Outline

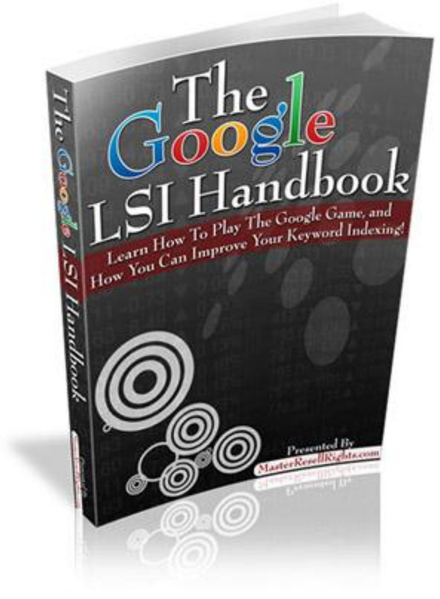
- Introduction to topic models
- VSM (Vector Space Model)
- LSA (Latent Semantic Analysis)
- pLSA (probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

Latent Semantic Analysis (1988)

- In the context of its application to information retrieval, it is sometimes called Latent Semantic Indexing (LSI)

- Patented in 1988 by Scott Deerwester, Susan Dumais, George Furnas, Richard Harshman, Thomas Landauer, Karen Lochbaum and Lynn Streeter

- See Wiki for details of the algorithm


The Google LSI Handbook
Learn How To Play The Google Game, and How You Can Improve Your Keyword Indexing!
Presented By
MasterResellRights.com
The Google LSI Handbook

- General Idea
  - Map documents (and terms) to a low-dimensional representation
  - Design a mapping such that the low-dimensional space reflects semantic association (latent semantic space)
  - Compute document similarity based on the inner product in the latent semantic space
- Goals
  - Similar terms map to similar location in low dimensional space
  - Noise reduction by dimension reduction

SVD

$$
\begin{array}{ccccccc}
 & X & & & U & \Sigma & V^T \\
 & \begin{matrix} (d_j) \\ \downarrow \end{matrix} & & & & & \begin{matrix} (\widehat{d_j}) \\ \downarrow \end{matrix} \\
({t_i}^T) \rightarrow & \begin{bmatrix} x_{1,1} & \cdots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \cdots & x_{m,n} \end{bmatrix} & = & (\widehat{\mathrm{t}_i}^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ u_1 \\ \\ \end{bmatrix} & \cdots & \begin{bmatrix} \\ u_l \\ \\ \end{bmatrix} \end{bmatrix} \cdot & \begin{bmatrix} \delta_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta_l \end{bmatrix} \cdot & \begin{bmatrix} [& v_1 &] \\ & \vdots & \\ [& v_l &] \end{bmatrix}
\end{array}
$$

- Given the Term-document matrix $X(m \times n)$, do the singular decomposition

$$X = U\Sigma V^T$$

- Notice how the only part of U that contributes to t_i is the *i*'th row. Let this row vector be called $\widehat{t_i}$. Likewise, the only part of V^T that contributes to d_j is the *j*'th column, $\widehat{d_j}$

- Selecting the k largest singular values, and corresponding singular vectors from U and V, you get the rank k approximation to X with the smallest error (Frobenius norm).
- This approximation has a minimal error.

$$\begin{array}{cccccc} & U_k & & \Sigma_k & & {V_k}^T \\ X_k = & \begin{bmatrix} \begin{bmatrix} \\ u_1 \\ \\ \end{bmatrix} & \cdots & \begin{bmatrix} \\ u_k \\ \\ \end{bmatrix} \end{bmatrix} & \cdot & \begin{bmatrix} \delta_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta_k \end{bmatrix} & \cdot & \begin{bmatrix} [& v_1 &] \\ & \vdots & \\ [& v_k &] \end{bmatrix} \end{array}$$

$$\overset{m}{\begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{bmatrix}}
=
\overset{u}{\begin{bmatrix}
0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & 0.41 \\
0.2 & -0.07 & 0.14 & -0.55 & 0.28 & 0.5 & -0.07 & -0.01 & 0.11 \\
0.24 & 0.04 & -0.16 & -0.59 & -0.11 & -0.25 & -0.3 & 0.06 & -0.49 \\
0.4 & 0.06 & -0.34 & 0.1 & 0.33 & 0.38 & 0 & 0 & -0.01 \\
0.64 & -0.17 & 0.36 & 0.33 & -0.16 & -0.21 & -0.17 & 0.03 & -0.27 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & 0.05 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & 0.05 \\
0.3 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & 0.17 \\
0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & 0.58 \\
0.01 & 0.49 & 0.23 & 0.02 & 0.59 & -0.39 & -0.29 & 0.25 & 0.23 \\
0.04 & 0.62 & 0.22 & 0 & -0.07 & 0.11 & 0.16 & -0.68 & -0.23 \\
0.03 & 0.45 & 0.14 & -0.01 & -0.3 & 0.28 & 0.34 & 0.68 & -0.18
\end{bmatrix}}
\overset{\Sigma}{\begin{bmatrix}
3.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2.54 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2.35 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.64 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1.31 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.85 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.56 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.36
\end{bmatrix}}
\overset{v}{\begin{bmatrix}
0.2 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & 0.06 \\
0.61 & 0.17 & -0.5 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & -0.24 \\
0.46 & -0.13 & 0.21 & 0.04 & 0.38 & 0.72 & -0.24 & 0.01 & -0.02 \\
0.54 & -0.23 & 0.57 & 0.27 & -0.21 & -0.37 & 0.26 & -0.02 & 0.08 \\
0.28 & 0.11 & -0.51 & 0.15 & 0.33 & 0.03 & 0.67 & -0.06 & 0.26 \\
0 & 0.19 & 0.1 & 0.02 & 0.39 & -0.3 & -0.34 & 0.45 & 0.62 \\
0.01 & 0.44 & 0.19 & 0.02 & 0.35 & -0.21 & -0.15 & -0.76 & -0.02 \\
0.02 & 0.62 & 0.25 & 0.01 & 0.15 & 0 & 0.25 & 0.45 & -0.52 \\
0.08 & 0.53 & 0.08 & -0.02 & -0.6 & 0.36 & 0.04 & -0.07 & 0.45
\end{bmatrix}}$$

$$\Downarrow$$

$$\overset{m}{\begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1
\end{bmatrix}}
=
\overset{u}{\begin{bmatrix}
0.22 & -0.11 & 0.29 & -0.41 & -0.11 & -0.34 & 0.52 & -0.06 & 0.41 \\
0.2 & -0.07 & 0.14 & -0.55 & 0.28 & 0.5 & -0.07 & -0.01 & 0.11 \\
0.24 & 0.04 & -0.16 & -0.59 & -0.11 & -0.25 & -0.3 & 0.06 & -0.49 \\
0.4 & 0.06 & -0.34 & 0.1 & 0.33 & 0.38 & 0 & 0 & -0.01 \\
0.64 & -0.17 & 0.36 & 0.33 & -0.16 & -0.21 & -0.17 & 0.03 & -0.27 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & 0.05 \\
0.27 & 0.11 & -0.43 & 0.07 & 0.08 & -0.17 & 0.28 & -0.02 & 0.05 \\
0.3 & -0.14 & 0.33 & 0.19 & 0.11 & 0.27 & 0.03 & -0.02 & 0.17 \\
0.21 & 0.27 & -0.18 & -0.03 & -0.54 & 0.08 & -0.47 & -0.04 & 0.58 \\
0.01 & 0.49 & 0.23 & 0.02 & 0.59 & -0.39 & -0.29 & 0.25 & 0.23 \\
0.04 & 0.62 & 0.22 & 0 & -0.07 & 0.11 & 0.16 & -0.68 & -0.23 \\
0.03 & 0.45 & 0.14 & -0.01 & -0.3 & 0.28 & 0.34 & 0.68 & -0.18
\end{bmatrix}}
\overset{\Sigma}{\begin{bmatrix}
3.34 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 2.54 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 2.35 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1.64 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1.5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1.31 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0.85 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.56 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.36
\end{bmatrix}}
\overset{v}{\begin{bmatrix}
0.2 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & 0.06 \\
0.61 & 0.17 & -0.5 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & -0.24 \\
0.46 & -0.13 & 0.21 & 0.04 & 0.38 & 0.72 & -0.24 & 0.01 & -0.02 \\
0.54 & -0.23 & 0.57 & 0.27 & -0.21 & -0.37 & 0.26 & -0.02 & 0.08 \\
0.28 & 0.11 & -0.51 & 0.15 & 0.33 & 0.03 & 0.67 & -0.06 & 0.26 \\
0 & 0.19 & 0.1 & 0.02 & 0.39 & -0.3 & -0.34 & 0.45 & 0.62 \\
0.01 & 0.44 & 0.19 & 0.02 & 0.35 & -0.21 & -0.15 & -0.76 & -0.02 \\
0.02 & 0.62 & 0.25 & 0.01 & 0.15 & 0 & 0.25 & 0.45 & -0.52 \\
0.08 & 0.53 & 0.08 & -0.02 & -0.6 & 0.36 & 0.04 & -0.07 & 0.45
\end{bmatrix}}$$

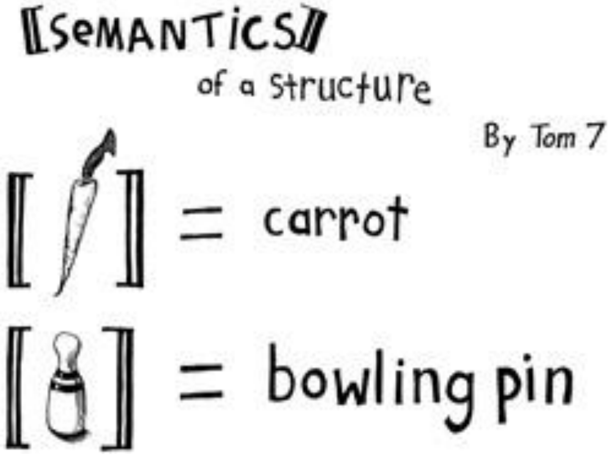
$$\Downarrow$$

$$\overset{m}{\begin{bmatrix}
0.16 & 0.4 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\
0.14 & 0.37 & 0.33 & 0.4 & 0.16 & -0.03 & -0.07 & -0.1 & -0.04 \\
0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\
0.26 & 0.84 & 0.61 & 0.7 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\
0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\
0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\
0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\
0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.2 & -0.11 \\
0.1 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\
-0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & 0.66 \\
-0.06 & 0.34 & -0.15 & -0.3 & 0.2 & 0.31 & 0.69 & 0.98 & 0.85 \\
-0.04 & 0.25 & -0.1 & -0.21 & 0.15 & 0.22 & 0.5 & 0.71 & 0.62
\end{bmatrix}}
=
\overset{u}{\begin{bmatrix}
0.22 & -0.11 \\
0.2 & -0.07 \\
0.24 & 0.04 \\
0.4 & 0.06 \\
0.64 & -0.17 \\
0.27 & 0.11 \\
0.27 & 0.11 \\
0.3 & -0.14 \\
0.21 & 0.27 \\
0.01 & 0.49 \\
0.04 & 0.62 \\
0.03 & 0.45
\end{bmatrix}}
\overset{\Sigma}{\begin{bmatrix}
3.34 & 0 \\
0 & 2.54
\end{bmatrix}}
\overset{v}{\begin{bmatrix}
0.2 & -0.06 & 0.11 & -0.95 & 0.05 & -0.08 & 0.18 & -0.01 & 0.06 \\
0.61 & 0.17 & -0.5 & -0.03 & -0.21 & -0.26 & -0.43 & 0.05 & -0.24
\end{bmatrix}}$$

concept space

- We can now treat the term and document vectors as a "concept space":
  - The vector then has k entries, each giving the occurrence of term i in one of the k concepts. Likewise, the vector gives the relation between document j and each concept.

$$(\widehat{t_i}^T) \rightarrow \begin{bmatrix} \begin{bmatrix} u_1 \end{bmatrix} & \cdots & \begin{bmatrix} u_k \end{bmatrix} \end{bmatrix} \cdot \begin{bmatrix} \delta_0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \delta_k \end{bmatrix} \cdot \overset{\overset{(\widehat{d_j})}{\downarrow}}{\begin{bmatrix} [\qquad v_1 \qquad] \\ \vdots \\ [\qquad v_k \qquad] \end{bmatrix}}$$



- Here is what you can do with the so-called "concept space"

1. See how related documents j and q are in the concept space by comparing the vectors $\widehat{d_j}$ and $\widehat{d_q}$(typically by cosine similarity).
2. Comparing terms i and p by comparing the vectors $\widehat{t_i}$ and $\widehat{t_p}$.
3. Given a query, view this as a mini document, and compare it to your documents in the concept space.

- To do the latter, you must first translate your query into the concept space.

$$d_j = U_k \Sigma_k \widehat{d_j}$$

$$\widehat{d_j} = {\Sigma_k}^{-1} {U_k}^T d_j$$

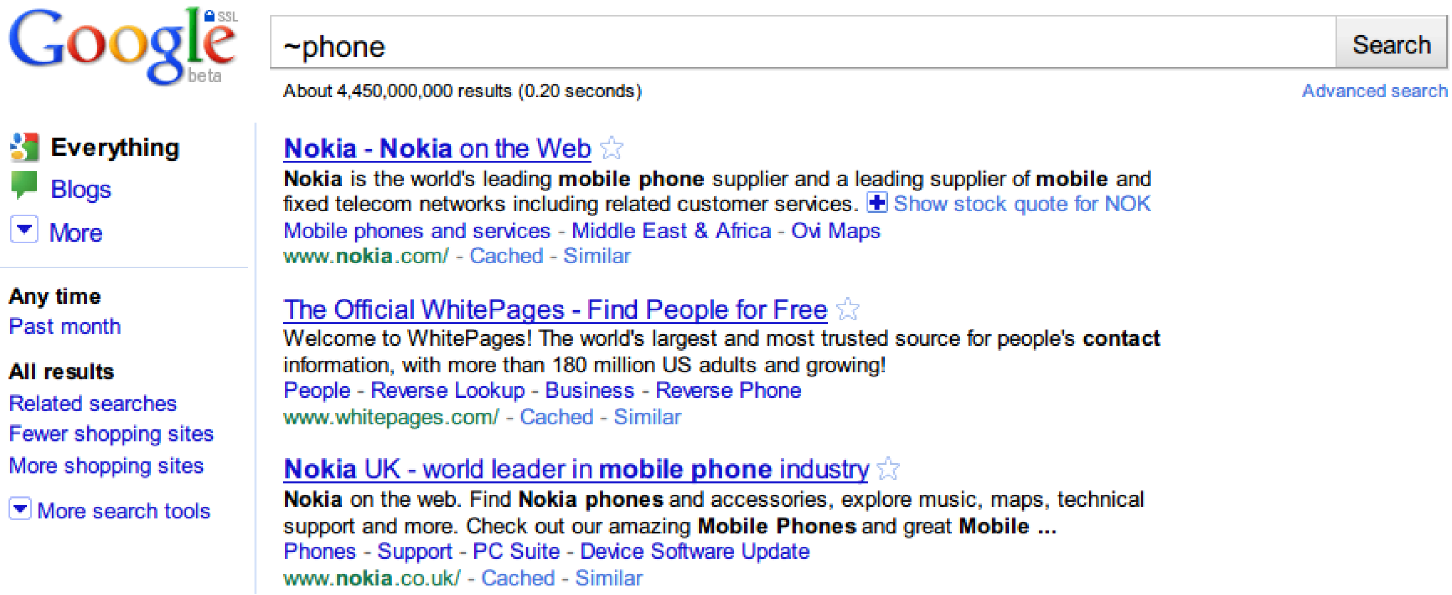
- You can do the same for pseudo term vectors.

$${t_i}^T = \widehat{t_i}^T \Sigma_k {V_k}^T$$

$$\widehat{t_i} = {\Sigma_k}^{-1} {V_k}^T t_i$$

Application of LSA

- Google
  - "~" sign before the search term stands for the semantic search, for instance "~phone"



About 4,450,000,000 results (0.20 seconds)

Nokia - Nokia on the Web

Nokia is the world's leading **mobile phone** supplier and a leading supplier of **mobile** and fixed telecom networks including related customer services. Show stock quote for NOK
Mobile phones and services - Middle East & Africa - Ovi Maps
www.**nokia**.com/ - Cached - Similar

The Official WhitePages - Find People for Free

Welcome to WhitePages! The world's largest and most trusted source for people's **contact** information, with more than 180 million US adults and growing!
People - Reverse Lookup - Business - Reverse Phone
www.whitepages.com/ - Cached - Similar

Nokia UK - world leader in **mobile phone** industry

Nokia on the web. Find **Nokia phones** and accessories, explore music, maps, technical support and more. Check out our amazing **Mobile Phones** and great **Mobile** ...
Phones - Support - PC Suite - Device Software Update
www.**nokia**.co.uk/ - Cached - Similar

  - Google Adsense sandbox
    - http://www.technolinks.co.uk/2010/05/seo-and-lsi-how-to-use-latent-semantic-indexing/

- TOEFL
  - a word is given
  - the most similar in meaning should be selected from the four words
  - scored %65 correct

Discussion of LDA

- pros:
  - Low-dimensional document representation is able to capture synonyms
  - Noise removal and robustness by dimension reduction
  - Experimentally: advantages over naive vector space model
- cons:
  - L2 norm is inappropriate as a distance function for count vectors (reconstruction may contain negative entries)
  - “Conceptually”:
    - Problem of polysemy is not addressed; principle of linear superposition, no active disambiguation
    - Context of terms is not taken into account (BOW)
    - Direction in latent space are hard to interpret
    - No probabilistic model of term occurrences
  - Ad hoc selection of the number of dimensions…

Outline

- Introduction to topic models
- VSM (Vector Space Model)
- LSA (Latent Semantic Analysis)
- pLSA (probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

probabilistic Latent Semantic Analysis

- PLSA evolved from Latent semantic analysis, adding a sounder probabilistic model
- It was introduced in 1999 by Thomas Hofmann (UAI’99)
- It is related to non-negative matrix factorization (NMF)

- Bayes rule: probability of relevance of document w.r.t query, w means word

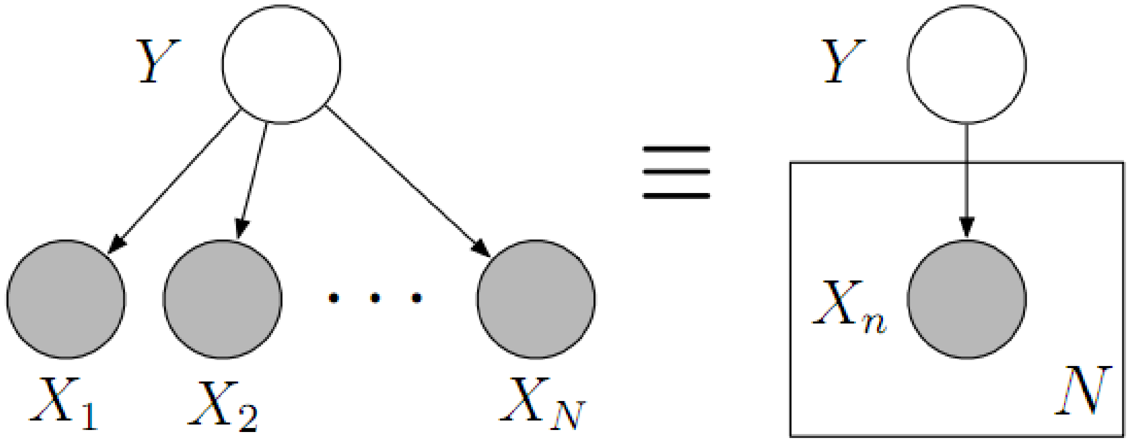
$$P(d|q) \propto P(q|d)P(d)$$

$$P(q|d) = \prod_{w \in q} P(w|d)$$

- Probabilistic dimension reduction techniques to overcome data sparseness problem, where z is a latent variable, stands for topic

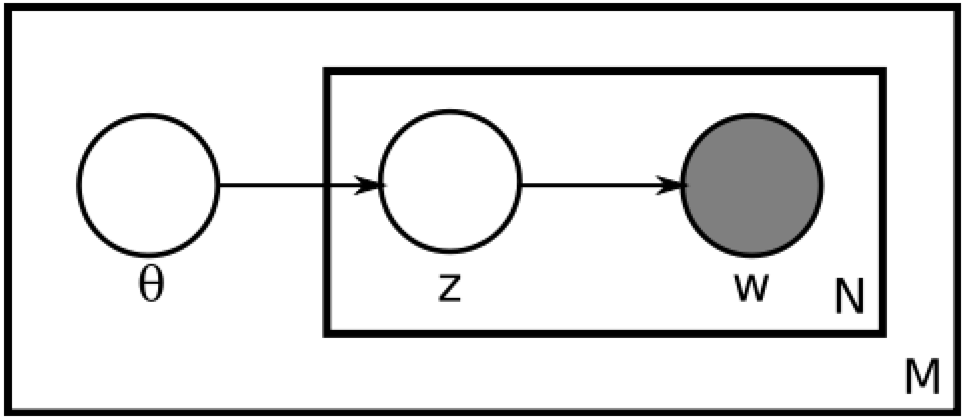
$$P(w|d) = \sum_{z} P(w|z)P(z|d)$$

Graphic model



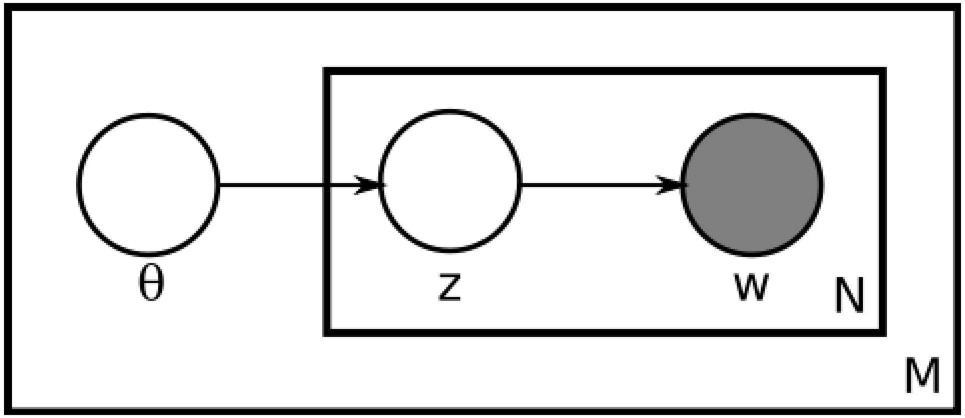
- Nodes are random variables
- Edges denote possible dependence
- Observed variables are shaded
- Plate denote replicated structure

Graphic model of pLSA



$$P(w,d) = \sum_z P(z)\, P(w|z) P(d|z) = P(d) \sum_z P(w|z) P(z|d)$$

- θ is the document variable (d in the text), z is a topic drawn from the topic distribution for this document, $P(z|d)$, and w is a word drawn from the word distribution for this topic, $P(w|z)$. The θ and w are observable variables, the topic z is a latent variable.
- Suppose there exist M documents and the bag of words consist of N words.



- $P(z|d)$ is shared by all words in a document
- $P(w|z)$ is shared by all documents in collection
- It is possible to derive the equations for computing these parameters by Maximum Likelihood

Maximum Likelihood

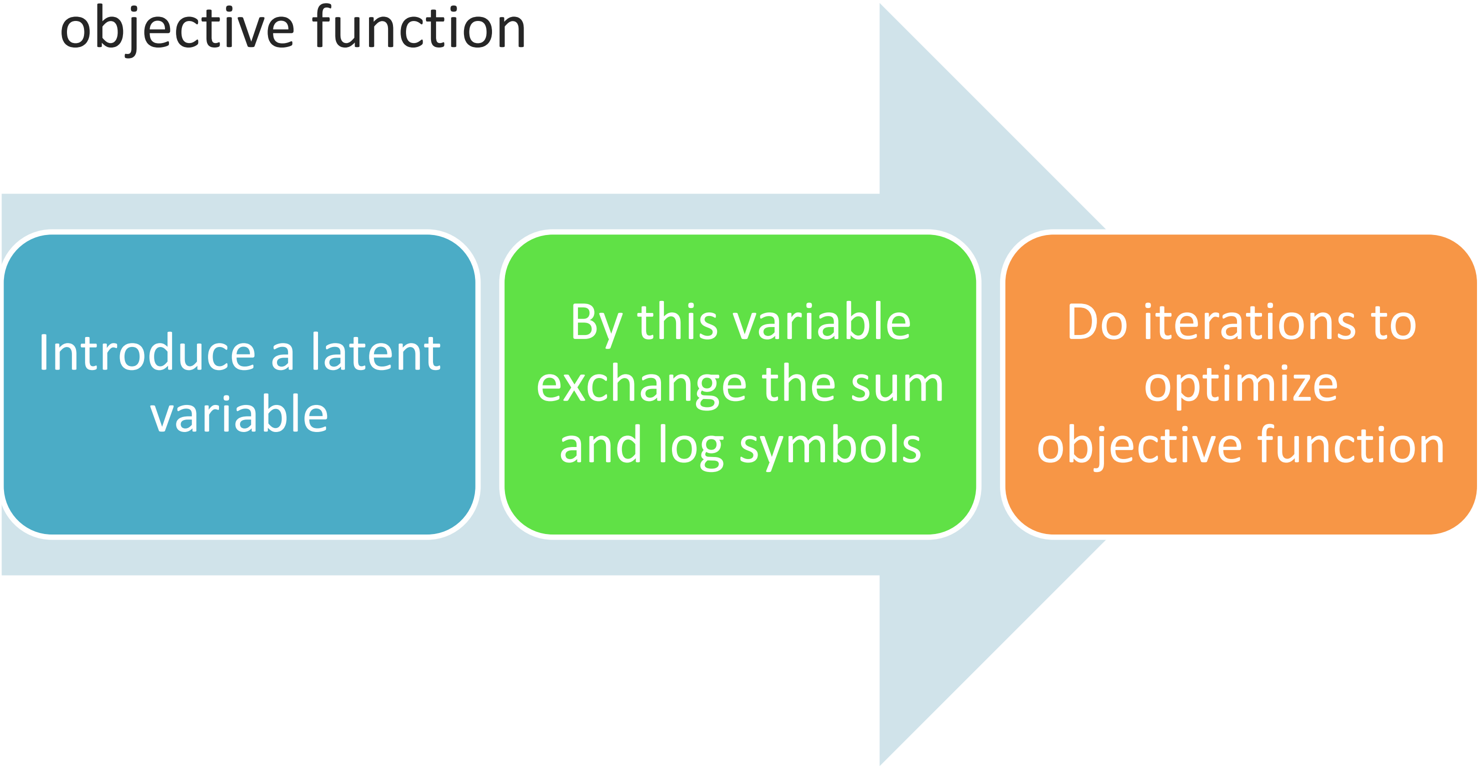
- The log likelihood of this model is the log probability of the entire collection:

$$L = \sum_{i,j} n\left(w_j, d_i\right) \log P(w, d) = \sum_{i,j} n\left(w_j, d_i\right) \log \sum_{z} P(z)\, P\left(w_j|z\right) P(d_i|z)$$

- Which is to be maximised w.r.t. parameters $P(w|z)$ and also $P(d|z)$, subject to the constraints that $\sum_{j=1}^{N} P\left(w_j|z\right) = 1$ and $\sum_{\mathrm{k}=1}^{\mathrm{K}} P(\mathrm{z}_{\mathrm{k}}|\mathrm{d}) = 1$

Expectation Maximization (EM)

- It is a process of iteration which consists of Expectation step and Maximization step with latent variables
- We need this algorithm to find solutions to the objective function



Introduce a latent variable

By this variable exchange the sum and log symbols

Do iterations to optimize objective function

Expectation Maximization (EM)

- Alternating the two steps:
- Expectation-step:

$$P(z|\mathrm{d},\mathrm{w}) = \frac{P(w|z)P(d|z)P(z)}{\sum_z P(w|z)P(d|z)P(z)}$$

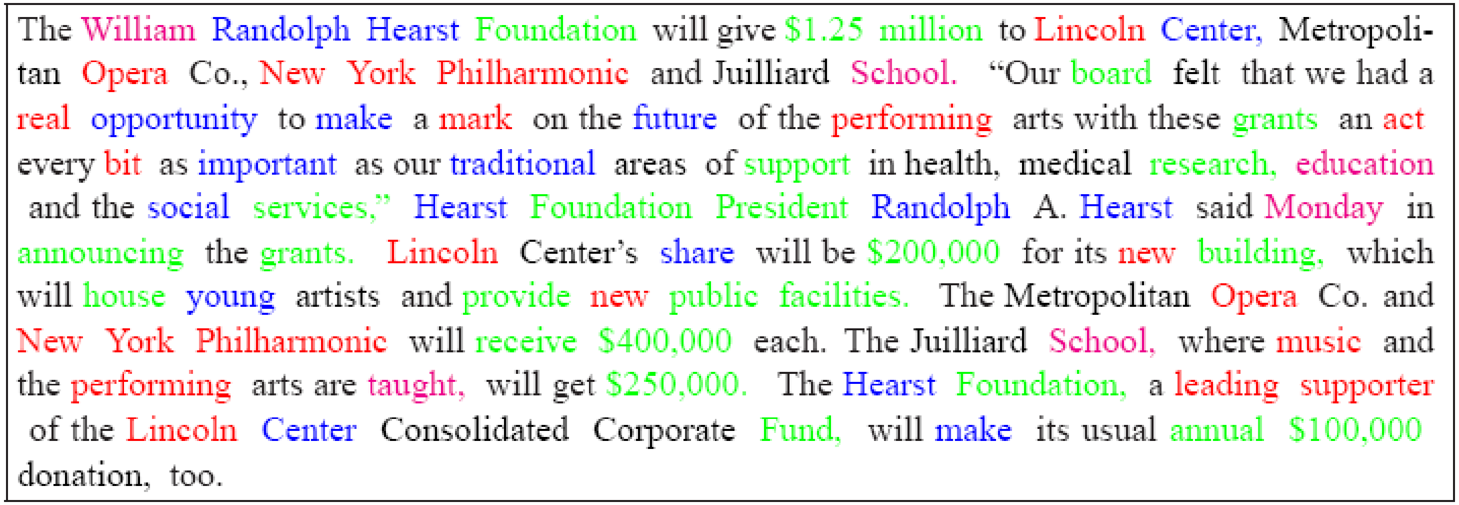
- Maximization-step:

$$P(z) = \frac{\sum_d \sum_w n(w,d)P(z|\mathrm{d},\mathrm{w})}{\sum_d \sum_w \sum_z n(w,d)P(z|\mathrm{d},\mathrm{w})}$$

$$P(w|\mathrm{z}) = \frac{\sum_d n(w,d)P(z|\mathrm{d},\mathrm{w})}{\sum_d \sum_w n(w,d)P(z|\mathrm{d},\mathrm{w})} \qquad P(d|\mathrm{z}) = \frac{\sum_w n(w,d)P(z|\mathrm{d},\mathrm{w})}{\sum_d \sum_w n(w,d)P(z|\mathrm{d},\mathrm{w})}$$

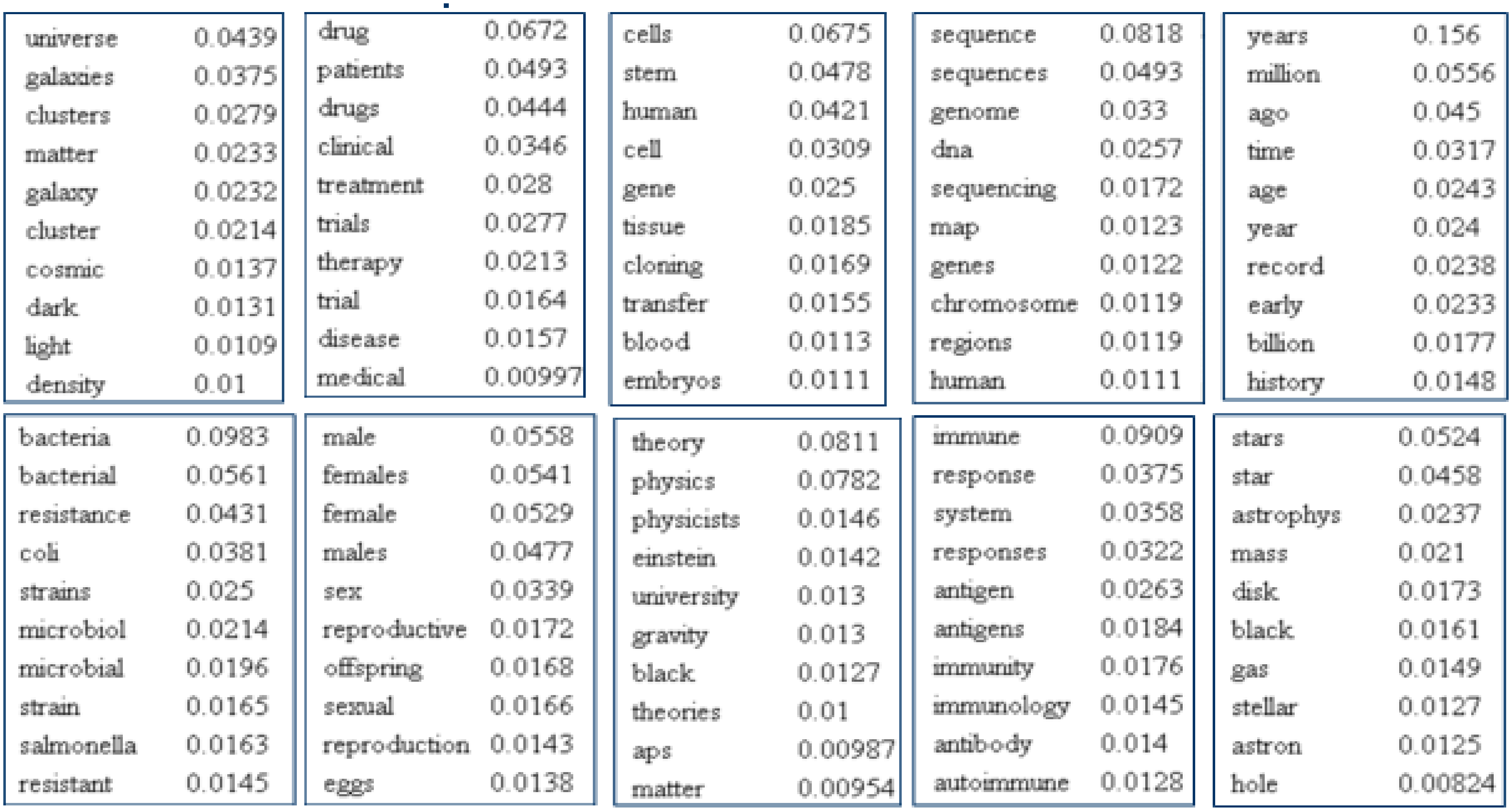
For detail, please refer to http://www.hongliangjie.com/2010/01/04/notes-on-probabilistic-latent-semantic-analysis-plsa/

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

$P(w|z)$ Matrix



universe	0.0439
galaxies	0.0375
clusters	0.0279
matter	0.0233
galaxy	0.0232
cluster	0.0214
cosmic	0.0137
dark	0.0131
light	0.0109
density	0.01

drug	0.0672
patients	0.0493
drugs	0.0444
clinical	0.0346
treatment	0.028
trials	0.0277
therapy	0.0213
trial	0.0164
disease	0.0157
medical	0.00997

cells	0.0675
stem	0.0478
human	0.0421
cell	0.0309
gene	0.025
tissue	0.0185
cloning	0.0169
transfer	0.0155
blood	0.0113
embryos	0.0111

sequence	0.0818
sequences	0.0493
genome	0.033
dna	0.0257
sequencing	0.0172
map	0.0123
genes	0.0122
chromosome	0.0119
regions	0.0119
human	0.0111

years	0.156
million	0.0556
ago	0.045
time	0.0317
age	0.0243
year	0.024
record	0.0238
early	0.0233
billion	0.0177
history	0.0148

bacteria	0.0983
bacterial	0.0561
resistance	0.0431
coli	0.0381
strains	0.025
microbiol	0.0214
microbial	0.0196
strain	0.0165
salmonella	0.0163
resistant	0.0145

male	0.0558
females	0.0541
female	0.0529
males	0.0477
sex	0.0339
reproductive	0.0172
offspring	0.0168
sexual	0.0166
reproduction	0.0143
eggs	0.0138

theory	0.0811
physics	0.0782
physicists	0.0146
einstein	0.0142
university	0.013
gravity	0.013
black	0.0127
theories	0.01
aps	0.00987
matter	0.00954

immune	0.0909
response	0.0375
system	0.0358
responses	0.0322
antigen	0.0263
antigens	0.0184
immunity	0.0176
immunology	0.0145
antibody	0.014
autoimmune	0.0128

stars	0.0524
star	0.0458
astrophys	0.0237
mass	0.021
disk	0.0173
black	0.0161
gas	0.0149
stellar	0.0127
astron	0.0125
hole	0.00824

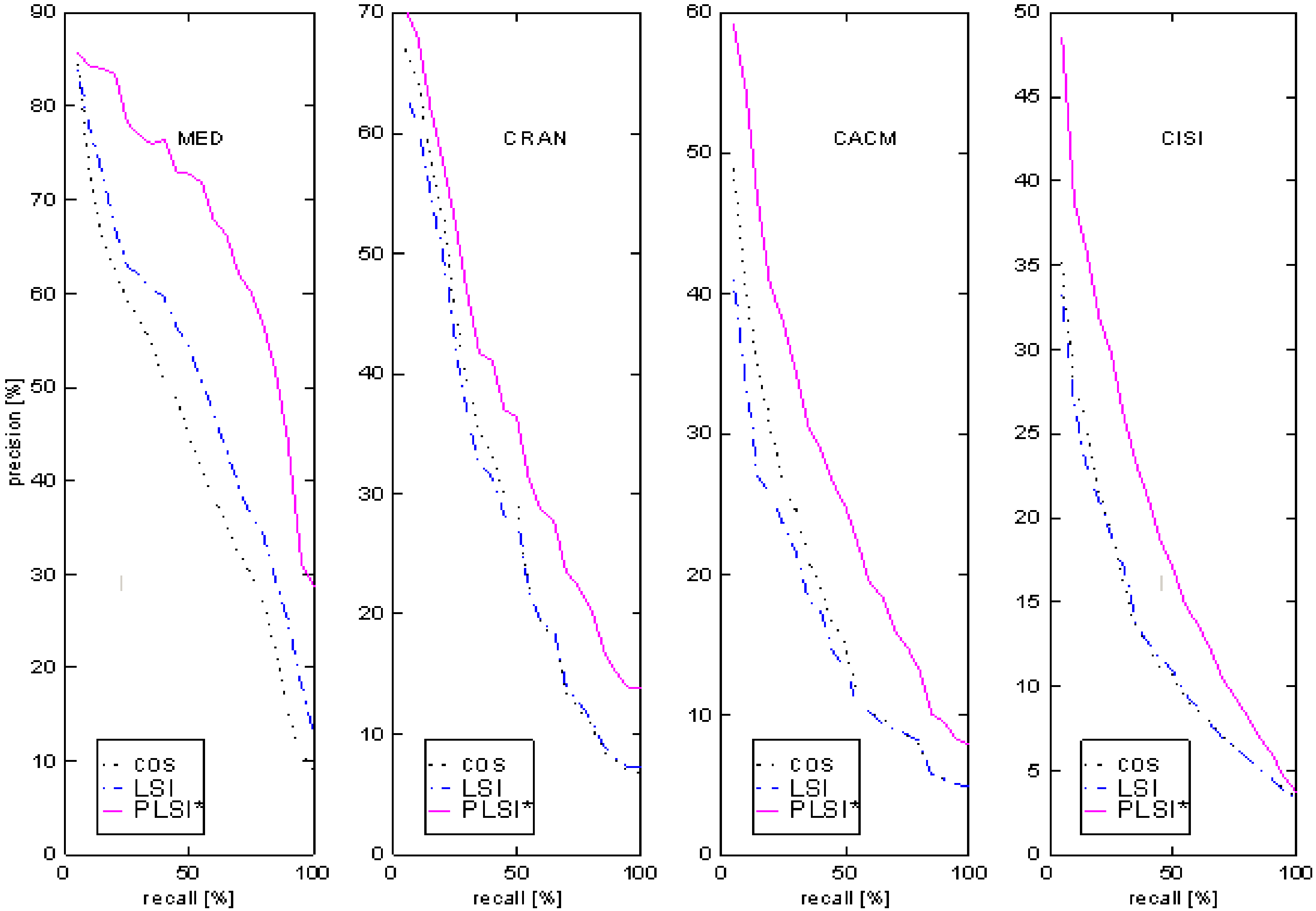
Example of topics found from a Science Magazine papers collection

pLSA v.s LSA

(probabilistic approach v.s matrix decomposition)

- Conditional independence assumption “replaces” outer product
- Class-conditional distributions “replace” left/right eigenvectors
- Maximum likelihood instead of minimum L2 Norm

The performance of a retrieval system based on this model (PLSI) was found superior to that of both the vector space based similarity (cos) and a non-probabilistic latent semantic indexing (LSI) method. (We skip details here.)



From Th. Hofmann, 2000

variations of pLSA

- Hierarchical extensions:
  - Asymmetric: MASHA ("Multinomial Asymmetric Hierarchical Analysis")
  - Symmetric: HPLSA ("Hierarchical Probabilistic Latent Semantic Analysis")
- Manifold regularizer:
  - Probabilistic Dyadic Data Analysis with Local and Global Consistency
- Generative models:
  - Latent Dirichlet allocation - adds a Dirichlet prior on the per-document topic distribution, trying to address an often-criticized shortcoming of PLSA, namely that it is not a proper generative model for new documents and at the same time avoid the overfitting problem.

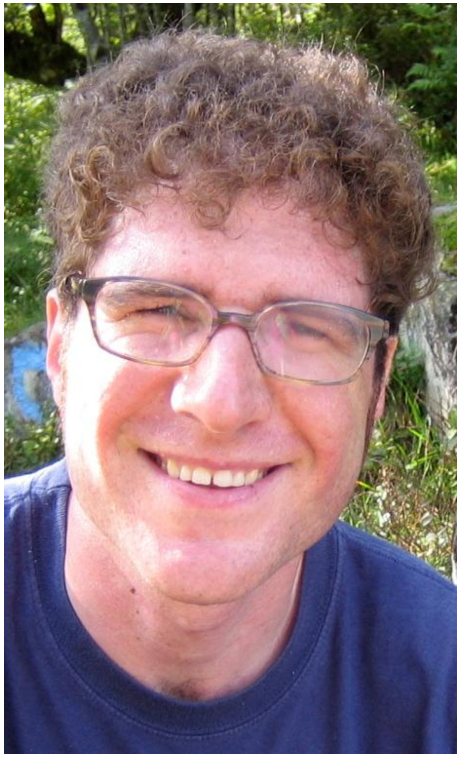
Outline

- Introduction to topic models
- VSM (Vector Space Model)
- LSA (Latent Semantic Analysis)
- pLSA (probabilistic Latent Semantic Analysis)
- LDA (Latent Dirichlet Allocation)

Latent Dirichlet Allocation (LDA)

- By 2003 Hofman’s PLSI model was put into question, this time by David Blei, Andrew Ng and Michael Jordan, who proposed that year the Latent Dirichlet Allocation Model (LDA).
- As noted by Blei, et al. (and quote)
  - pLSI “is incomplete in that it provides no probabilistic model at the level of documents. In pLSI, each document is represented as a list of numbers (the mixing proportions for topics), and there is no generative probabilistic model for these numbers.”

LDA (2003)

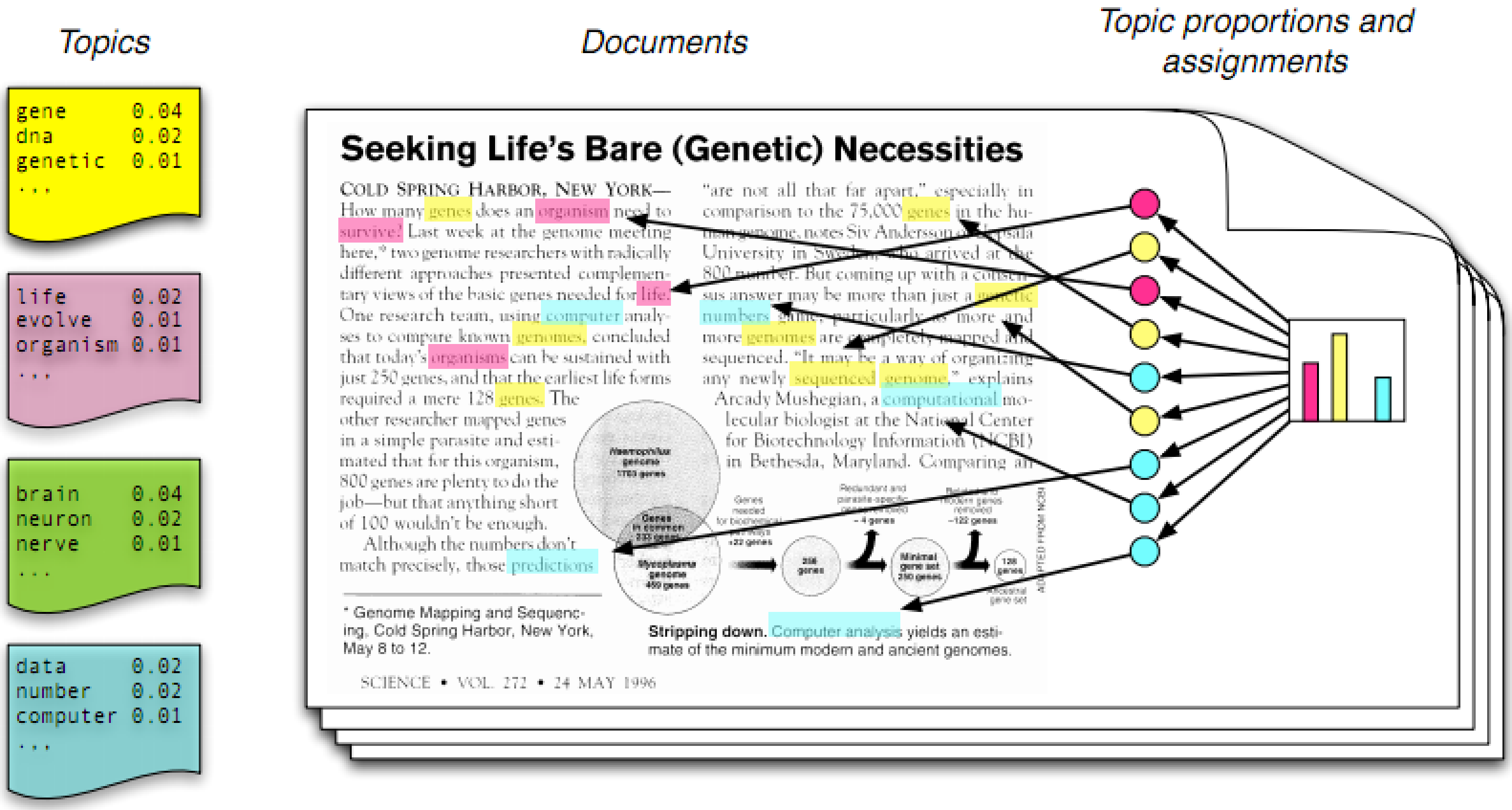


David M. Blei

Andrew Ng

Michael Jordon

Generative Model



Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Topic proportions and assignments

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

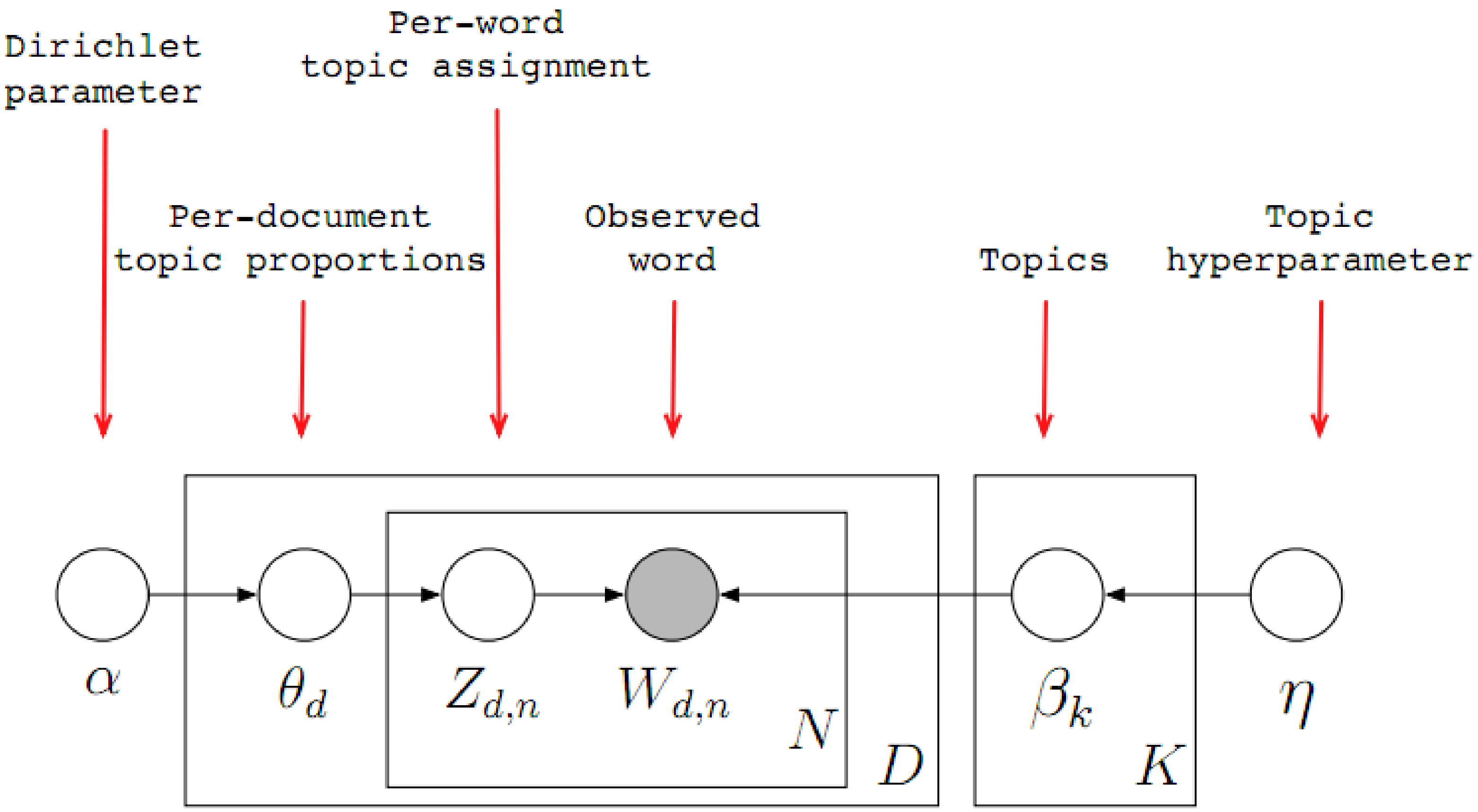
Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 468 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Redundant modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

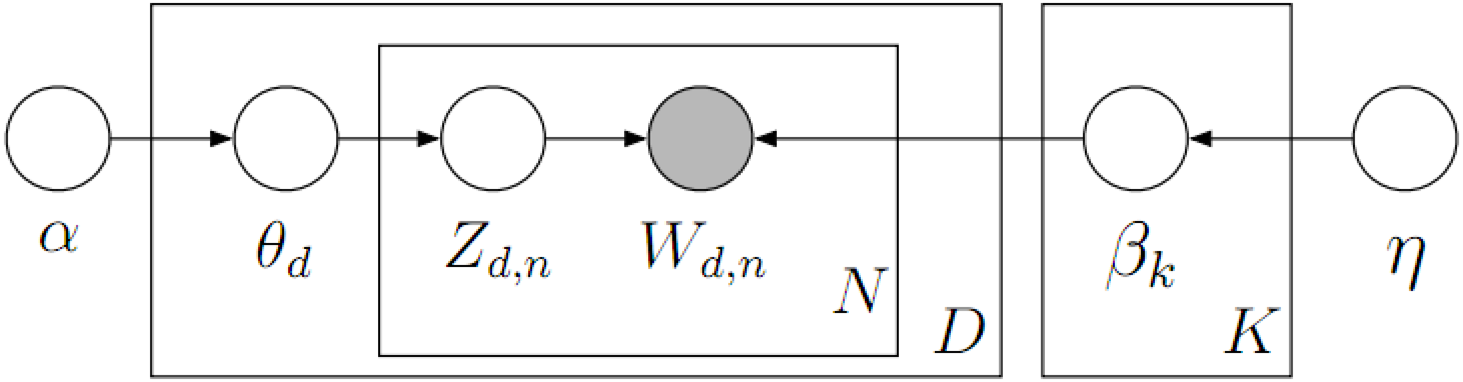
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Graphic model of LDA





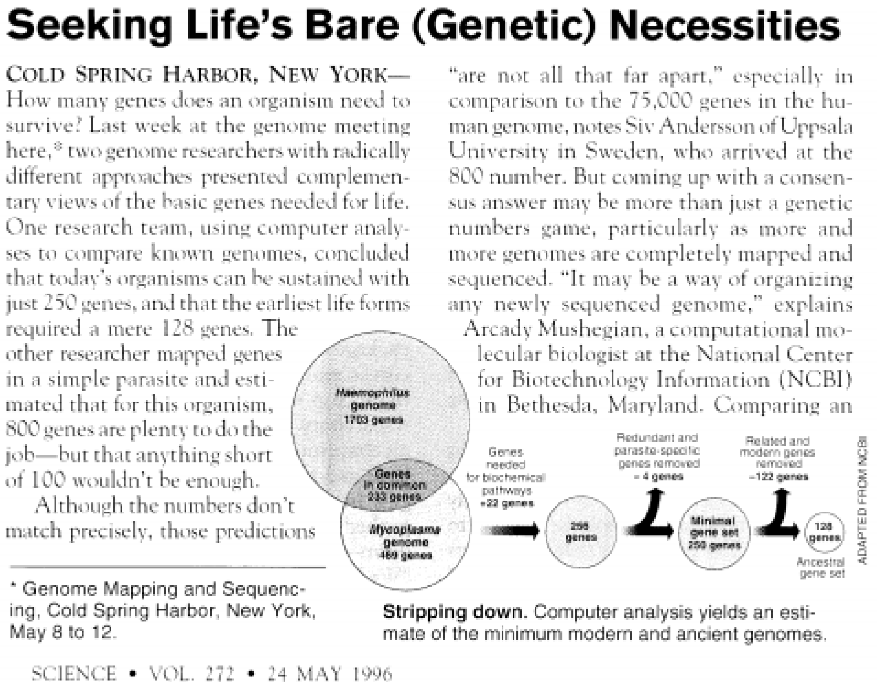
From a collection of documents, infer:

- Per-word topic assignment $z_{d,n}$
- Per-document topic proportions θ_d
- Per-corpus topic distribution β_k

Use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, etc.

- For example, an LDA model might have topics **CAT** and **DOG**. The **CAT** topic has probabilities of generating various words: the words *milk*, *meow*, *kitten* and of course *cat* will have high probability given this topic. The **DOG** topic likewise has probabilities of generating each word: *puppy*, *bark* and *bone* might have high probability.
- A document is generated by picking a distribution over topics (ie, mostly about **DOG**, mostly about **CAT**, or a bit of both), and given this distribution, picking the topic of each specific word. Then words are generated given
- LDA is similar to probabilistic latent semantic analysis (pLSA), except that in LDA the topic distribution is assumed to have a Dirichlet prior their topics.
- If you are interested, please find Blei's paper and slides.

Example of LDA



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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

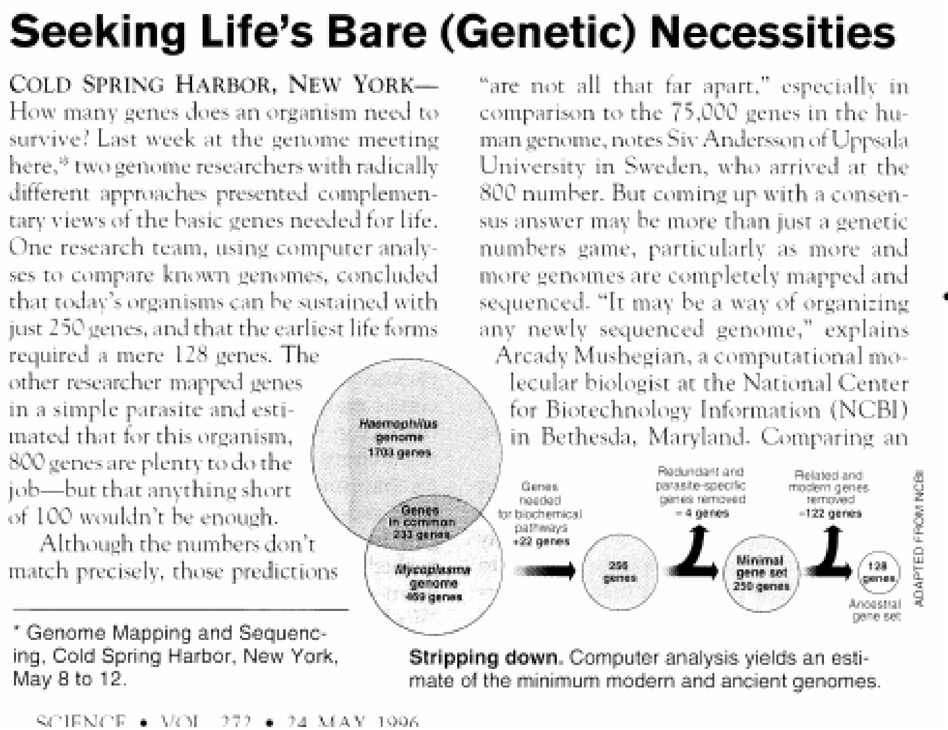
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- Data: The OCR'ed collection of Science from 1990-2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)
- Model: 100-topic LDA model using variational inference.

Example inference 1



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.

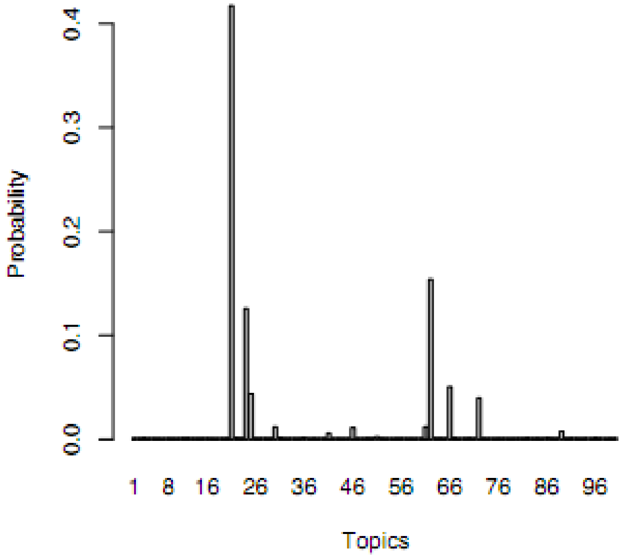
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just 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

Haemophilus genome 1703 genes
Genes in common 233 genes
Mycoplasma genome 469 genes
Genes needed for biochemical pathways +22 genes
256 genes
Redundant and parasite-specific genes removed −4 genes
Minimal gene set 250 genes
Related and modern genes removed −122 genes
128 genes
Ancestral gene set
ADAPTED FROM NCBI

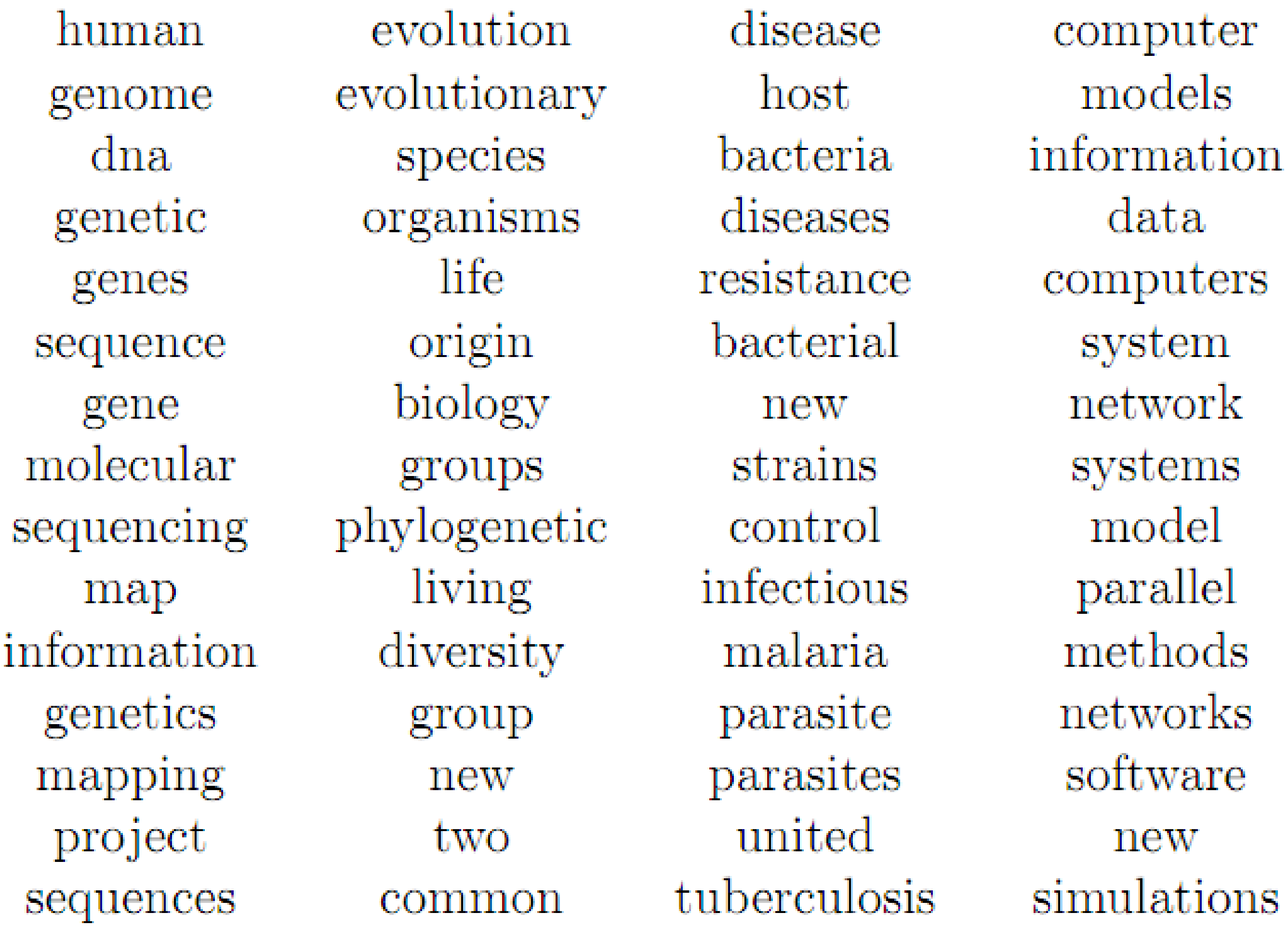
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

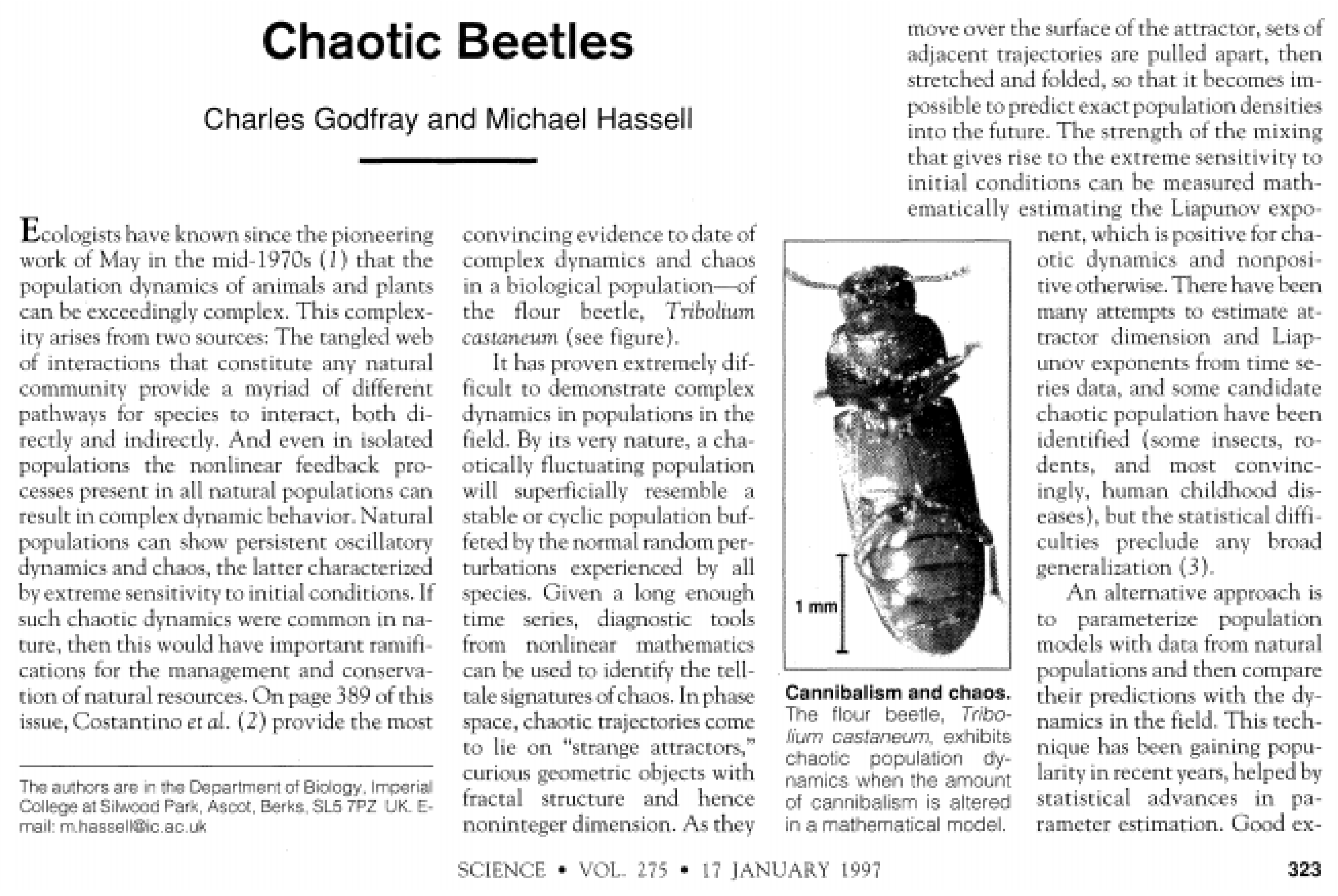


Example inference 1



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 inference 2



Chaotic Beetles

Charles Godfray and Michael Hassell

Ecologists have known since the pioneering work of May in the mid-1970s (1) that the population dynamics of animals and plants can be exceedingly complex. This complexity arises from two sources: The tangled web of interactions that constitute any natural community provide a myriad of different pathways for species to interact, both directly and indirectly. And even in isolated populations the nonlinear feedback processes present in all natural populations can result in complex dynamic behavior. Natural populations can show persistent oscillatory dynamics and chaos, the latter characterized by extreme sensitivity to initial conditions. If such chaotic dynamics were common in nature, then this would have important ramifications for the management and conservation of natural resources. On page 389 of this issue, Costantino *et al.* (2) provide the most convincing evidence to date of complex dynamics and chaos in a biological population—of the flour beetle, *Tribolium castaneum* (see figure).

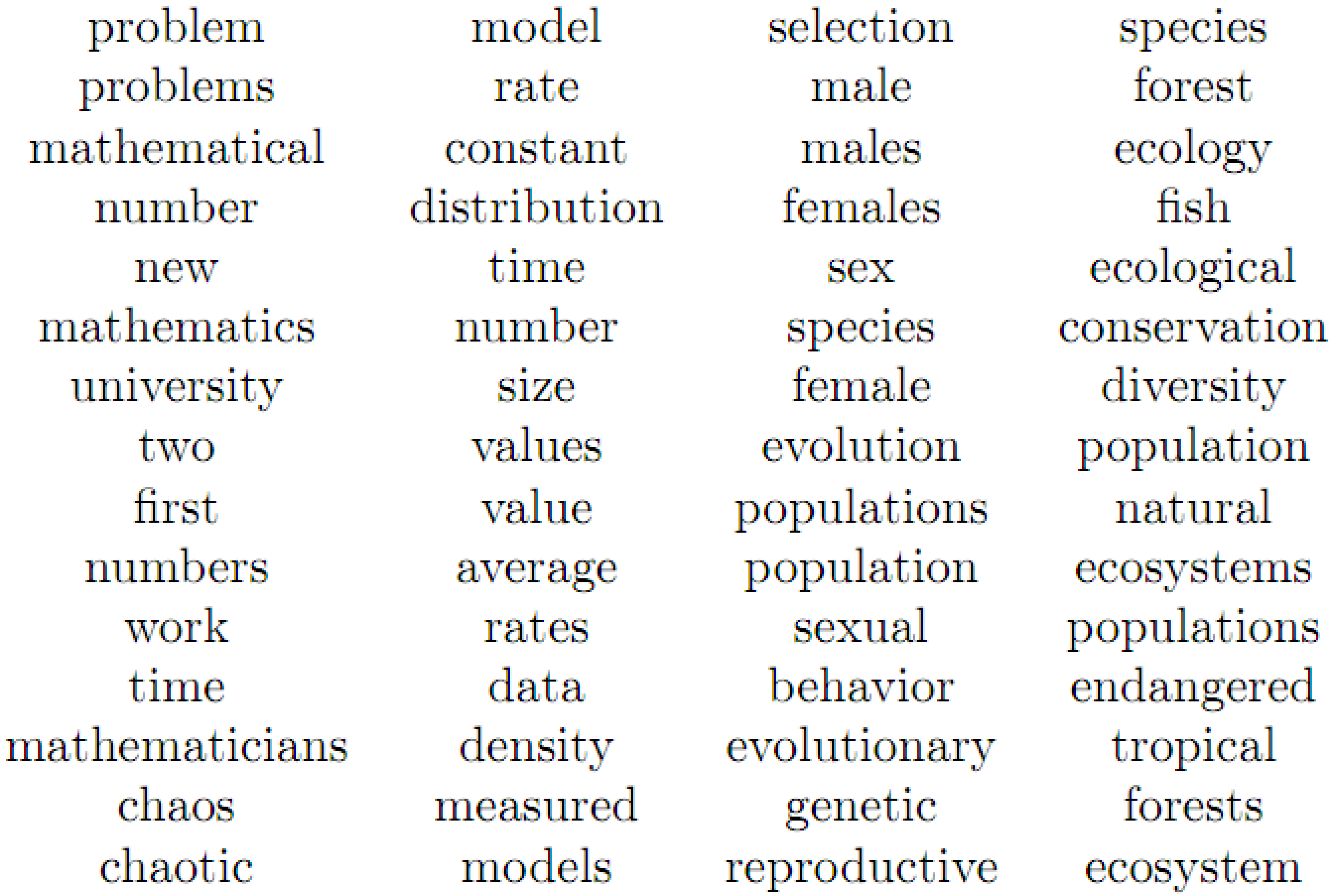
It has proven extremely difficult to demonstrate complex dynamics in populations in the field. By its very nature, a chaotically fluctuating population will superficially resemble a stable or cyclic population buffeted by the normal random perturbations experienced by all species. Given a long enough time series, diagnostic tools from nonlinear mathematics can be used to identify the telltale signatures of chaos. In phase space, chaotic trajectories come to lie on "strange attractors," curious geometric objects with fractal structure and hence noninteger dimension. As they move over the surface of the attractor, sets of adjacent trajectories are pulled apart, then stretched and folded, so that it becomes impossible to predict exact population densities into the future. The strength of the mixing that gives rise to the extreme sensitivity to initial conditions can be measured mathematically estimating the Liapunov exponent, which is positive for chaotic dynamics and nonpositive otherwise. There have been many attempts to estimate attractor dimension and Liapunov exponents from time series data, and some candidate chaotic population have been identified (some insects, rodents, and most convincingly, human childhood diseases), but the statistical difficulties preclude any broad generalization (3).

An alternative approach is to parameterize population models with data from natural populations and then compare their predictions with the dynamics in the field. This technique has been gaining popularity in recent years, helped by statistical advances in parameter estimation. Good ex-

Cannibalism and chaos. The flour beetle, *Tribolium castaneum*, exhibits chaotic population dynamics when the amount of cannibalism is altered in a mathematical model.

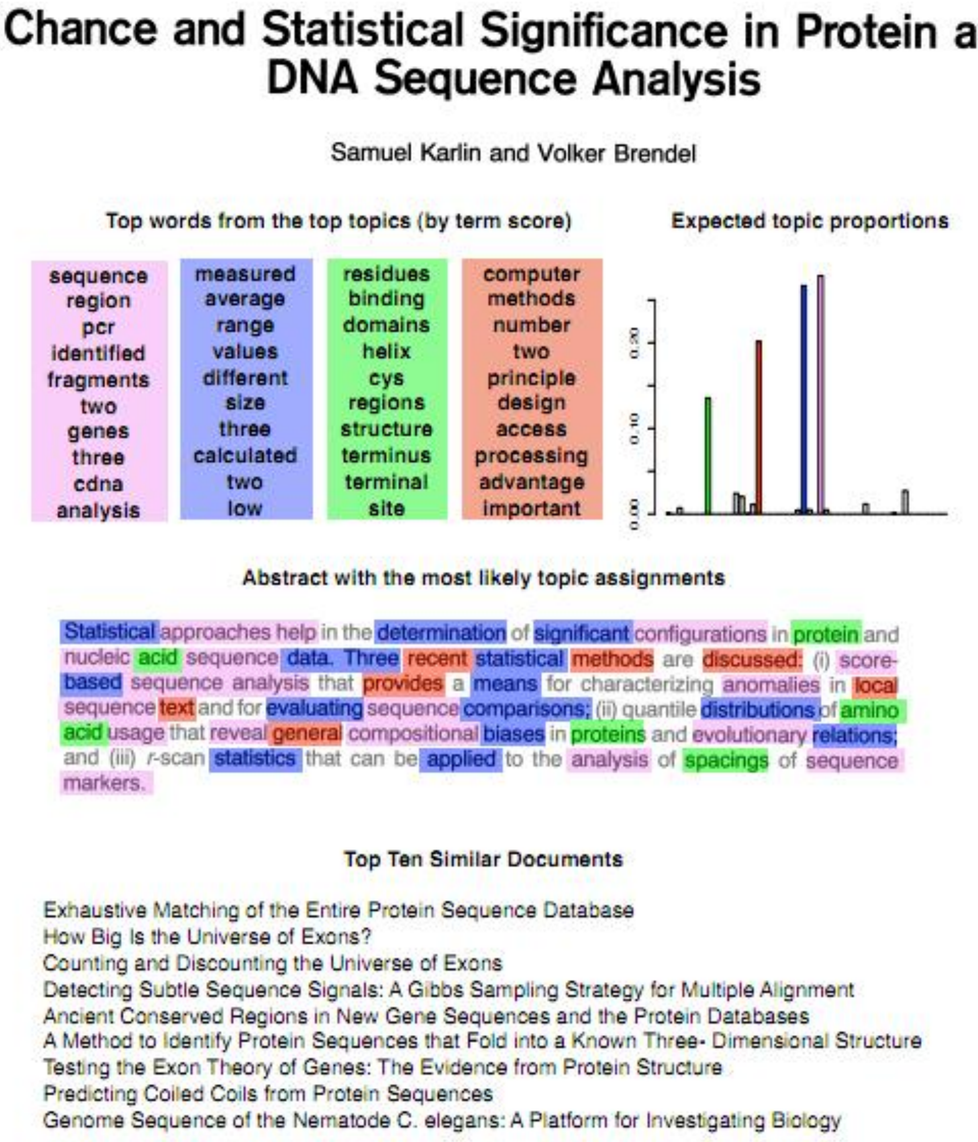
The authors are in the Department of Biology, Imperial College at Silwood Park, Ascot, Berks, SL5 7PZ UK. E-mail: m.hassell@ic.ac.uk

Example inference 2



problem	model	selection	species
problems	rate	male	forest
mathematical	constant	males	ecology
number	distribution	females	fish
new	time	sex	ecological
mathematics	number	species	conservation
university	size	female	diversity
two	values	evolution	population
first	value	populations	natural
numbers	average	population	ecosystems
work	rates	sexual	populations
time	data	behavior	endangered
mathematicians	density	evolutionary	tropical
chaos	measured	genetic	forests
chaotic	models	reproductive	ecosystem

Used to explore and browse document collections



Chance and Statistical Significance in Protein and DNA Sequence Analysis

Samuel Karlin and Volker Brendel

Top words from the top topics (by term score)

sequence	measured	residues	computer
region	average	binding	methods
pcr	range	domains	number
identified	values	helix	two
fragments	different	cys	principle
two	size	regions	design
genes	three	structure	access
three	calculated	terminus	processing
cdna	two	terminal	advantage
analysis	low	site	important

Expected topic proportions

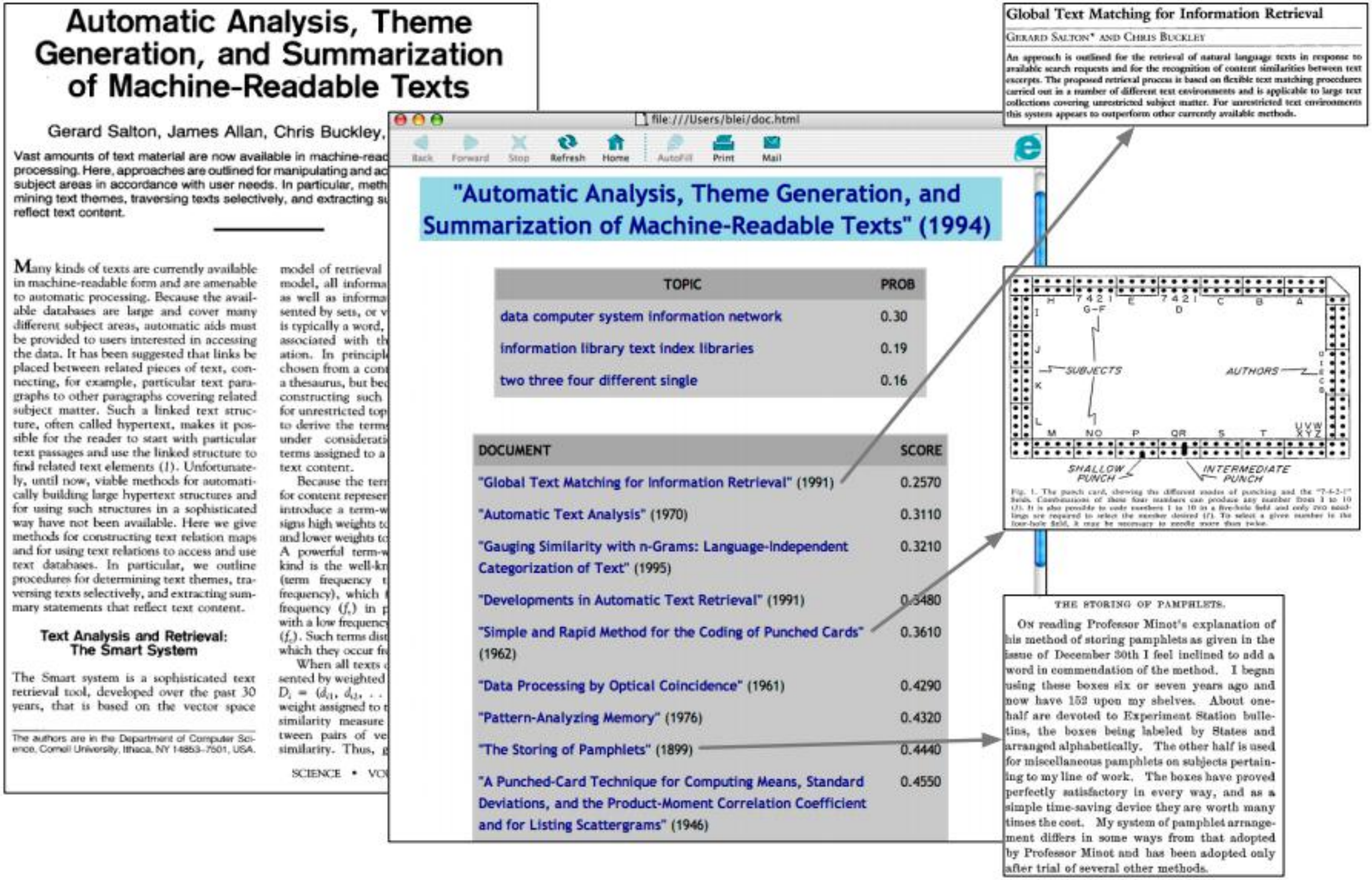
Abstract with the most likely topic assignments

Statistical approaches help in the determination of significant configurations in protein and nucleic acid sequence data. Three recent statistical methods are discussed: (i) score-based sequence analysis that provides a means for characterizing anomalies in local sequence text and for evaluating sequence comparisons; (ii) quantile distributions of amino acid usage that reveal general compositional biases in proteins and evolutionary relations; and (iii) *r*-scan statistics that can be applied to the analysis of spacings of sequence markers.

Top Ten Similar Documents

Exhaustive Matching of the Entire Protein Sequence Database
How Big Is the Universe of Exons?
Counting and Discounting the Universe of Exons
Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment
Ancient Conserved Regions in New Gene Sequences and the Protein Databases
A Method to Identify Protein Sequences that Fold into a Known Three- Dimensional Structure
Testing the Exon Theory of Genes: The Evidence from Protein Structure
Predicting Coiled Coils from Protein Sequences
Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology

Used in exploratory tools of document collections



Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts

Gerard Salton, James Allan, Chris Buckley,

Vast amounts of text material are now available in machine-read processing. Here, approaches are outlined for manipulating and ac subject areas in accordance with user needs. In particular, meth mining text themes, traversing texts selectively, and extracting su reflect text content.

Many kinds of texts are currently available in machine-readable form and are amenable to automatic processing. Because the available databases are large and cover many different subject areas, automatic aids must be provided to users interested in accessing the data. It has been suggested that links be placed between related pieces of text, connecting, for example, particular text paragraphs to other paragraphs covering related subject matter. Such a linked text structure, often called hypertext, makes it possible for the reader to start with particular text passages and use the linked structure to find related text elements (1). Unfortunately, until now, viable methods for automatically building large hypertext structures and for using such structures in a sophisticated way have not been available. Here we give methods for constructing text relation maps and for using text relations to access and use text databases. In particular, we outline procedures for determining text themes, traversing texts selectively, and extracting summary statements that reflect text content.

Text Analysis and Retrieval: The Smart System

The Smart system is a sophisticated text retrieval tool, developed over the past 30 years, that is based on the vector space

model of retrieval
model, all informa
as well as informa
sented by sets, or v
is typically a word,
associated with th
ation. In principl
chosen from a con
a thesaurus, but be
constructing such
for unrestricted to
to derive the term
under considerati
terms assigned to a
text content.

Because the ter
for content represe
introduce a term-w
signs high weights t
and lower weights t
A powerful term-w
kind is the well-kn
(term frequency t
frequency), which
frequency (f_i) in p
with a low frequenc
(f_c). Such terms dis
which they occur fr

When all texts
sented by weighted
$D_i = (d_{i1}, d_{i2}, ..$
weight assigned to t
similarity measure
tween pairs of ve
similarity. Thus, g

The authors are in the Department of Computer Science, Cornell University, Ithaca, NY 14853–7501, USA.

"Automatic Analysis, Theme Generation, and Summarization of Machine-Readable Texts" (1994)

TOPIC	PROB
data computer system information network	0.30
information library text index libraries	0.19
two three four different single	0.16

DOCUMENT	SCORE
"Global Text Matching for Information Retrieval" (1991)	0.2570
"Automatic Text Analysis" (1970)	0.3110
"Gauging Similarity with n-Grams: Language-Independent Categorization of Text" (1995)	0.3210
"Developments in Automatic Text Retrieval" (1991)	0.3480
"Simple and Rapid Method for the Coding of Punched Cards" (1962)	0.3610
"Data Processing by Optical Coincidence" (1961)	0.4290
"Pattern-Analyzing Memory" (1976)	0.4320
"The Storing of Pamphlets" (1899)	0.4440
"A Punched-Card Technique for Computing Means, Standard Deviations, and the Product-Moment Correlation Coefficient and for Listing Scattergrams" (1946)	0.4550

Global Text Matching for Information Retrieval

Gerard Salton* and Chris Buckley

An approach is outlined for the retrieval of natural language texts in response to available search requests and for the recognition of content similarities between text excerpts. The proposed retrieval process is based on flexible text matching procedures carried out in a number of different text environments and is applicable to large text collections covering unrestricted subject matter. For unrestricted text environments this system appears to outperform other currently available methods.

Fig. 1. The punch card, showing the different modes of punching and the "7-4-2-1" fields. Combinations of these four numbers can produce any number from 1 to 10 (1). It is also possible to code numbers 1 to 10 in a five-hole field and only two needlings are required to select the number desired (1). To select a given number in the four-hole field, it may be necessary to needle more than twice.

THE STORING OF PAMPHLETS.

On reading Professor Minot's explanation of his method of storing pamphlets as given in the issue of December 30th I feel inclined to add a word in commendation of the method. I began using these boxes six or seven years ago and now have 152 upon my shelves. About one-half are devoted to Experiment Station bulletins, the boxes being labeled by States and arranged alphabetically. The other half is used for miscellaneous pamphlets on subjects pertaining to my line of work. The boxes have proved perfectly satisfactory in every way, and as a simple time-saving device they are worth many times the cost. My system of pamphlet arrangement differs in some ways from that adopted by Professor Minot and has been adopted only after trial of several other methods.

Variations of LDA

- Multimodal Dirichlet Priors
- Correlated Topic Models
- Hierarchical Dirichlet Processes
- Abstract Tagging in Scientific Journals
- Object Detection/Recognition

Related resources

- **Probabilistic Latent Semantic Analysis.** Thomas Hofmann. Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99)
- **Indexing by latent semantic analysis.** Scott Deerwester et al. Journal of te American Society for Information Science, vol 41, no 6, pp. 391—407, 1990.
- **Latent Dirichlet allocation.** D. Blei, A. Ng, and M. Jordan. Journal of Machine Learning Research, 3:993-1022, January 2003.
- Wiki page for LSA, pLSA, LDA
- Liangjie Hong's personal page for **Notes on Probabilistic Latent Semantic Analysis (PLSA)**
- Thomas Hofmann's and Ata Kaban's slides about pLSA
- Blei's slides about LDA