Topic Model

Pulak Purkait,
SRF, ISI Kolkata

Information overload

As more information becomes available, it becomes more difficult to find and discover what we need.

We need new tools to help us organize, search, and understand these vast amounts of information.

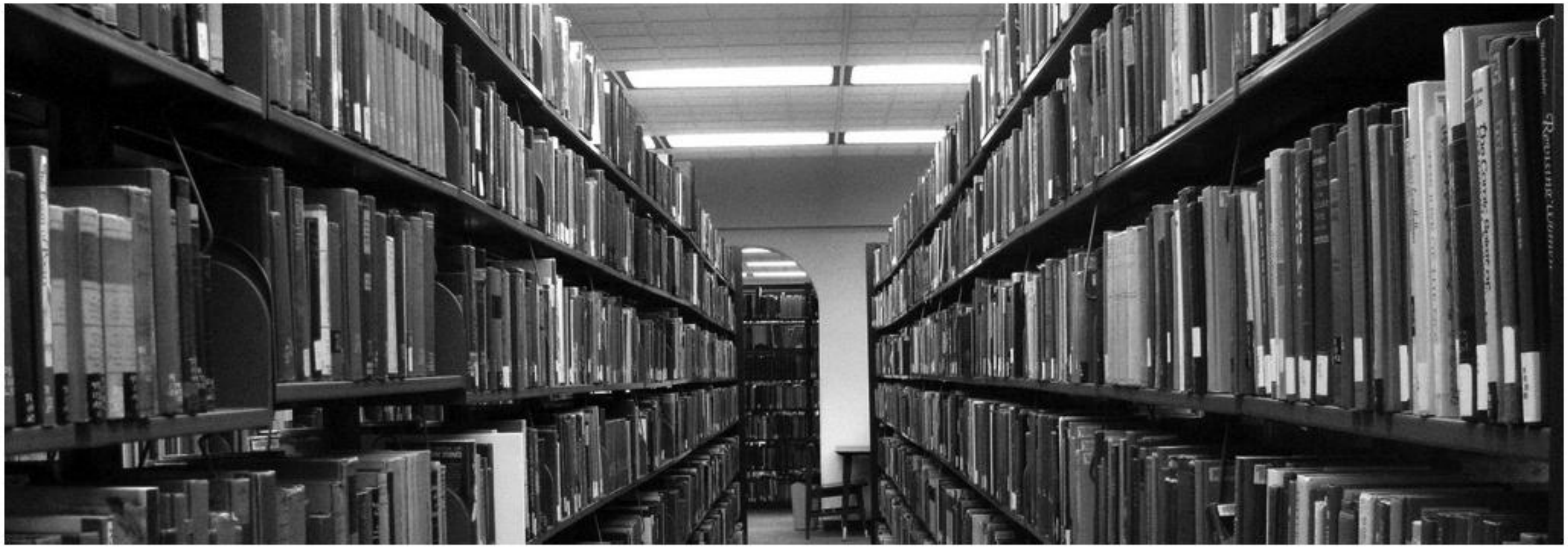
David M. Blei

pLSA(1999) :

- The main challenge a machine learning system has to address roots in the distinction between the lexical level of “what actually has been said or written” and the semantical level of “what was intended” or “what was referred to” in a text or an utterance.

------Thomas Hofmann

Topic modeling



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

1. Discover the hidden themes that pervade the collection.
2. Annotate the documents according to those themes.
3. Use annotations to organize, summarize, and search the texts.

David M. Blei

Annotate images

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

The Netflix problem

- Netflix database
    - About half a million users
    - About 18,000 movies
- People rate movies
- Sparsely sampled entries

The #1 Online DVD Rental Service with over 2,000,000 members!

is the best way to rent movies.

Rent all the DVDs you want for $21.99 a month — **NO LATE FEES!**

- No Late Fees - Ever!
- Over 25,000 Titles - Classics to New Releases.
- Free Shipping Both Ways.
- No Driving. No Lines. No Hassles.
- Always have up to 3 DVDs at home.
- No Commitments. Cancel Anytime.

Have a special offer? Enter Code:

© 1997-2004 Netflix, Inc. All rights reserved. U.S. Patent No. 6584450 (113) USA

Emmanuel Candes

The Netflix problem

- Netflix database
    - About half a million users
    - About 18,000 movies
- People rate movies
- Sparsely sampled entries

$$\begin{array}{cc} & \text{Movies} \\ \text{Users} & \begin{bmatrix} \times & & & \times & \\ & \times & \times & & \\ \times & & \times & & \\ & \times & & & \times \\ \times & & & & \\ & \times & \times & & \end{bmatrix} \end{array}$$

Challenge

Complete the “Netflix matrix”

Many such problems → collaborative filtering, partially filled out surveys...

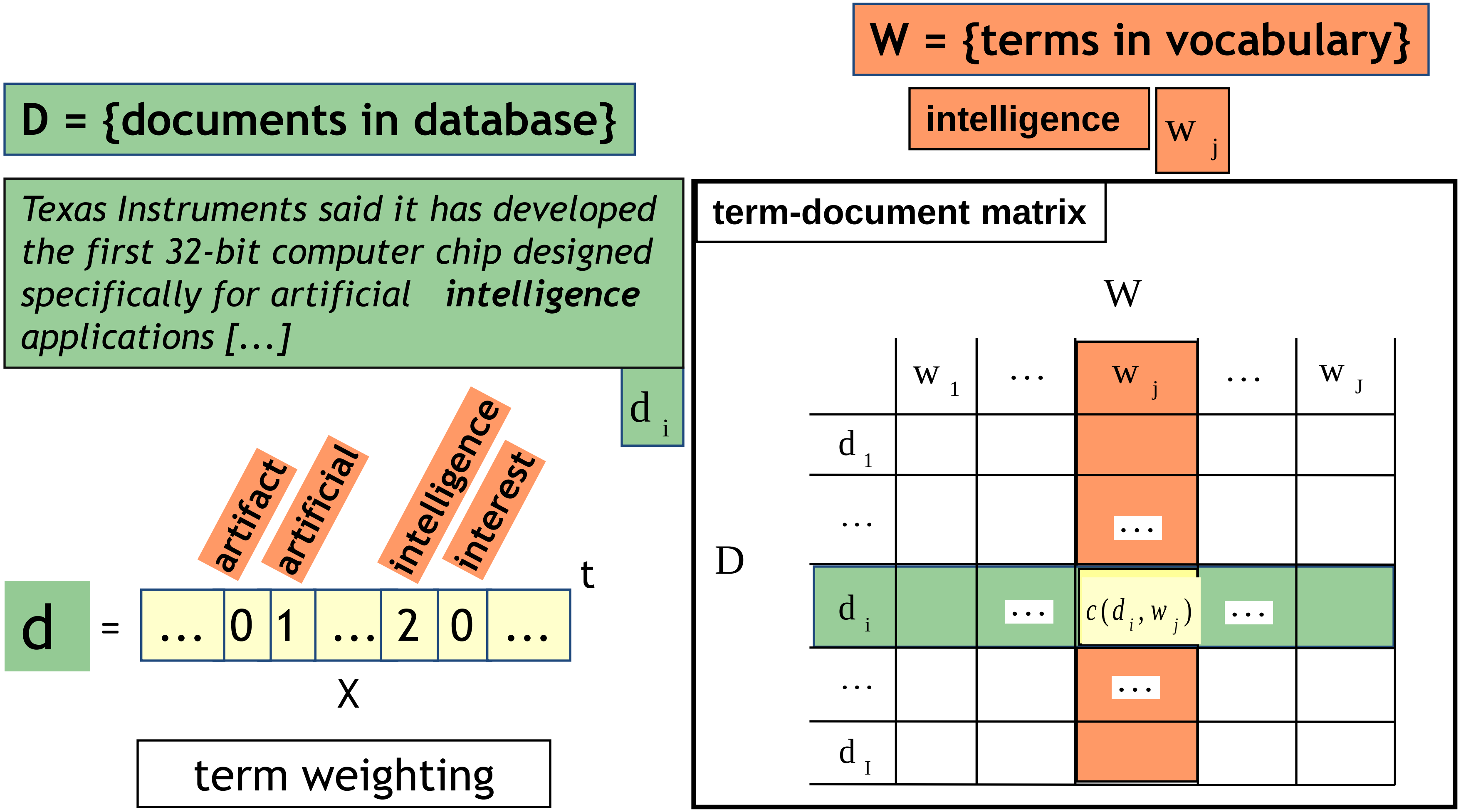
Emmanuel Candes

- Clustering --> Dimensionality reduction
- Low rank approximation
- Collaborative filter
- Matrix completion
- Topic modeling ---> LSA, pLSA, LDA

Vector Space Model

a traditional approach

Term-Document Matrix



W = {terms in vocabulary}

intelligence w_j

D = {documents in database}

Texas Instruments said it has developed the first 32-bit computer chip designed specifically for artificial ***intelligence*** *applications […]*

d_i

$$d = \begin{pmatrix} \ldots & 0 & 1 & \ldots & 2 & 0 & \ldots \end{pmatrix}^t$$

X

term weighting

term-document matrix

$$\begin{array}{c|c|c|c|c|c|c} & & & W & & & \\ & & w_1 & \ldots & w_j & \ldots & w_J \\ \hline & d_1 & & & & & \\ \hline & \ldots & & & \ldots & & \\ \hline D & d_i & & \ldots & c(d_i, w_j) & \ldots & \\ \hline & \ldots & & & \ldots & & \\ \hline & d_I & & & & & \end{array}$$

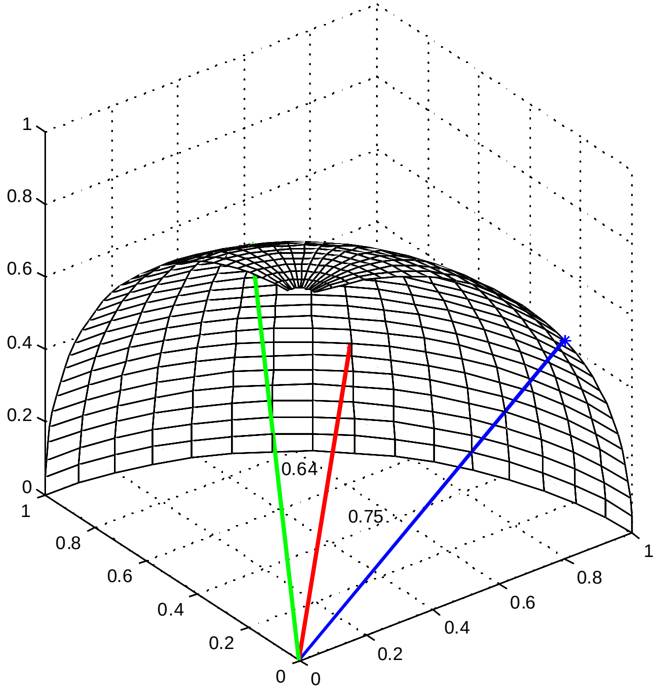
Documents in “Inner” Space

similarity between
document and query

$$\operatorname{sim}(d,q) = \cos\left(\angle(\vec{d},\vec{q})\right) = \frac{\langle \vec{d},\vec{q} \rangle}{\|\vec{d}\|\|\vec{q}\|}$$

- Retrieval method
  - rank documents according to similarity with query
  - term weighting schemes, for example, TFIDF
  - used in SMART system and many successor systems, high popularity

cosine of angle between
query and document(s)



Advantages of the Vector Space Model

- No subjective selection of index terms
- Partial matching of queries and documents *(dealing with the case where no document contains all search terms)*
- Ranking according to similarity score *(dealing with large result sets)*
- Term weighting schemes *(improves retrieval performance)*
- Various extensions
  - Document clustering
  - Relevance feedback (modifying query vector)
- Geometric foundation

Limitations of the Vector Space Model

- 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 …

- Ambiguity and association in natural language
  - **Polysemy**: Words often have a multitude of meanings and different types of usage *(more urgent for very heterogeneous collections)*.
  - The vector space model is unable to discriminate between different meanings of the same word.

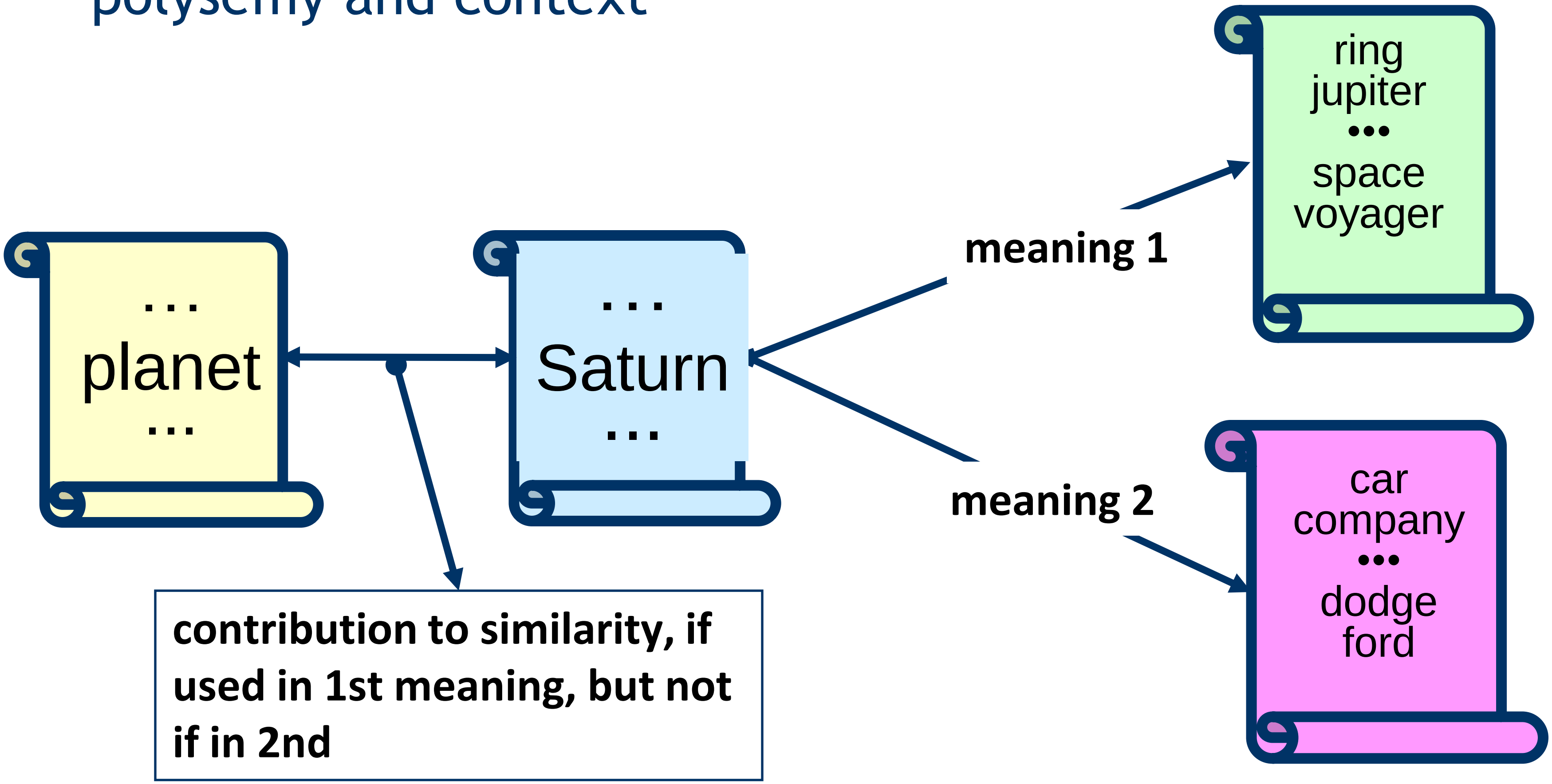
$$\mathrm{sim}(d,q) < \cos\angle(\vec{d},\vec{q})$$

  - **Synonymy**: Different terms may have an identical or a similar meaning (weaker: words indicating the same topic).
  - No associations between words are made in the vector space representation.

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

Polysemy and Context

- Document similarity on single word level: polysemy and context



...
planet
...

...
Saturn
...

meaning 1

ring
jupiter
•••
space
voyager

meaning 2

car
company
•••
dodge
ford

contribution to similarity, if used in 1st meaning, but not if in 2nd

Latent Semantic Analysis (LSA)

An algebraic approach

Latent Semantic Analysis

- ❑ General idea
  - Map documents (and terms) to a low-dimensional representation.
  - Design a mapping such that the low-dimensional space reflects semantic associations (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.

LSA: Matrix Decomposition by SVD

- Dimension reduction by singular value decomposition of term-document matrix

$$\mathbf{C} = (c_{ij}),\ c_{ij} = c(d_i, w_j)$$

word frequencies (possibly transformed)

- Document length normalization
- Sublinear transformation (e.g., log)
- Global term weight

original td matrix

$$\mathbf{C} = \mathbf{U}\Sigma\mathbf{V}^t \approx \mathbf{U}\Sigma\mathbf{V}^t = \hat{\mathbf{C}}$$

reconstructed td matrix

term/document vectors

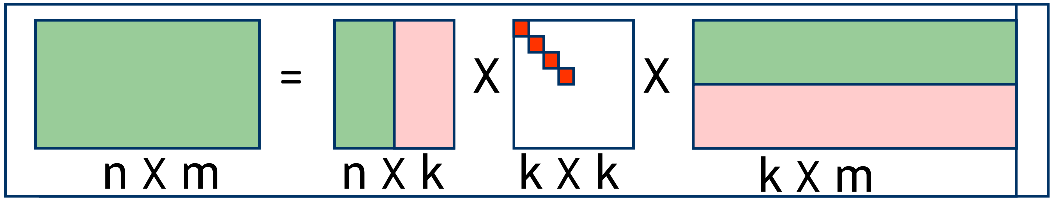
thresholded singular values

L_2 optimal approximation

Background: SVD

- Singular Value Decomposition, definition

$$\hat{\mathbf{C}} = \mathbf{U}\Sigma\mathbf{V}^{t}$$



  - $\mathbf{U}$, $\mathbf{V}$: orthonormal columns
  - Σ : diagonal with singular values (ordered)

- Properties:
  - Existence & uniqueness
  - Thresholding small singular values yields an optimal low-rank approximation (in the sense of the Frobenius norm)

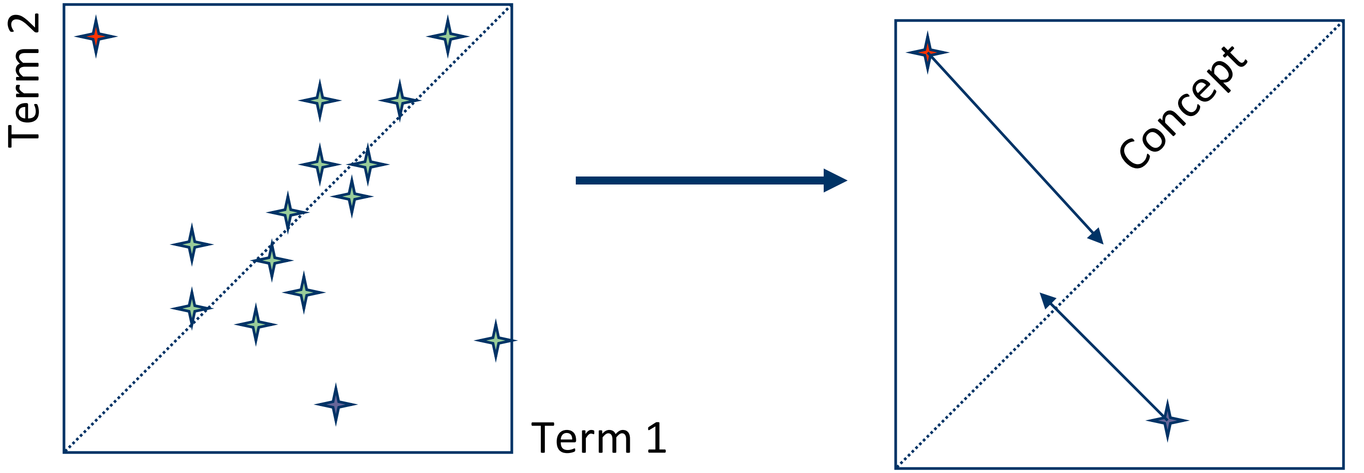
SVD and PCA

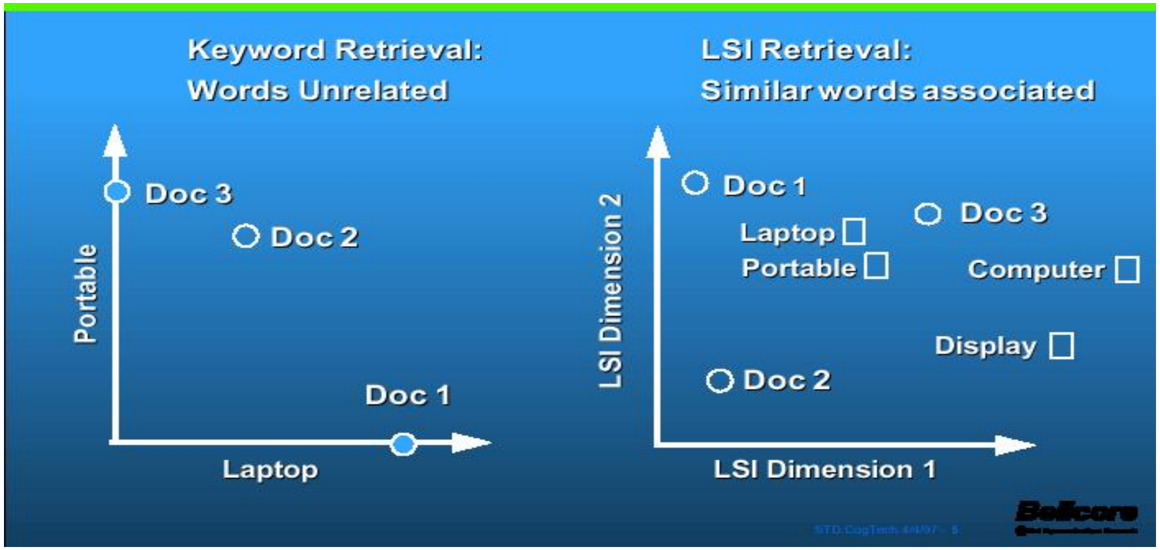
- If (!) the rows of $\mathbf{C}$ would be shifted such that their mean is zero, then:

$$\mathbf{C}\mathbf{C}^{t} = \mathbf{U}\Sigma\mathbf{V}^{t}(\mathbf{U}\Sigma\mathbf{V}^{t})^{t} = \mathbf{U}\Sigma^{2}\mathbf{U}^{t}$$

- Then, one would essentially perform a projection on the principal axis defined by the columns of $\mathbf{U}$
- Yet, this would destroy the sparseness of the term-document matrix (and consequently might hurt the performance of SVD methods)

Term Associations from LSA




Keyword Retrieval:
Words Unrelated

LSI Retrieval:
Similar words associated

(taken from slide by S. Dumais)

LSA: Discussion

- pros:
  - Low-dimensional document representation is able to capture synonyms.
  - Noise removal and robustness by dimension reduction
  - Experimentally: advantages over naïve vector space model
- cons:
  - “Formally”: L_2 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.**
    - **Directions in latent space are hard to interpret.**
    - **No probabilistic model of term occurrences.**
  - [ad hoc selection of the number of dimensions, ...]

Probabilistic Latent Semantic Analysis (pLSA)

A probabilistic approach

Thomas Hofmann

Documents as Information Sources

- ❑ **“real” document**: empirical probability distrib. ≅ relative frequencies

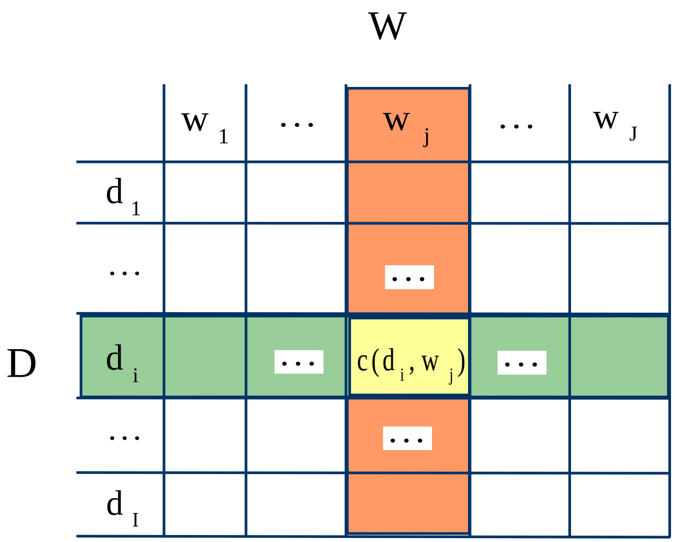
$$\hat{P}(w \mid d) = \frac{c(d,w)}{c(d)}$$

- ❑ **“ideal” document:** (memoryless) information source

other documents → $P(w \mid d) = ?$ **sample**

D = {documents in database}

W = {words in vocabulary}



W

		w_1	…	w_j	…	w_J
	d_1					
	…			…		
D	d_i		…	$c(d_i, w_j)$	…	
	…			…		
	d_I					

Information Source Models in IR

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

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

prior probability of relevance

- Query translation model
  - Probability that q is “generated” from d
  - Probability that query term is generated

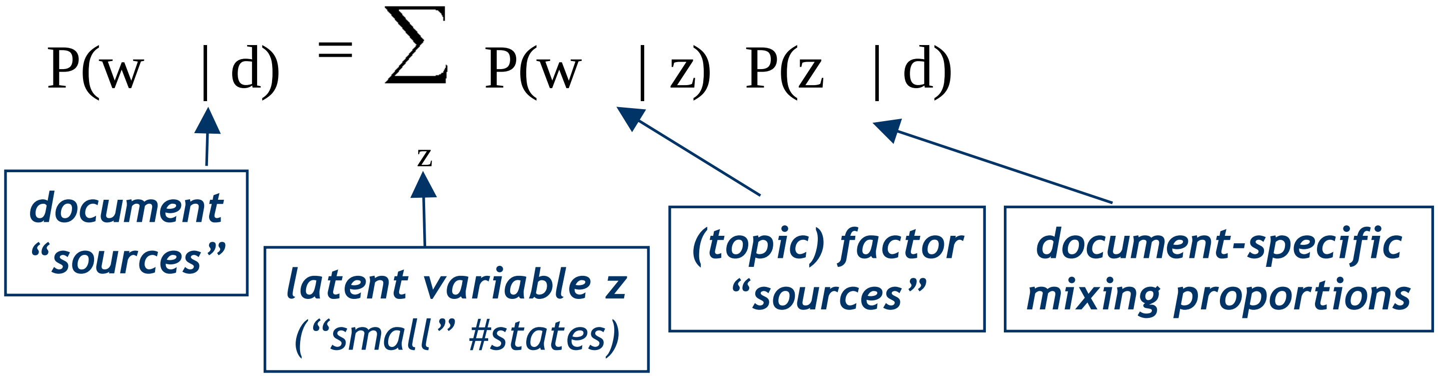
$$P(q \mid d) = \prod_{t \in q} P(t \mid d),$$

$$P(t \mid d) = \sum_{w} \underbrace{P(t \mid w)}_{\text{Translation model}} \; \underbrace{P(w \mid d)}_{\text{Language model}}$$

J. Ponte & W.B. Croft, ”A Language Model Approach to Information Retrieval”, SIGIR 1998.
A. Berger & J. Lafferty, “Information Retrieval as Statistical Translation, SIGIR 1999.

Probabilistic Latent Semantic Analysis

- How can we learn document-specific language models? Sparseness problem, even for unigrams.
- Probabilistic dimension reduction techniques to overcome data sparseness problem.
- Factor analysis for count data: factors ≡ concepts



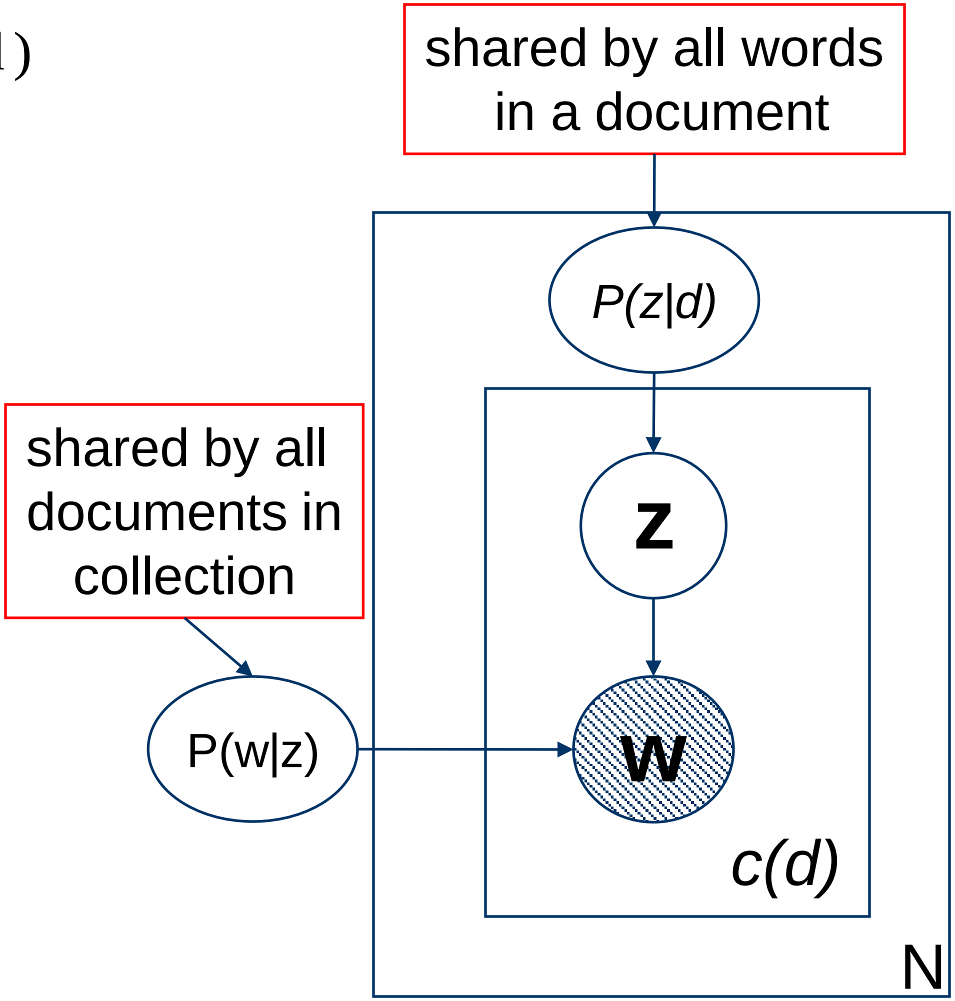
$$P(w \mid d) = \sum_{z} P(w \mid z)\, P(z \mid d)$$

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

T. Hofmann, “Probabilistic Latent Semantic Analysis”, UAI 1999.

PLSA: Graphical Model

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



shared by all words in a document

shared by all documents in collection

Positive Matrix Decomposition

- mixture decomposition in matrix notation

$$\mathbf{C} \approx \mathbf{P}_d \Pi \mathbf{P}_w^t = \tilde{\mathbf{C}} \quad \begin{cases} \Pi = \mathrm{diag}(P(z_1), \ldots, P(z_K)) \\ (\mathbf{P}_d)_{i,k} = P(d_i \mid z_k) \\ (\mathbf{P}_w)_{j,k} = P(w_j \mid z_k) \end{cases}$$

- constraints
  - Non-negativity of all matrices
  - Normalization according to L_1-norm
  - (no orthogonality)

D.D. Lee & H.S. Seung, “Learning the parts of objects by non-negative matrix factorization”, Nature, 1999.

Positive Matrix Decomposition & SVD

- mixture decomposition in matrix notation

$$\mathbf{C} \approx \mathbf{P}_d \, \Pi \, \mathbf{P}_w^t = \tilde{\mathbf{C}} \quad \left\{ \begin{array}{l} \Pi = \mathrm{diag}(P(z_1), \ldots, P(z_K)) \\ (\mathbf{P}_d)_{i,k} = P(d_i \mid z_k) \\ (\mathbf{P}_w)_{j,k} = P(w_j \mid z_k) \end{array} \right.$$

compare to

$$\boxed{\mathbf{C} = \mathbf{U} \Sigma \mathbf{V}^t \approx \mathbf{U} \Sigma \mathbf{V}^t = \hat{\mathbf{C}}}$$

- probabilistic approach vs. linear algebra decomposition
  - conditional independence assumption “replaces” outer product
  - class-conditional distributions “replace” left/right eigenvectors
  - **maximum likelihood instead of minimum L_2 norm criterion**

$$L = \sum_{i,j} c_{ij} \log \tilde{c}_{ij} = \sum_{i,j} c_{ij} \log \sum_{z} P(w_j \mid z) P(d_i \mid z) P(z)$$

Expectation Maximization Algorithm

- Maximizing log-likelihood by (tempered) EM iterations
- E-step (posterior probabilities of latent variables)
- M-step (max. of expected complete log-likelihood)

$$P(z \mid d, w) = \frac{P(d \mid z) P(w \mid z) P(z)}{\sum_{z'} P(d \mid z') P(w \mid z') P(z')}$$

probability that a term occurrence w within d is explained by topic z“

$$P(w \mid z) \propto \sum_{d} c(d, w) P(z \mid d, w)$$

$$P(d \mid z) \propto \sum_{w} c(d, w) P(z \mid d, w) \qquad P(z) \propto \sum_{d, w} c(d, w) P(z \mid d, w)$$

Example: Science Magazine Papers

- Dataset with approx. 12K papers from Science Magazine
- Selected concepts from model with K=200

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

Collaborative Filtering/ Content-Based Filtering/ Matrix Completion

Collaborative and Content-Based Filtering

- **Collaborative/social filtering**
  - Properties of persons or similarities between persons are used to improve predictions.
  - Makes use of user profile data
  - Formally: starting point is sparse matrix with user ratings

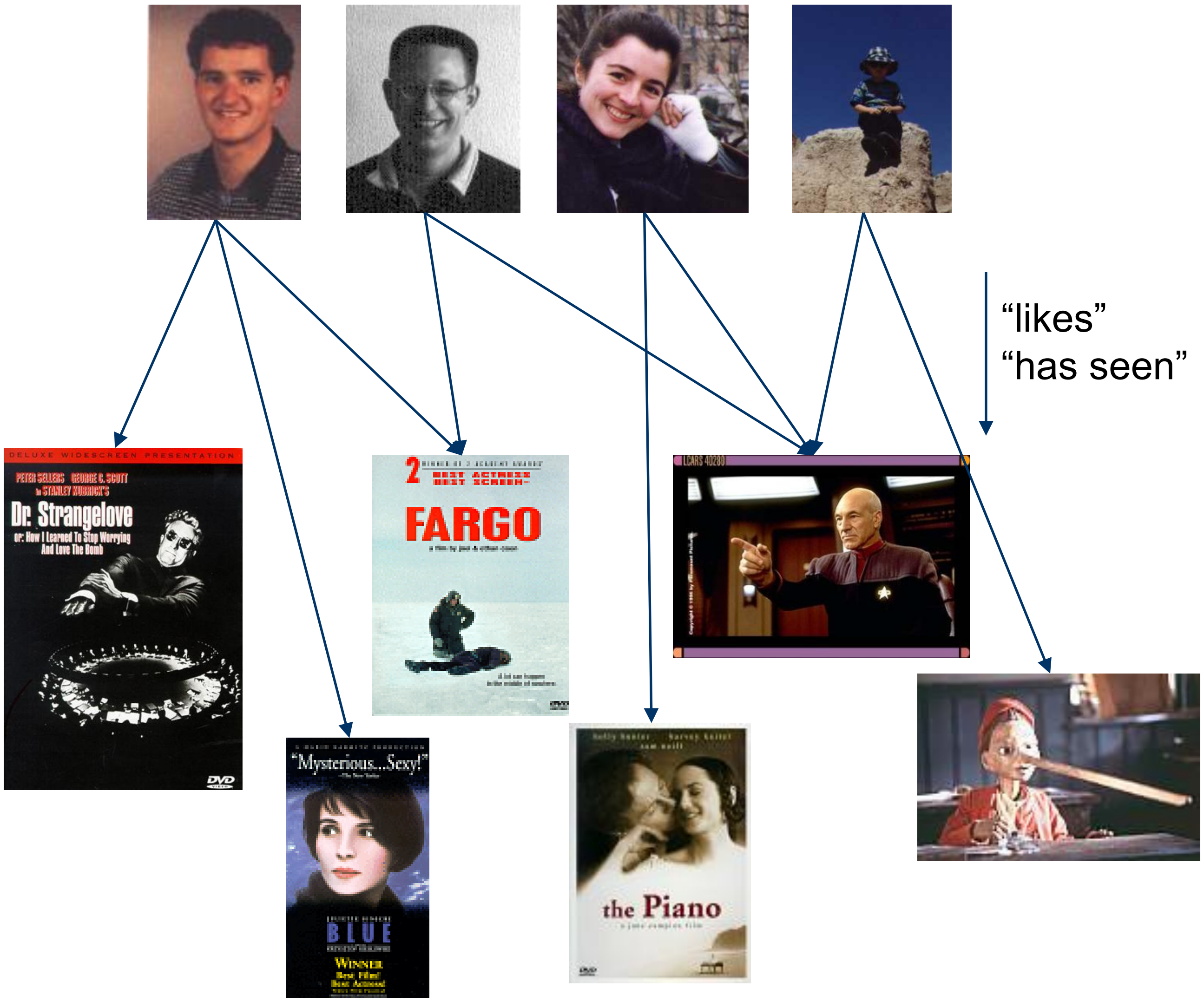
- **Content-based filtering**
  - properties of objects or similarities between objects are used to improve predictions

Personalized Information Filtering:

Users/ Customers

Judgement/ Selection

Objects



"likes"
"has seen"

Predicting Preferences and Actions

User Profile

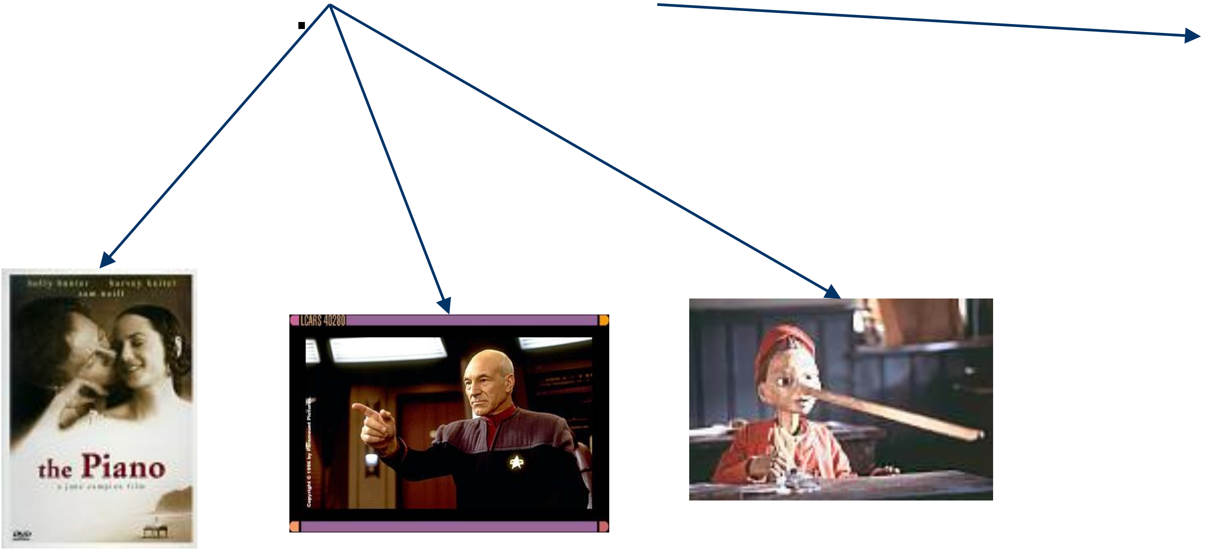
Dr. Strangeloves *****

Three Colors: Blue *****

Fargo *****

Pretty Woman *

Movie? Rating?



*

**

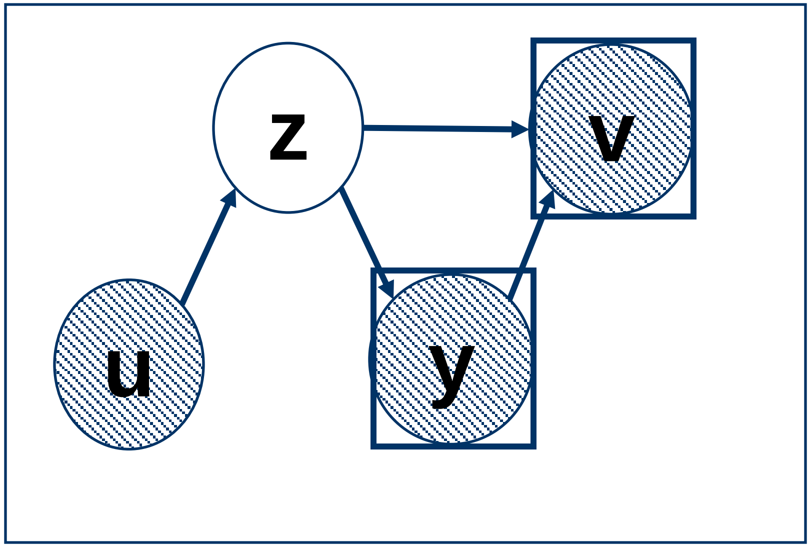
***

****

*****

PLSA for Predicting User Ratings

Multi-valued (or real-valued) rating $v \in \{0,1,2,3,4,5\}$



preference v is independent of person u, given latent state z
"community-based" variant

- Each user is represented by a specific probability distribution
- Analogy to IR [user=document], [items=terms]

Probabilistic Topic Model

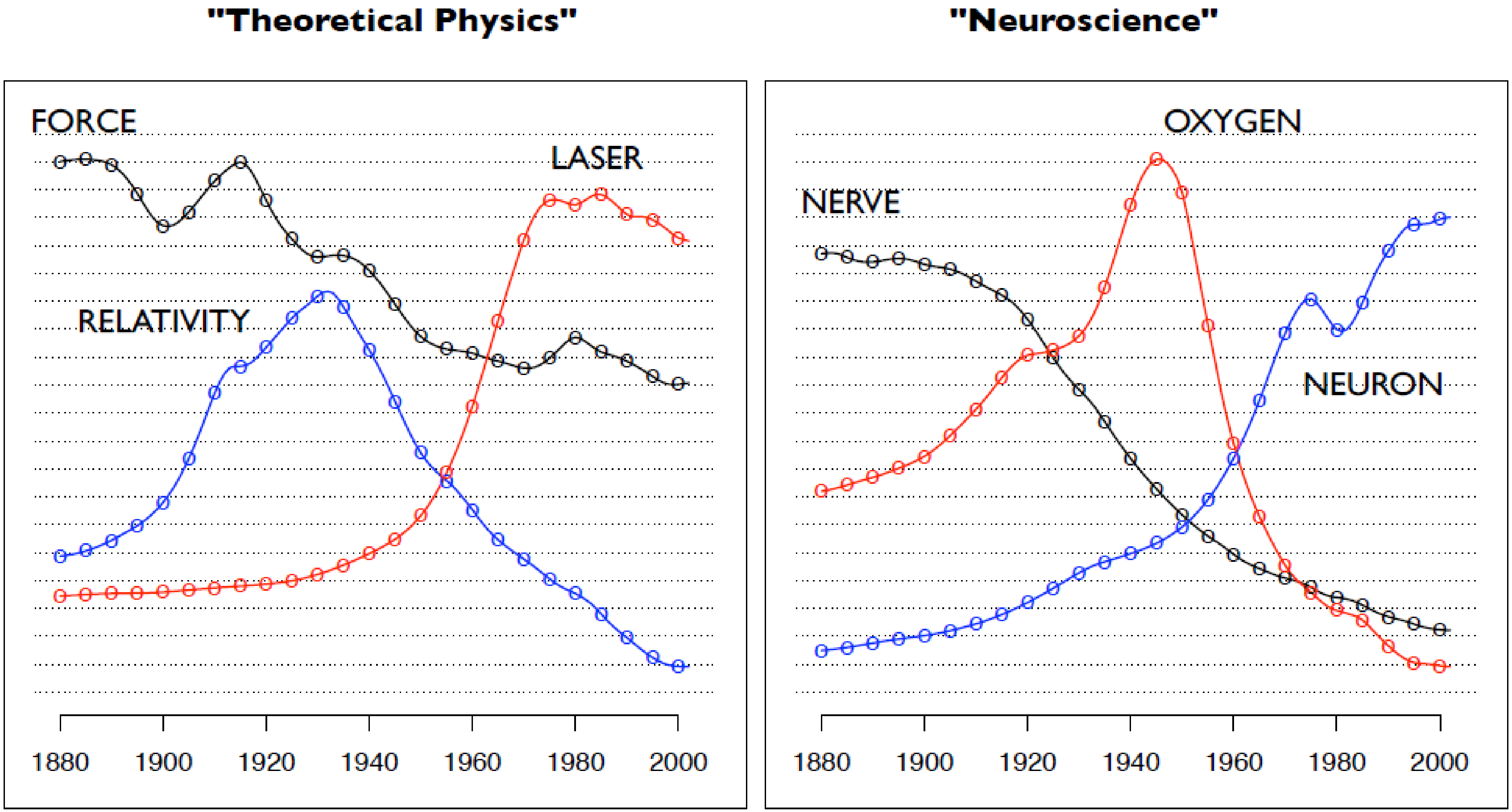
Latent Dirichlet Allocation

Devid M. Blei ‘03

Discover topics from a corpus

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

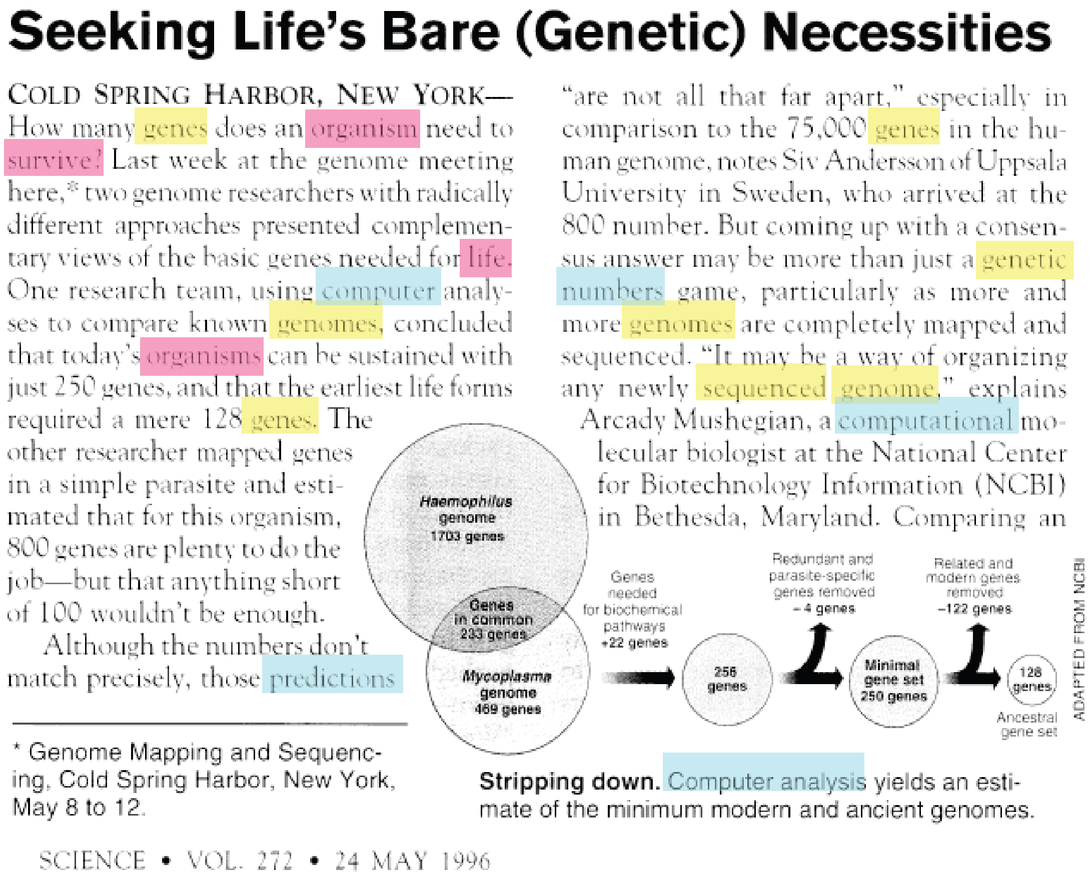
Model the evolution of topics over time



Probabilistic modeling

1. Data are assumed to be observed from a generative probabilistic process that includes hidden variables.
    - *In text, the hidden variables are the thematic structure.*
2. Infer the hidden structure using posterior inference
    - *What are the topics that describe this collection?*
3. Situate new data into the estimated model.
    - *How does a new document fit into the topic structure?*

Latent Dirichlet allocation (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

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

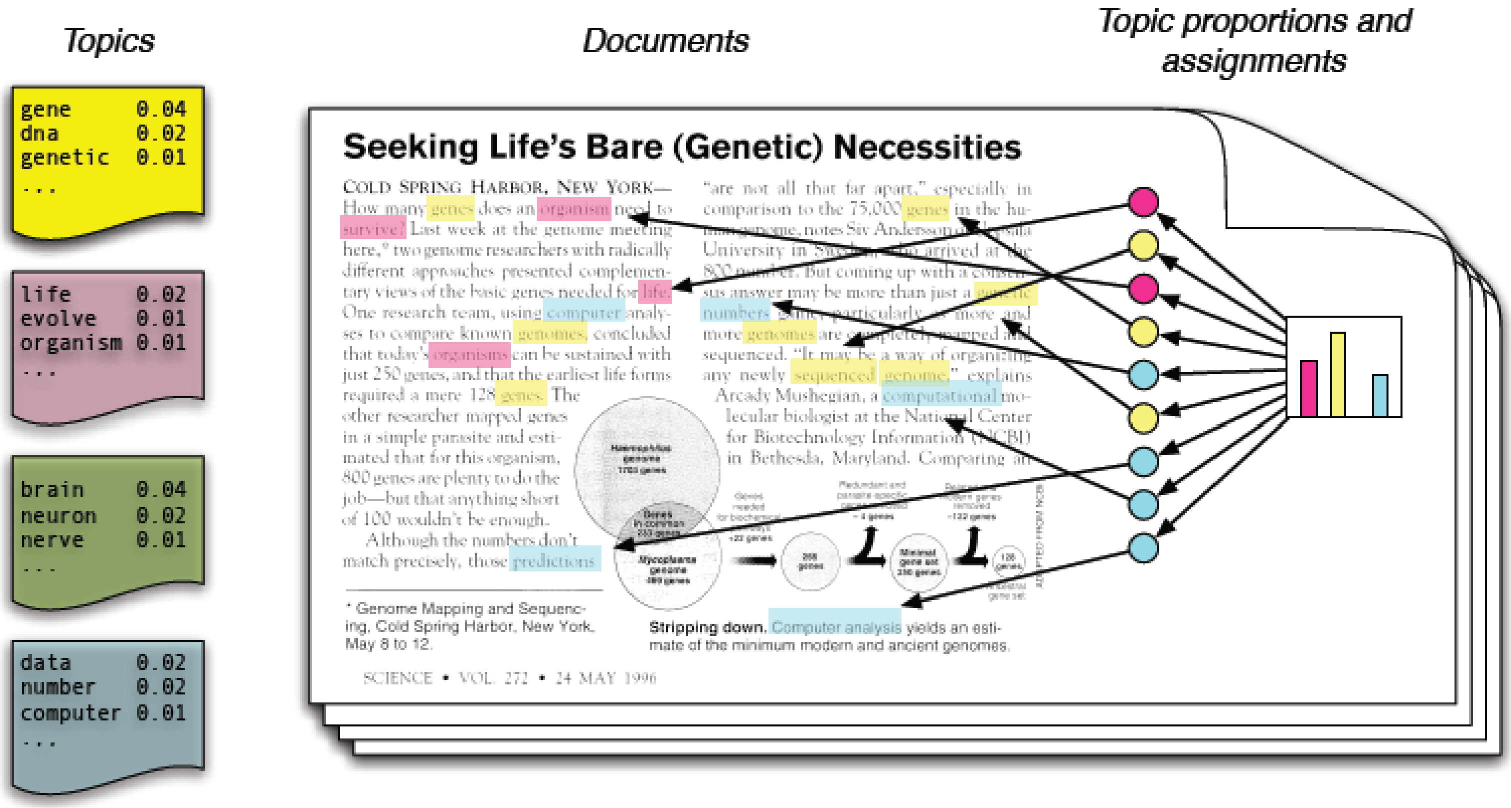
---

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

SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

Generative model for LDA



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

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

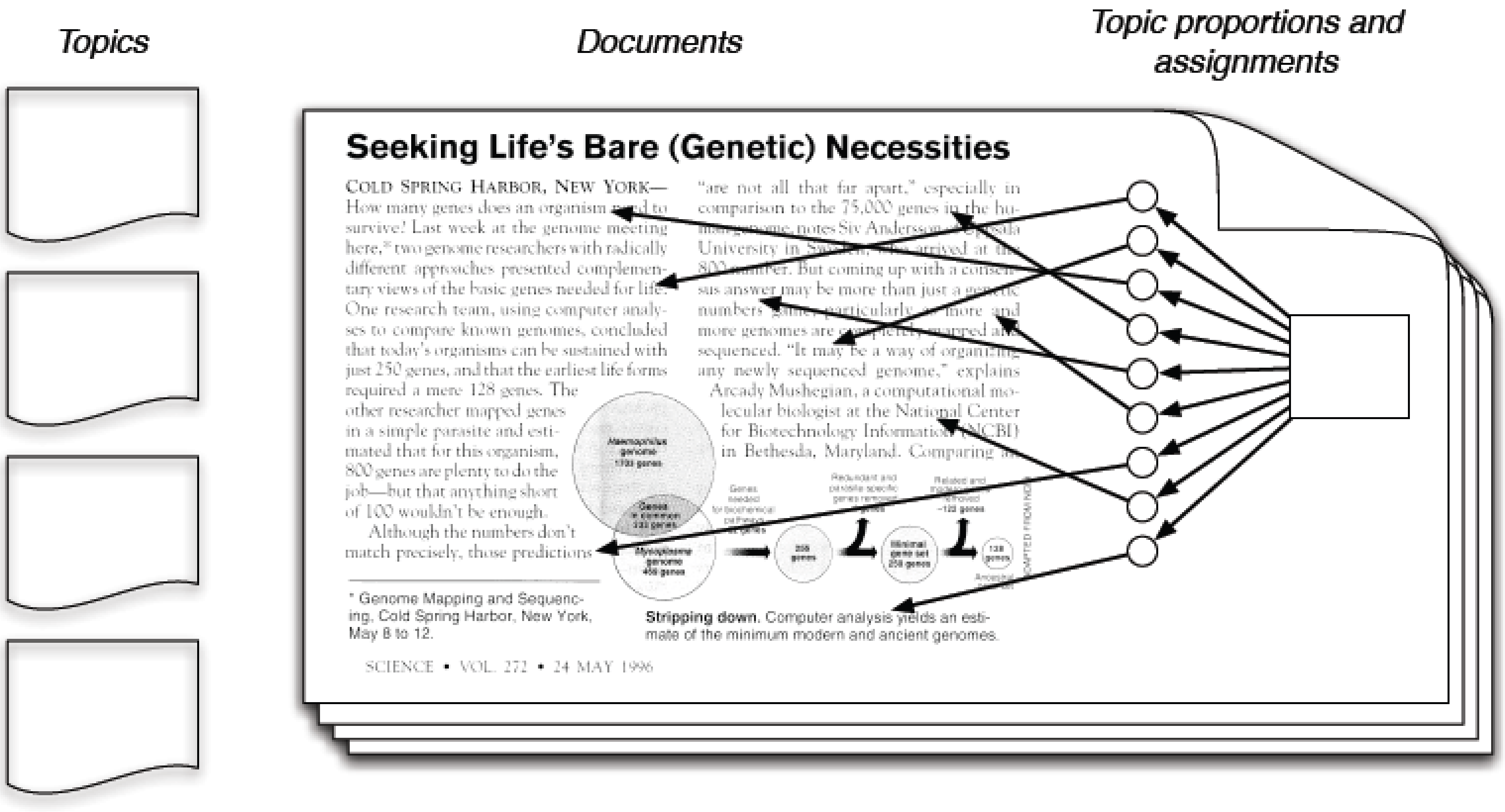
---

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

SCIENCE • VOL. 272 • 24 MAY 1996

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

The posterior distribution



Topics

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

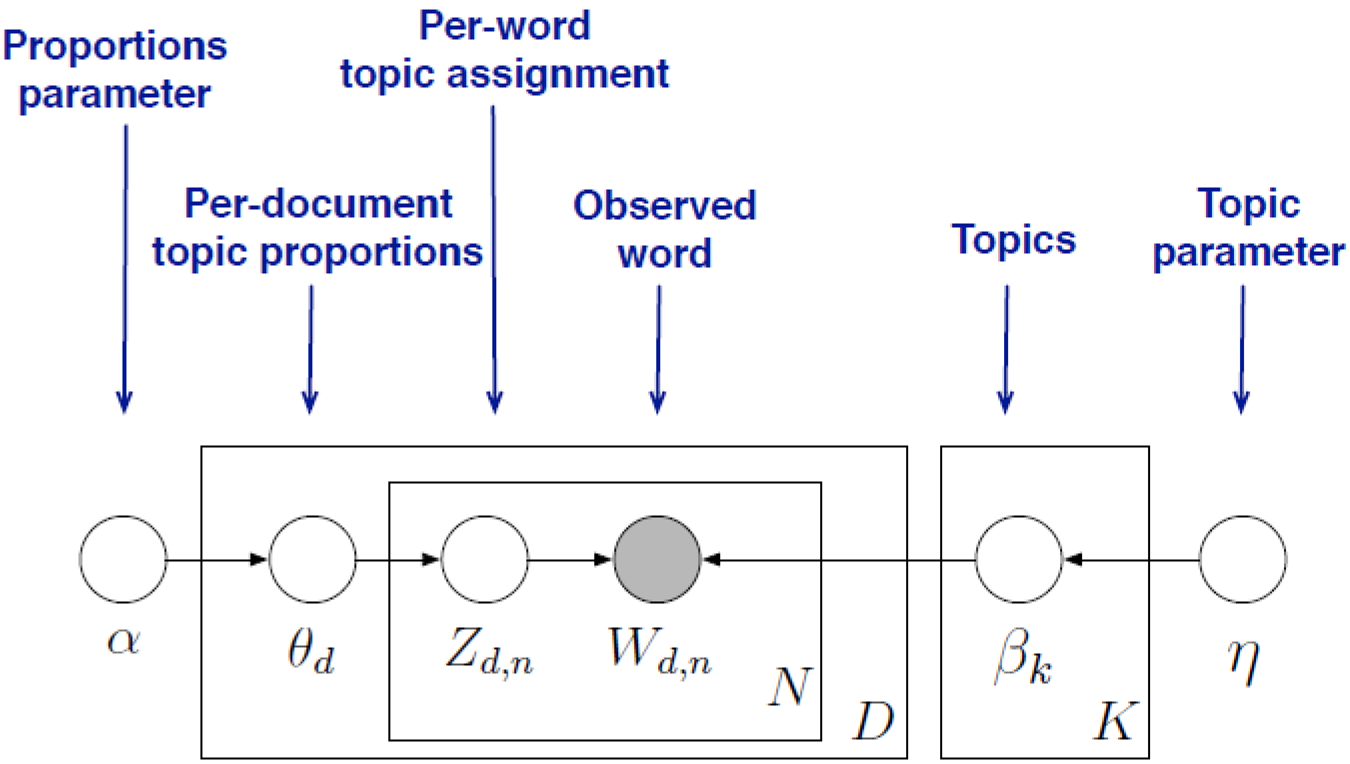
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- In reality, we only observe the documents
- The other structure are **hidden variables**

LDA as a graphical model



Proportions parameter

Per-word topic assignment

Per-document topic proportions

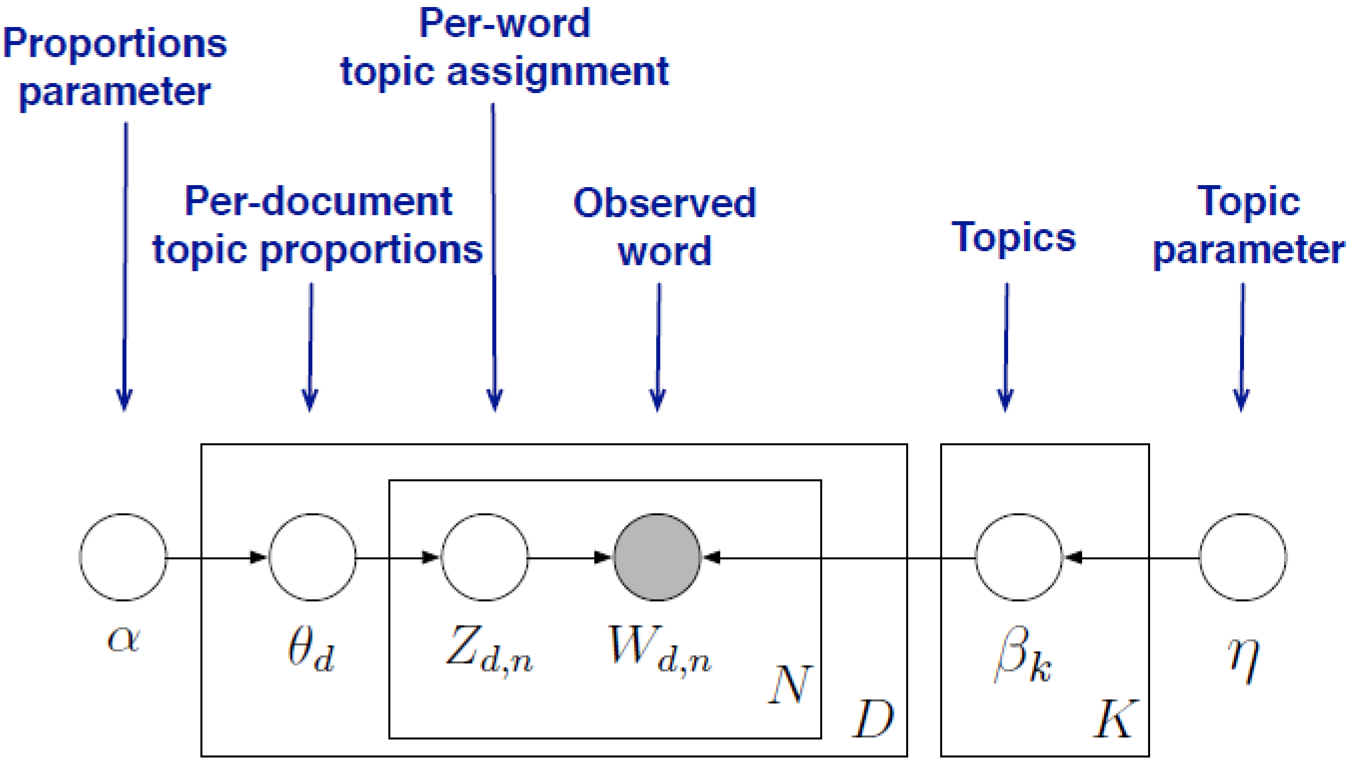
Observed word

Topics

Topic parameter

- Encodes our assumptions about the data
- Connects to algorithms for computing with data
- See *Pattern Recognition and Machine Learning* (Bishop, 2006).

LDA as a graphical model



Proportions parameter

Per-document topic proportions

Per-word topic assignment

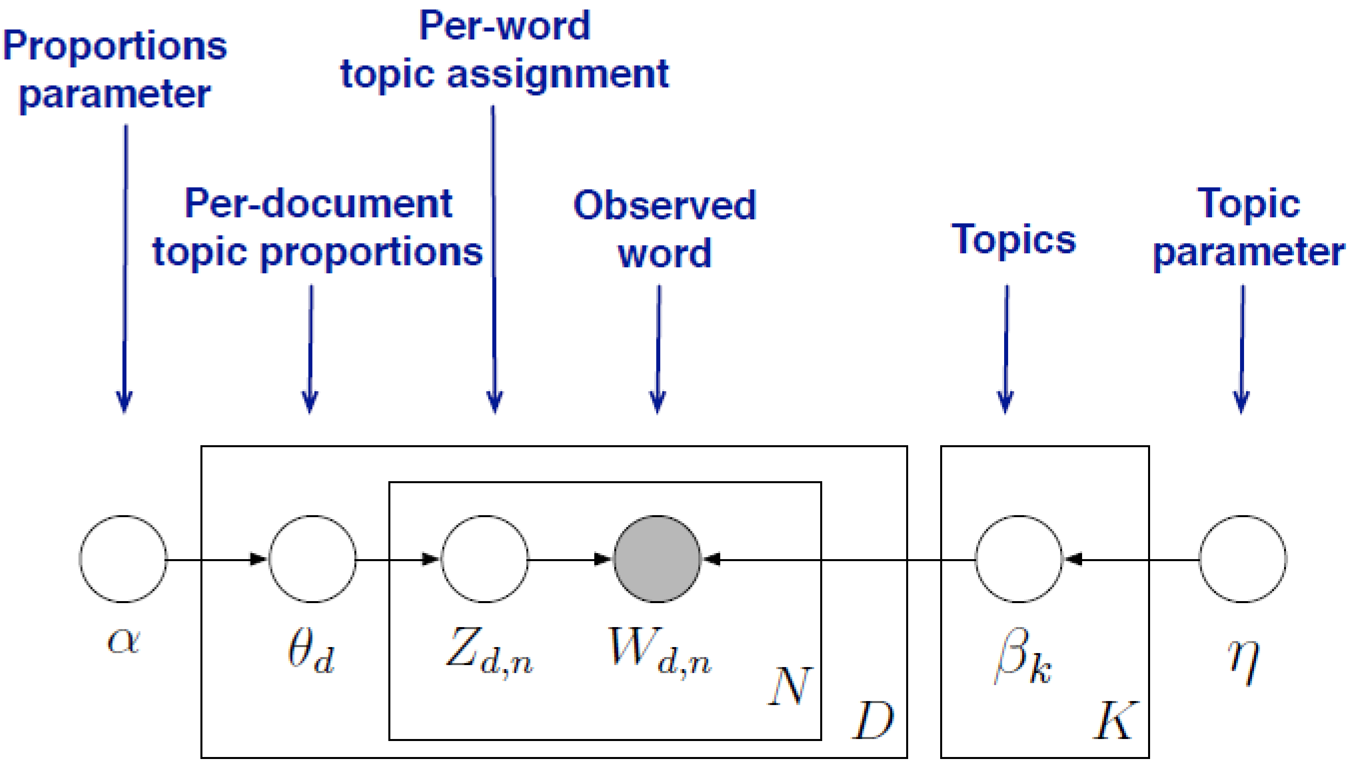
Observed word

Topics

Topic parameter

- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed.
- Plates indicate replicated variables.

LDA as a graphical model



Proportions parameter

Per-document topic proportions

Per-word topic assignment

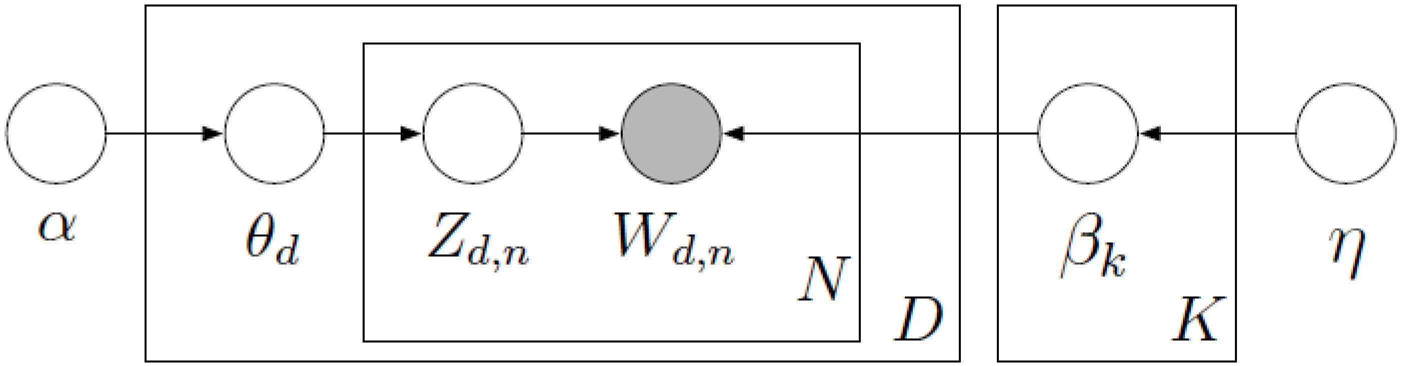
Observed word

Topics

Topic parameter

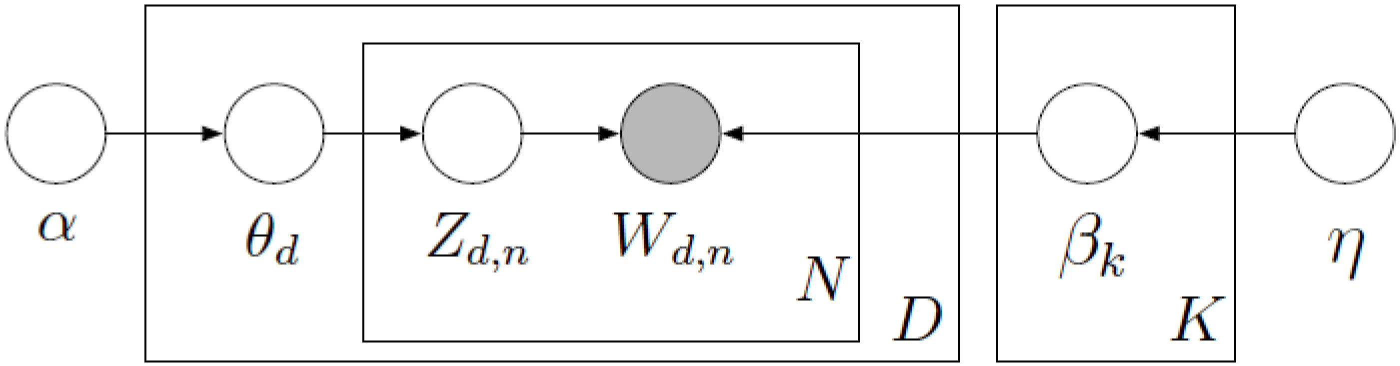
$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

LDA



- This joint defines a posterior.
- From a collection of documents, infer
  - Per-word topic assignment $z_{d,n}$
  - Per-document topic proportions θ_d
  - Per-corpus topic distributions β_k
- Then use posterior expectations to perform the task at hand, e.g., information retrieval, document similarity, exploration, ...

LDA



Approximate posterior inference algorithms

- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (Teh et al., 2006)
- Online variational inference (Hoffman et al., 2010)

Also see Mukherjee and Blei (2009) and Asuncion et al. (2009).

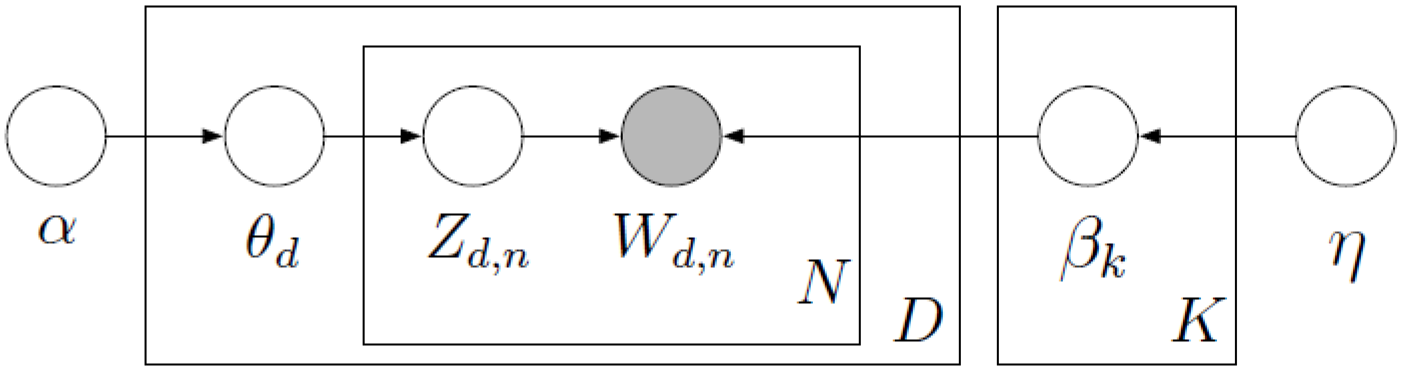
Aside: The Dirichlet distribution

- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

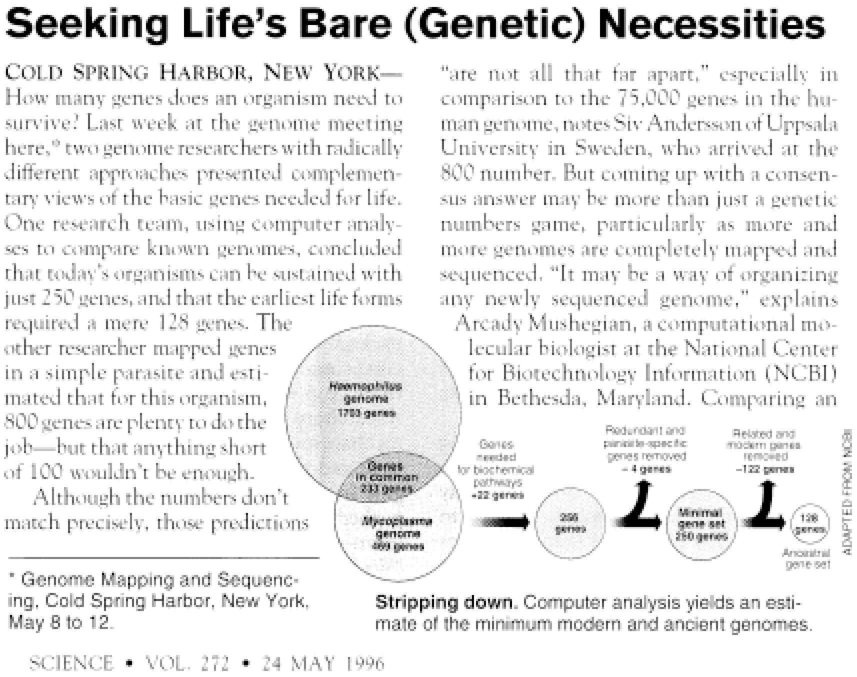
- It is **conjugate** to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ.
- The topic proportions are a K dimensional Dirichlet.
  The topics are a V dimensional Dirichlet.

Example inference



- **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



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.

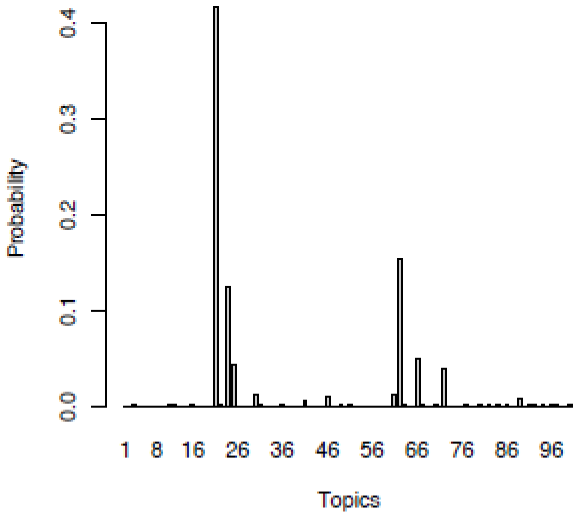
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

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

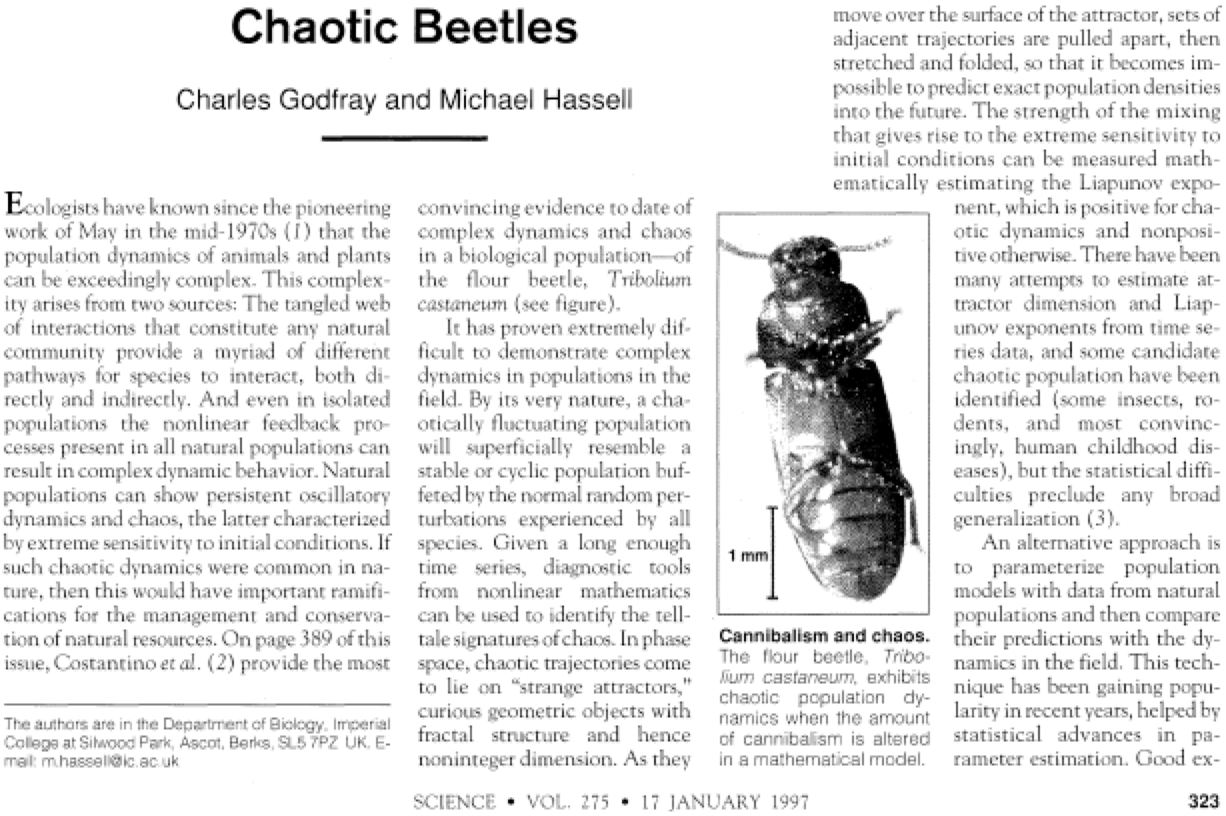
human
genome
dna
genetic
genes
sequence
gene
molecular
sequencing
map
information
genetics
mapping
project
sequences

evolution
evolutionary
species
organisms
life
origin
biology
groups
phylogenetic
living
diversity
group
new
two
common

disease
host
bacteria
diseases
resistance
bacterial
new
strains
control
infectious
malaria
parasite
parasites
united
tuberculosis

computer
models
information
data
computers
system
network
systems
model
parallel
methods
networks
software
new
simulations

Example inference (II)



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-

Canniballism 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 (II)

problem
problems
mathematical
number
new
mathematics
university
two
first
numbers
work
time
mathematicians
chaos
chaotic

model
rate
constant
distribution
time
number
size
values
value
average
rates
data
density
measured
models

selection
male
males
females
sex
species
female
evolution
populations
population
sexual
behavior
evolutionary
genetic
reproductive

species
forest
ecology
fish
ecological
conservation
diversity
population
natural
ecosystems
populations
endangered
tropical
forests
ecosystem

Implementations of LDA

There are many available implementations of topic modeling—

LDA-C[*]	A C implementation of LDA
HDP[*]	A C implementation of the HDP (“infinite LDA”)
Online LDA[*]	A python package for LDA on massive data
LDA in R[*]	Package in R for many topic models
LingPipe	Java toolkit for NLP and computational linguistics
Mallet	Java toolkit for statistical NLP
TMVE[*]	A python package to build browsers from topic models

[*] available at www.cs.princeton.edu/∼blei/