Latent Semantic Analysis and Topic Modeling: Roads to Text Meaning

Hồ Tú Bảo

Japan Advanced Institute of Science and Technology

Vietnamese Academy of Science and Technology

Archeology of computational linguistics

- **1990s–2000s: Statistical learning**
  - algorithms, evaluation, corpora
- **1980s: Standard resources and tasks**
  - Penn Treebank, WordNet, MUC
- **1970s: Kernel (vector) spaces**
  - clustering, information retrieval (IR)
- **1960s: Representation Transformation**
  - Finite state machines (FSM) and Augmented transition networks (ATNs)
- **1960s: Representation**—beyond the word level
  - lexical features, tree structures, networks

Internet and Web in 1990s

- **Natural language processing**
- **Information retrieval and extraction on the Web**

(adapted from E. Hovy, COLING 2004)

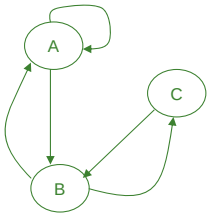
PageRank algorithm (Google)

- Google the word 'weather forecast' → Answer: 4.2 million pages.
- How does Google know which pages are the most important?
- Google assigns a number to each individual page (PageRank number) computed via the eigenvalue problem

$$\mathbf{Pw = \lambda w}$$

- Current size of P: $\mathbf{4.2 \times 10^9}$

Larry Page, Sergey Brin



	A	B	C
A	1/2	1/2	0
B	1/2	0	1
C	0	1/2	0

Latent semantic analysis & topic models

- The LSA approach makes three claims

  (1) semantic information can be derived from a word-document co-occurrence matrix;

  (2) dimensionality reduction is an essential part of this derivation;

  (3) words and documents can be represented as points in Euclidean space.

- Different from (3), topic models express the semantic information of words and documents by 'topics'.

'Latent' = 'hidden', 'unobservable', 'presently inactive', …

What is topic?

- The subject matter of a speech, text, meeting, discourse, etc.
- The topic of a text captures "what a document is about", i.e., the meaning of the text.
- A text can be represented by a "bag of words" for several purposes and you can see the words.
- But how can you see (know) the topics of the text? How a topic is represented, discovered, etc.?

Topic modeling = Finding 'word patterns' of topic

A 'topic' consists of a cluster of words that frequently occur together.

Notation and terminology

- A word is the basic unit of discrete data, from vocabulary indexed by {1,…,V} = **V**. The vth word is represented by a V-vector w such that $w^v = 1$ and $w^u = 0$ for $u \neq v$
- A document is a sequence of N words denote by **d** = $(w_1, w_2, \ldots, w_N)$
- A corpus is a collection of M documents denoted by **D** = $\{\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_M\}$

Term frequency–inverse document frequency

- tf-idf of a word t_i in document d_j (Salton & McGill, 1983)

$$\frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

- Results in a txd matrix – thus reducing the corpus to a fixed-length list
- Used for search engines

$n_{i,j}$ = # times t_i occurs in d_j

Vector space model in IR

	d1	d2	d3	d4	d5	d6	q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

$$\cos(x, y) = \frac{x.y}{|x||y|}$$

cos(d3, q1) = 0
cos(d5, q1) = 0
cos(d4, q1) ≠ 0
cos(d6, q1) ≠ 0

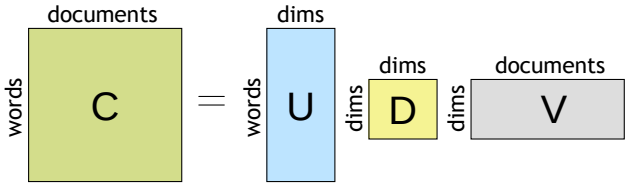
- Given a query, says, q1 = ('rock', 'marble') → d3 more relevant to q1 than d4, d6 even cos(d3, q1) = 0.
- Problem of **synonymy** (one meaning can be expressed by multiple words, e.g. 'group', 'cluster'), and **polysemy** (a word can have multiple meanings, e.g. 'rock').

LSI: Latent semantic indexing

(Deerwester et al., 1990)

- LSI is a dimensionality reduction technique that projects documents to a lower-dimensional semantic space and, in doing so, causes documents with similar topical content to be close to one another in the resulting space.
- In particular, two documents which share no terms with each other directly, but which do share many terms with third document, will end up being similar in the projected space.
- Similarity between LSI and PCA?

LSI: Latent semantic indexing



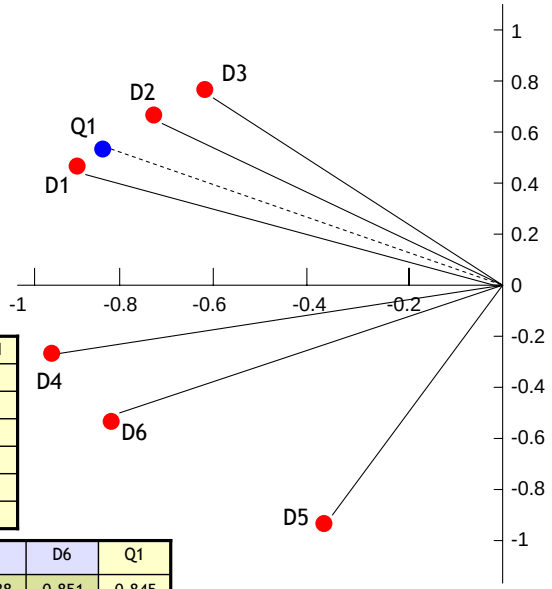
C = UDV^T

dims = # singular values = # (absolute) values of eigenvalues

- **C = UDV^T** by singular value decomposition such that $UU^T = I$ and $VV^T = I$ and D is a diagonal matrix whose diagonal entries are the singular values of C.
- **Idea of LSI**: to strip away most of dimensions and only keep those which capture the most variation in the document collection (typically, from |V| = hundreds of thousands to k = between 100 and 200).

LSI: Example

- LSI clusters documents in the reduced-dimension semantic space according to word co-occurrence patterns.
- Dimensions loosely correspond with topic boundaries.



	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0

	D1	D2	D3	D4	D5	D6	Q1
Dim. 1	-0.888	-0.759	-0.615	-0.961	-0.388	-0.851	-0.845
Dim. 2	0.460	0.652	0.789	-0.276	-0.922	-0.525	0.534

Exchangeability

- A finite set of random variables $\{x_1, \ldots, x_N\}$ is said to be exchangeable if the joint distribution is invariant to permutation. If π is a permutation of the integers from 1 to N:

$$p(x_1, \ldots x_N) = p(x_{\pi(1)}, \ldots, x_{\pi(N)})$$

- An infinite sequence of random is infinitely exchangeable if every finite subsequence is exchangeable

bag-of-words assumption

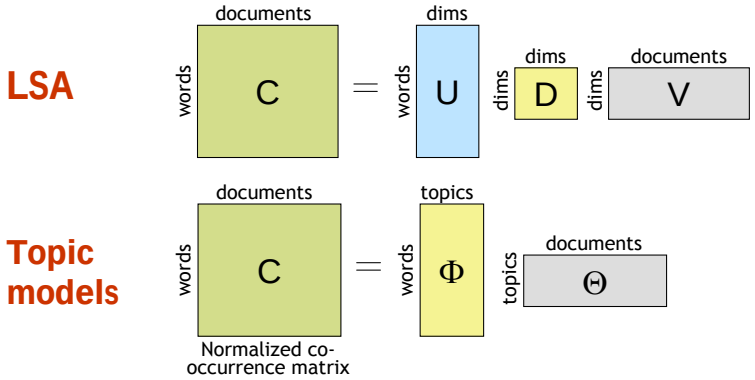
- Word order is ignored
- "bag-of-words" – exchangeability, not i.i.d
- Theorem (De Finetti, 1935): if $(x_1, x_2, \ldots, x_N)$ are infinitely exchangeable, then the joint probability has a representation as a mixture:

$$p(x_1, x_2, \ldots, x_N)$$

for some random variable θ

$$p(x_1, x_2, \ldots, x_N) = \int d\theta\, p(\theta) \prod_{i=1}^{N} p(x_i|\theta)$$

Probabilistic topic models: key ideas



- **Key idea**: *documents are mixtures of latent topics, where a topic is a probability distribution over words.*
- Hidden variables, generative processes, and statistical inference are the foundation of probabilistic modeling of topics.

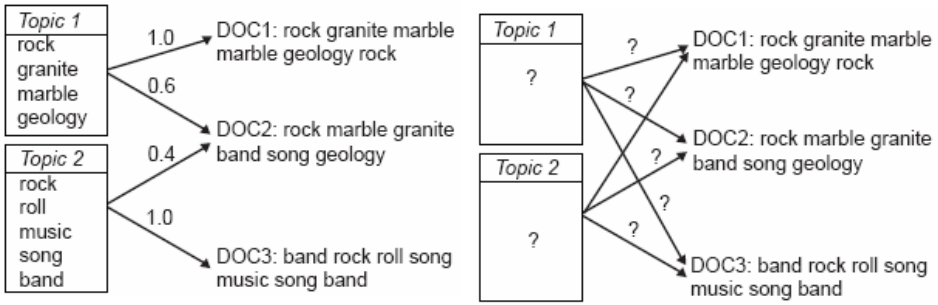
Probabilistic topic models: processes

Generative models: generating a document

- Choose a distribution over topics and the document length;
- For each word w_i, choose a topic at random according to this distribution, and choose a word from the topic-word distribution.

Statistical inference (invert): to know which topic model is most likely to have generated the data, it infers

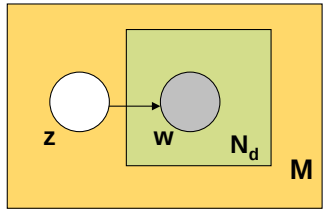
- Probability distribution over words associated with each topic
- Distribution over topics for each document
- Topic responsible for generating each word



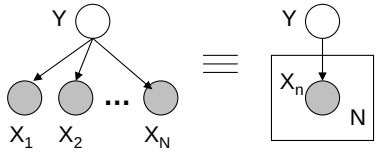
Mixture of unigrams model

(Nigam et al., 2000)

- Simple, each document one topic (appropriate for supervised classification).
- Generates a document by
  - choosing a topic z
  - generating N words independently from the conditional multinomial distribution p(w|z)
- A topic is associated with a specific language model that generates words appropriate to the topic.



$$p(d) = \sum_z p(z) \prod_{n=1}^{N_d} p(w_n \mid z)$$



- Nodes are random variables
- Edges denote possible dependence
- Plates denotes replicated structure
- Pattern of conditional dependence between the ensemble of random variables

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

Observable variables Latent

How to calculate?

$$p(d) = \sum_z p(z) \prod_{n=1}^{N_d} p(w_n \mid z)$$

- We must draw the multinomial distributions p(z) and p(w|z)
- If each document is annotated with a topic z
  - using maximum likelihood estimation → p(z)
  - count # times each word w appeared in all documents labeled with z and then normalize → p(w|z)
- If topics are not known for documents
  - EM algorithm can be used to estimate p(d)
- Once the model has been trained, inference can be performed using Bayes' rule to obtain the most likely topics for each document.

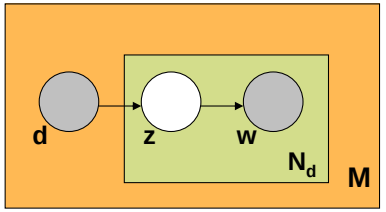
Limitations:

1. a document can only contain a single topic.
2. the distributions have no priors and are assumed to be learned completely from data

Probabilistic latent semantic indexing

(Hofmann, 1999)

- pLSI: Each word is generated from a single topic, different words in the document may be generated from different topics.
- Each document is represented as a list of mixing proportions for the mixture components.



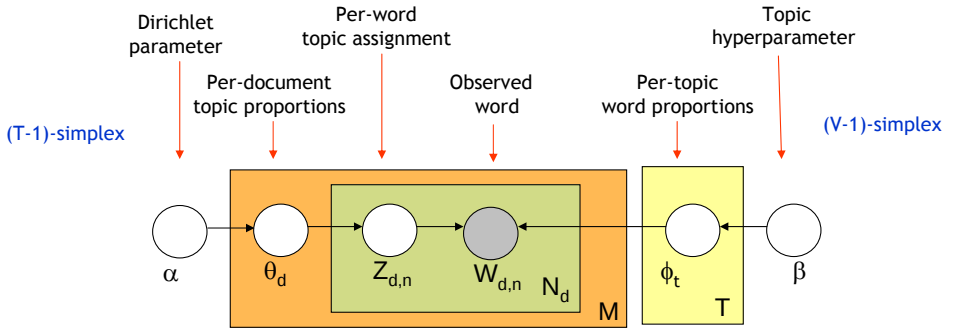
$$p(d, w_n) = p(d) \sum_z p(w_n \mid z) p(z \mid d)$$

- Generative process:
  - Choose a document d_m with p(d)
  - For each word w_n in the d_m
    - Choose a z_n from a multinomial conditioned on d_m, i.e., from $p(z|d_m)$
    - Choose a w_n from a multinomial conditioned on z_n, i.e., from $p(w|z_n)$.

Limitations

- The model allows multiple topics in each document, but
  - the possible topic proportions have to be learned from the document collection
  - pLSI does not make any assumptions about how the mixture weights θ are generated, making it difficult to test the generalizability of the model to new documents.
- Topic distribution must be learned for each document in the collection → # parameters grows with the number of documents (billion documents?).
- Blei et al. (2003) extended this model by introducing a Dirichlet prior on θ, calling Latent Dirichlet Allocation (LDA).

Latent Dirichlet allocation

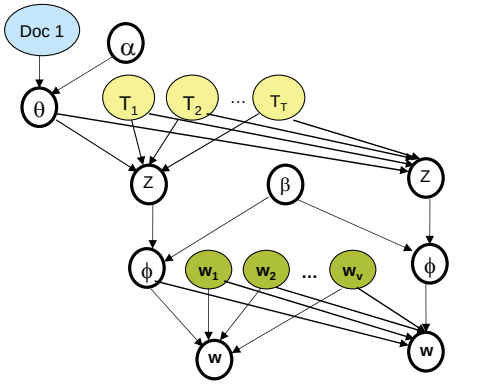


1. Draw each topic $\phi_t \sim Dir(\beta)$, t=1,..,T
2. For each document:
   1. Draw topic proportions $\theta_d \sim Dir(\alpha)$
   2. For each word:
      1. Draw $z_{d,n} \sim Mult(\theta_d)$
      2. Draw $w_{d,n} \sim Mult(\phi_{z_{d,n}})$

1. From collection of documents, infer
   - per-word topic assignment $z_{d,n}$
   - per-document topic proportions θ_d
   - per-topic word distribution ϕ_t
2. Use posterior expectations to perform the tasks: IR, similarity, ...

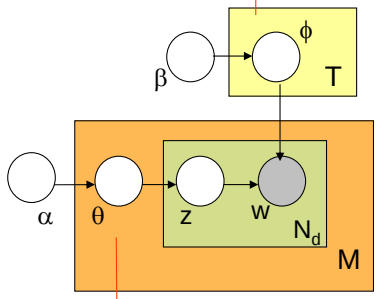
Choose N_d from a Poisson distribution with parameter ξ

Latent Dirichlet allocation

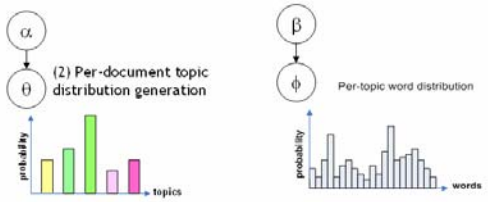


kxV matrix β, $\beta_{ij} = p(w^j=1,z^i=1)$

word distributions ϕ over each topic



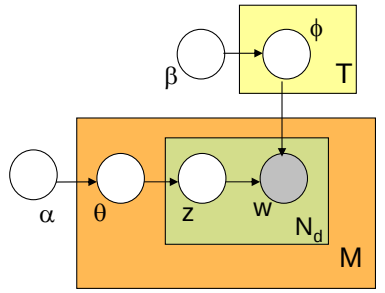
topic distributions θ over each document



LDA model

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

Dirichlet prior on the per-document topic distributions



$$p(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta) = p(\theta|\alpha)\prod_{n=1}^{N}p(z_n|\theta)p(w_n|z_n,\beta)$$

Joint distribution of topic mixture θ, a set of N topic **z**, a set of N words **w**

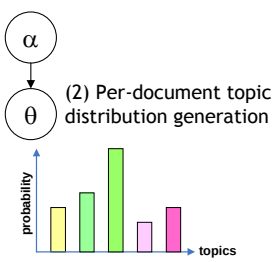
$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n}p(z_n|\theta)p(w_n|z_n,\beta)\right)d^k\theta$$

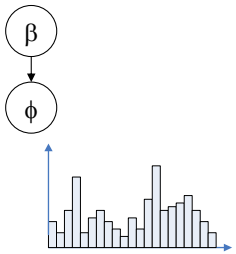
Marginal distribution of a document by integrating over θ and summing over z

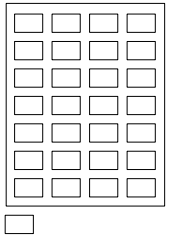
$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d^k\theta_d$$

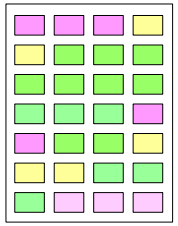
Probability of collection by product of marginal probabilities of single documents

Generative process

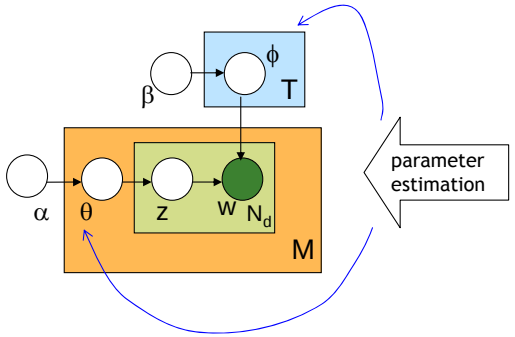








Inference: parameter estimation



Parameter estimation methods:

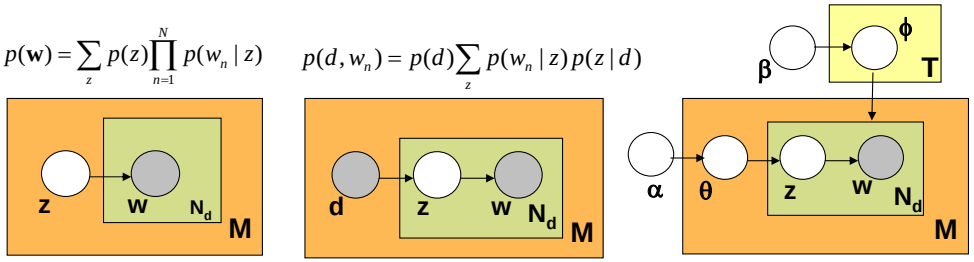
- Mean field variational methods (Blei et al., 2001, 2003)
- Expectation propagation (Minka & Lafferty 2002)
- Gibbs sampling (Griffiths & Steyvers 2004)
- Collapsed variational inference (Teh et al., 2006)

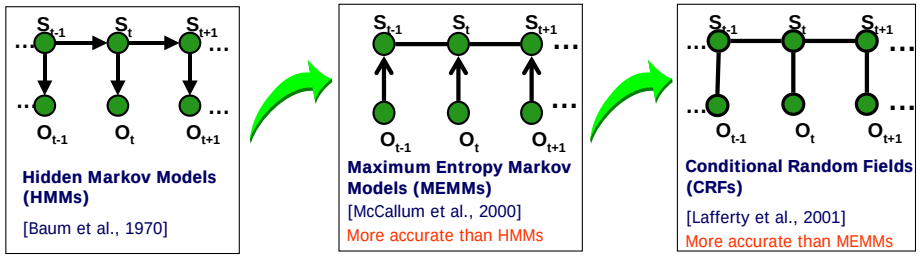
Evolution

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}$$

$$p(\mathbf{w}) = \sum_z p(z)\prod_{n=1}^{N} p(w_n \mid z)$$

$$p(d, w_n) = p(d)\sum_z p(w_n \mid z)\,p(z \mid d)$$





Hidden Markov Models (HMMs)

[Baum et al., 1970]

Maximum Entropy Markov Models (MEMMs)

[McCallum et al., 2000]

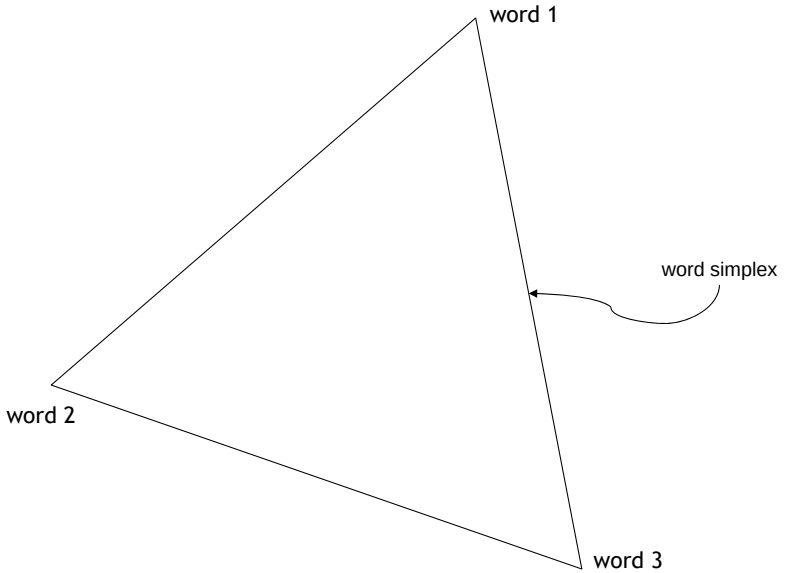
More accurate than HMMs

Conditional Random Fields (CRFs)

[Lafferty et al., 2001]

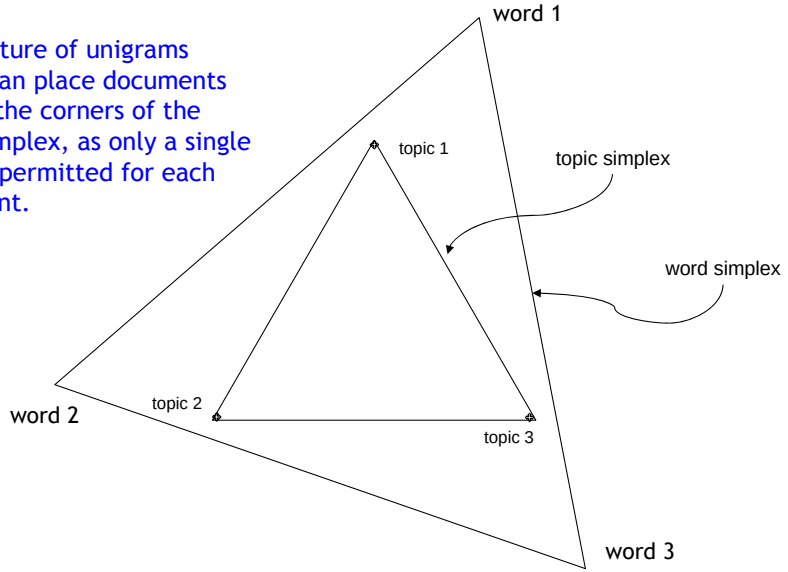
More accurate than MEMMs

A geometric interpretation



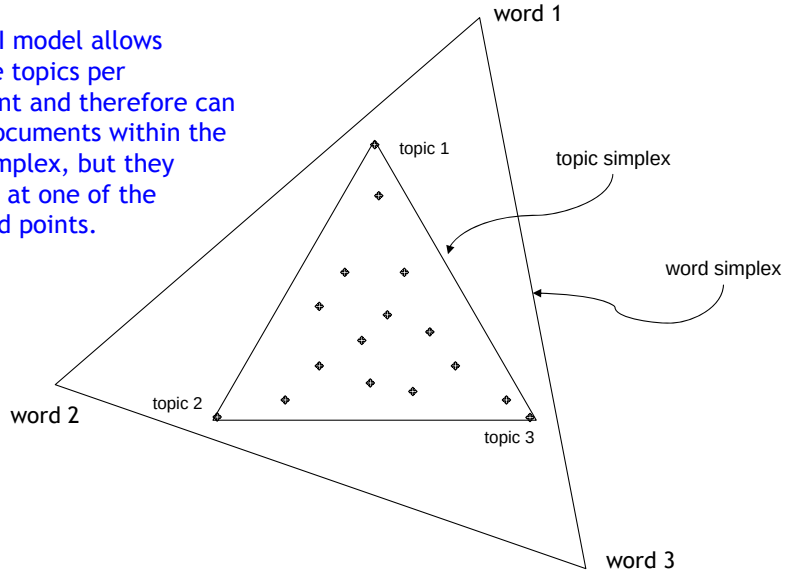
A geometric interpretation

The mixture of unigrams model can place documents only at the corners of the topic simplex, as only a single topic is permitted for each document.



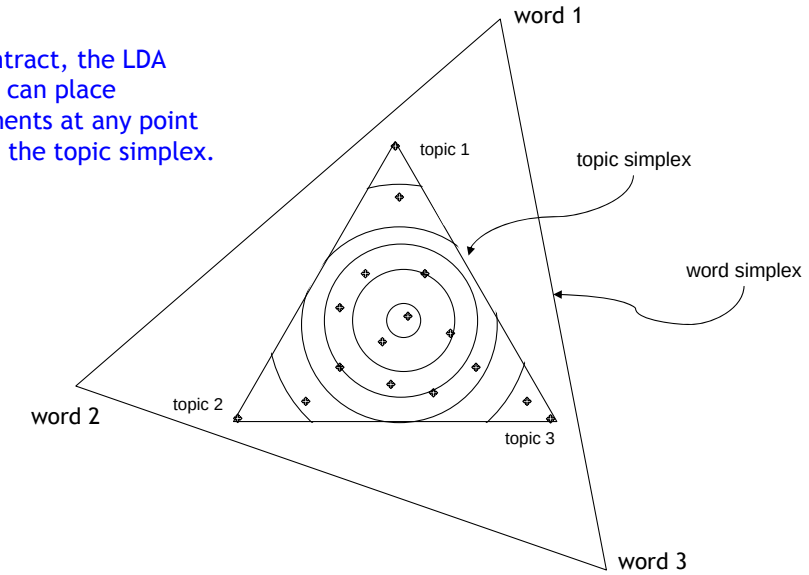
A geometric interpretation

The pLSI model allows multiple topics per document and therefore can place documents within the topic simplex, but they must be at one of the specified points.



A geometric interpretation

By contract, the LDA model can place documents at any point within the topic simplex.



Inference

- We want to use LDA to compute the posterior distribution of the hidden variables given a document:

$$p(\theta, \mathbf{z} \mid \mathbf{w}, \alpha, \beta) = \frac{p(\theta, z, w \mid \alpha, \beta)}{p(w \mid \alpha, \beta)}$$

- Unfortunately, this is intractable to compute in general. We marginalize over hidden variables and write (3) as:

$$p(\mathbf{w} \mid \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^{N} \sum_{i=1}^{k} \prod_{j=1}^{V} (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

- Variety of approximate inference algorithms for LDA

Inference

- Expectation Maximization
  - But poor results (local maxima)
- Gibbs Sampling
  - Parameters: ϕ, θ
  - Start with initial random assignment
  - Update parameter using other parameters
  - Converges after 'n' iterations
  - Burn-in time

Example

- From 16000 documents of AP corpus → 100-topic LDA model.
- An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated

"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.

TASA corpus, 37000 texts, 300 topics

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

By giving equal probability to the first two topics, one could construct a document about a person that has taken too many drugs, and how that affected color perception.

Topic models for text data

http://www.ics.uci.edu/~smyth/topics.html

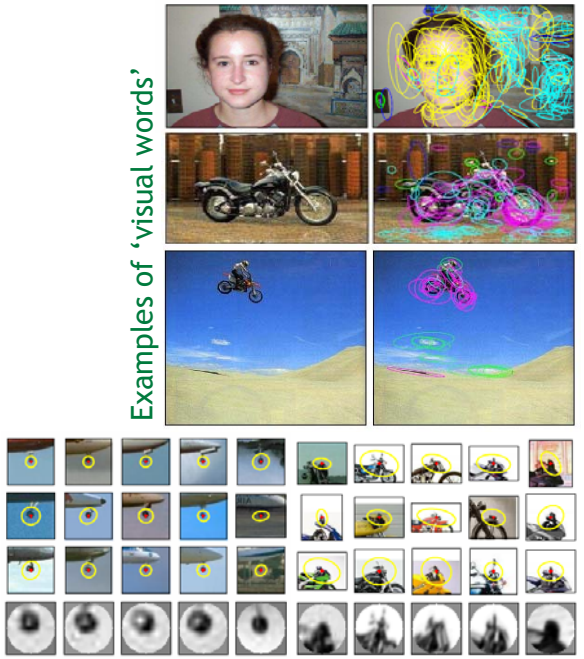
- David Blei's Latent Dirichlet Allocation (LDA) code in C, using variational learning.
- A topic-based browser of UCI and UCSD faculty research interests built by Newman and Asuncion at UCI in 2005.
- A topic-based browser for 330,000 New York Times articles, by Dave Newman, UCI.
- Wray Buntine's topic-based search interface to Wikipedia.
- Dave Blei and John Lafferty's browsable 100-topic model of journal articles from Science.
- LSA tools and application http://LSA.colorado.edu

Directions for hidden topic discovery

- Hidden Topic Discovery from Documents
- Application in Web Search Analysis & Disambiguation
- Application in Medical Information (Disease Classification)
- Application in Digital Library (Info. Navigation)
- Potential Applications in Intelligent Advertising & Recommendation
- Many others

Visual words

- Idea: Given a collection of images,
  - Think of each image as a document.
  - Think of feature patches of each image as words.
  - Apply the LDA model to extract topics.
- J. Sivic et al., Discovering object categories in image collections. *MIT AI Lab Memo AIM-2005-005*, Feb. 2005



Related works and problems

- Hyperlink modeling using LDA, Erosheva …, PNAS, 2004
- Finding scientific topics, Griffiths & Steyvers, PNAS, 2006
- Author-Topic model for scientific literature, Rozen-Zvi …, UAI, 2004
- Author-Topic-Recipient model for email data, McCallum …, IJCAI'05
- Modeling Citation Influences, Dietz et al., ICML 2007
- Word sense disambiguation, Blei …, 2007
- Classify short and sparse text & Web …, P.X. Hieu …, www2008
- Automatic Labeling of Multinomial Topic Models, Mei …, KDD 2007???
- DLA-based doc. models of ad-hoc retrieval, Croft …, SIGIR 2006???
- Connection to language modeling???
- Topic models and emerging trend detection???
- Similarity between words, between documents → clustering???
- etc.

Topic analysis of Wikipedia

- **Topic-oriented crawling to download documents from Wikipedia**
  - *Arts*: *architecture, fine art, dancing, fashion, film, museum, music, …*
  - *Business*: *advertising, e-commerce, capital, finance, investment, …*
  - *Computers*: *hardware, software, database, digital, multimedia, …*
  - *Education*: *course, graduate, school, professor, university, …*
  - *Engineering*: *automobile, telecommunication, civil engineering, …*
  - *Entertainment*: *book, music, movie, movie star, painting, photos, …*
  - *Health*: *diet, disease, therapy, healthcare, treatment, nutrition, …*
  - *Mass-media*: *news, newspaper, journal, television, …*
  - *Politics*: *government, legislation, party, regime, military, war, …*
  - *Science*: *biology, physics, chemistry, ecology, laboratory, patent, …*
  - *Sports*: *baseball, cricket, football, golf, tennis, olympic games, …*

Raw data: **3.5**GB, **471,177** docs
Preprocessing: remove duplicates, HTML, stop & rare words
Final data: **240**MB, **71,986** docs, **882,376** paragraphs, **60,649** unique words, **30,492,305** words

JWikiDocs: Java Wikipedia Document Crawling Tool http://jwebpro.sourceforge.net/
GibbsLDA++: C/C++ Latent Dirichlet Allocation http://gibbslda.sourceforge.net/

(source: next some slides from P.X. Hieu, www08)

Topic discovery from MEDLINE

PubMed www.pubmed.gov
A service of the National Library of Medicine and the National Institutes of Health

MEDLINE (Medical Literature Analysis and Retrieval System Online): database of life sciences and biomedical information. It covers the fields of medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care.

More than **15 million records** from approximately **5,000 selected publications** (NLM Systems, Feb 2007) covering biomedicine and health from **1950 to the present**.

Our topic analysis on **400MB** MEDLINE data including **348,566** medical document abstracts from 1987 to 1991. The outputs (i.e., hidden topics) will be used for "disease classification"

Disease name	ID
Neoplasms	C04
Digestive System Diseases	C06
Cardiovascular Diseases	C14
Immunologic Diseases	C20
Pathological Conditions, Signs and Symptoms	C23

The full list of topics at http://gibbslda.sourceforge.net/ohsumed-topics.txt

Topics most related to *Neoplasms*
T182: cancer breast cancers women er screening pr mammography carcinoma mammary mastectomy colorectal axilla tamoxifen estrogen tumor incidence detection status benign mammographic malignancy invasive tumors observed …
T149: chemotherapy median toxicity treatment therapy survival treated combination mg/m cisplatin regimen study r methotrexate cancer doxorubicin dose cyclophosphamide cr high-dose phase partial plus rate trial …
T193: carcinoma cell malignant melanoma carcinomas squamous tumour tumours gland tumor adenocarcinoma tumo salivary metastatic lesions benign glands neoplasms invasive melanomas parotid papillary basal malignancy adenomas …

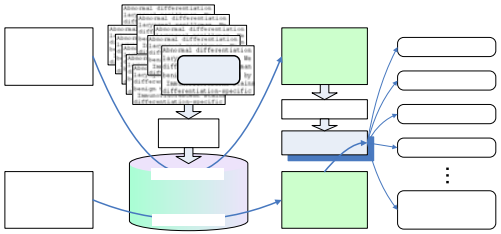
Topics most related to *Digestive System Diseases*
T60: liver hepatic portal cirrhosis hepatocytes hepatocellular shunt varices cirrhotic chronic livers alcoholic bleeding sclerotherapy function hepatitis hepatocyte variceal hcc encephalopathy intrahepatic ascites shunts hepatectomy hypo
T102: gastric ulcer duodenal gastrointestinal mucosal ulcers mucosa stomach cimetidine acid pylori bleeding endosco endoscopy gastritis gastrin ranitidine peptic sucralfate healing cp ulceration ph damage vagotomy …
T74: bile pancreatic biliary duct pancreatitis gallbladder pancreas endoscopic cholecystectomy bilirubin obstruction cholangitis jaundice ducts amylase stones gallstone sclerosing cholecystitis es drainage acute exocrine retrograde …

Topics most related to *Cardiovascular Diseases*
T164: myocardial infarction coronary acute angina ischemia artery cardiac ischemic ami infarct depression unstable c segment ecg heart mi events silent pectoris chest hours electrocardiographic thrombolytic …
T180: pressure blood hypertension hg hypertensive systolic diastolic pressures bp arterial normotensive antihypertens mean spontaneously mmhg wky elevated heart cardiovascular reduction hypotension rate mild supine …
T114: coronary artery angioplasty stenosis balloon arteries bypass angiography percutaneous occlusion ptca translum diameter vessel angiographic restenosis stenoses arterial anterior descending vessels lesions segments plaque surgery …

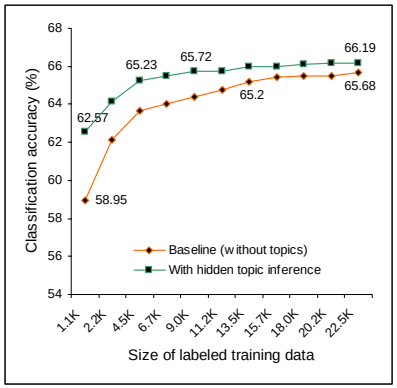
Topics most related to *Immunologic Diseases*
T137: hiv aids immunodeficiency infection virus human acquired syndrome infected risk seropositive transmission htlv-i immune drug sexual zidovudine hiv-infected homosexual kaposi sarcoma antibody transmitted infections asymptomatic …
T140: asthma histamine airway mast bronchial fev asthmatic inhaled responsiveness airways subjects methacholine inhalation pd bronchoconstriction function aerosol respiratory obstruction cough expiratory reactivity forced baseline ltd
T20: antibodies igg antibody sera serum iga lupus immune igm systemic assay erythematosus elisa sle antigen autoantibo immunoglobulin autoimmune antigens detected enzyme-linked immunosorbent complexes titers positive …

Topics most related to *Pathological Conditions, Signs and Symptoms*
T54: symptoms clinical diagnosis signs pain history symptomatic asymptomatic examination findings physical presenta fever recurrent laboratory symptom cause diagnosed features illness complaints severe manifestations evaluation diagno
T162: surgery cent postoperative surgical operation preoperative complications procedures operative operations intraop postoperatively resection procedure period follow-up perioperative preoperatively elective incidence operated rate …
T70: complications abdominal surgical fistula drainage management abscess complication perforation surgery fistulas t splenic treatment repair laparotomy abscesses bleeding managed complicated hemorrhage splenectomy hernia operative …

Application in disease classification



Comparison of disease classification accuracy using maximum entropy method



with the increment of training data

Training disease abstracts (moderate size)

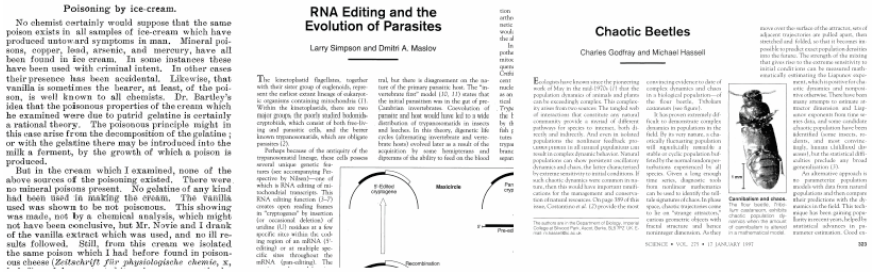
Ohsumed MEDLINE

T
d
abst
hidden topics

Application in digital library

Modeling *Science*

David M. Blei and John D. Lafferty
Princeton Unversity and Carnegie Mellon University
[Blei & Lafferty 2007]



- Our data are *Science* from 1880-2002, courtesy of JSTOR.
- JSTOR is an on-line archive that scans the original volumes and performs optical character recognition on the scans.
- This process results in 130K documents, 76M words.

Source: http://www.cs.princeton.edu/~blei/modeling-science.pdf

Digestive Syst. Dis.

WORDS
ice
climate
ocean
sea
temperature
years
surface
water
changes
global
change
data
atmospheric
time
warming
north
record
fig
pacific

RELATED TOPICS
volcanic years fig deposits rocks
age rates rate coral mortality
ozone atmospheric air measurements atmosphere
carbon co2 energy water gas
species forest forests populations population
ice climate ocean sea temperature

RELATED DOCUMENTS
"Nonstationary Phase of the Plio-Pleistocene Asian Monsoon" (1996)
"Northern Hemisphere Ice-Sheet Influences on Global Climate Change" (1999)
"Abrupt Climate Events 500,000 to 340,000 Years Ago: Evidence from Subpolar North Atlantic Sediments" (1998)
"Climate Records Covering the Last Deglaciation" (1995)
"A 0.5-Million-Year Record of Millennial-Scale Climate Variability in the North Atlantic" (1999)
"Simulations of Atmospheric Variability Induced by Sea Surface Temperatures and Implications for Global Warming" (1994)
"Interdecadal Climate Fluctuations That Depend on Exchanges Between the Tropics and Extratropics" (1997)
"Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation" (1995)
"El Nino-Southern Oscillation Displacements of the Western Equatorial Pacific Warm Pool" (1990)
"Mechanism of the Zonal Displacements of the Pacific Warm Pool: Implications for ENSO" (1996)
"16°C Rapid Temperature Variation in Central Greenland 70,000 Years Ago" (1999)
"Past and Present Subtropical Summer Monsoons" (1996)
"Climate Change During the Last Deglaciation in Antarctica" (1996)

Demo of Information Navigation in Journal *Science*

http://www.cs.cmu.edu/~lemur/science/

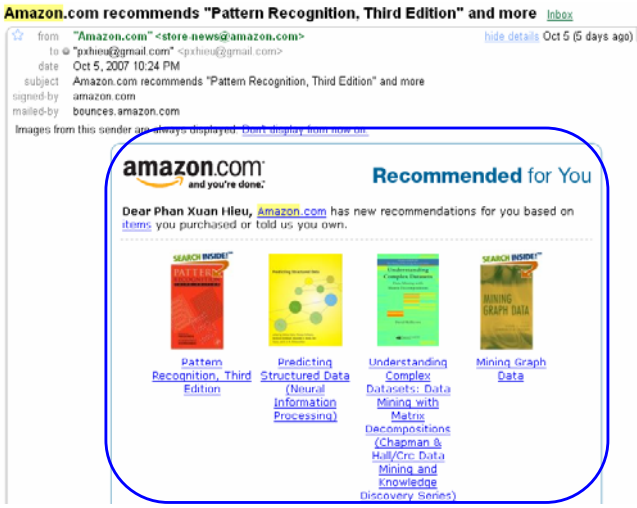
This was done with Correlated Topic Model (CTM) – a more advanced variant of LDA

T

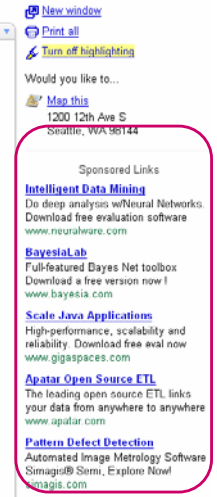
C

Nev
abst
hidc

Potential in intelligent advertising



Amazon recommendation

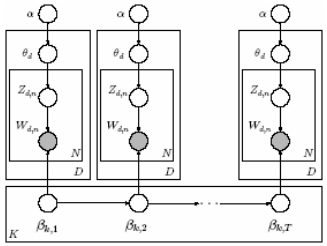


Google context sensitive advertising

Can Amazon and Gmail understand us?

Applications in scientific trends

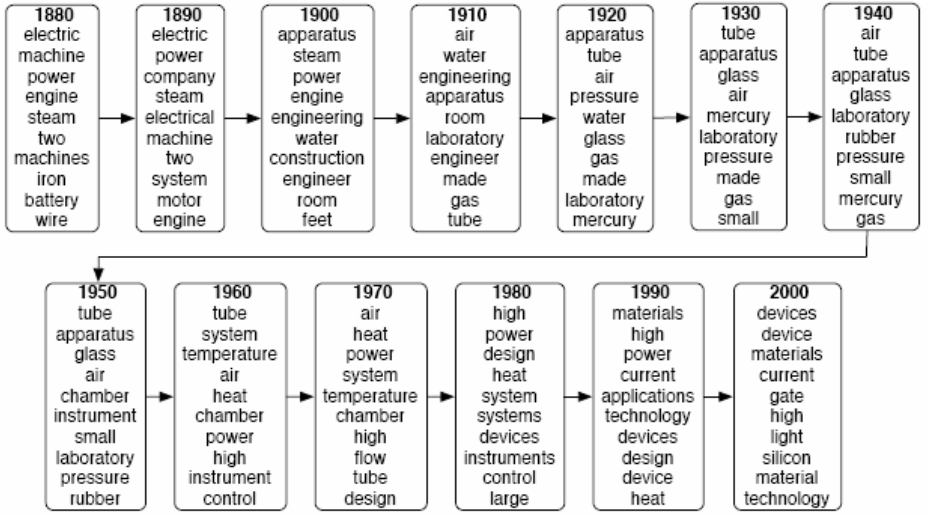
Dynamic Topic Models [Blei & Lafferty 2006]



Analyzed Data:

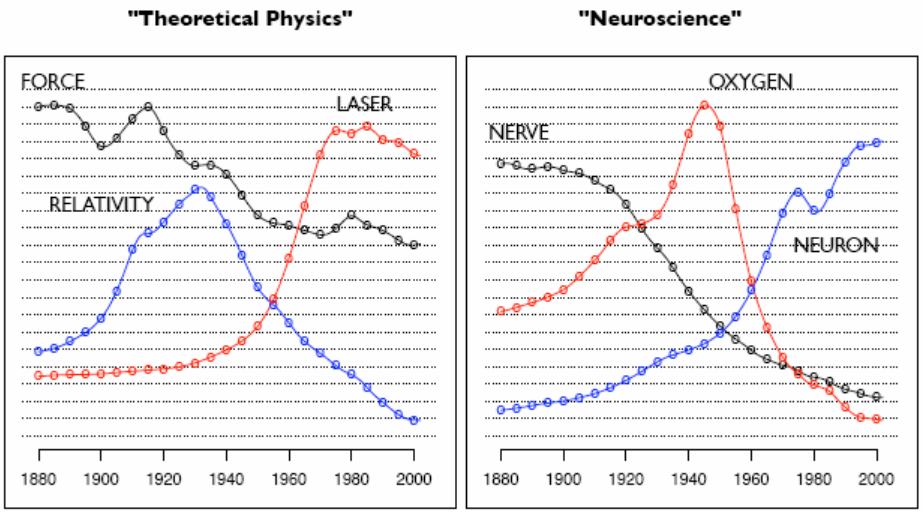
- JSTOR (www.jstor.org) scanned and ran optical character recognition on *Science* from 1880-2002.
- No reliable punctuation, meta-data, or references
- Restrict to 30K terms that occur more than ten times
- The data are 76M words in 130K documents

Analyzing a topic



Source: http://www.cs.princeton.edu/~blei/modeling-science.pdf

Visualizing trends within a topic



Summary

- LSA and topic models are roads to text meaning.
- Can be viewed as a dimensionality reduction technique.
- Exact inference is intractable, we can approximate instead.
- Various applications and fundamentals for digitalized era.
- Exploiting latent information depends on applications, the fields, researcher backgrounds, …

Key references

- S Deerwester, et al. (1990). Indexing by latent semantic analysis. Journal American Society for Information Science (citation 3574).
- Deerwester, T. (1999). Probabilistic Latent Semantic Analysis. Uncertainty in AI (citation 722).
- Nigam et al. (2000). Text classification from labeled and unlabeled documents using EM, Machine learning (citation 454).
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. J. of Machine Learning Research (citation 699).

Some other references

- Sergey Brin, Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, seventh international conference on World Wide Web 7, p.107-117, April 1998, Brisbane, Australia.
- Taher H. Haveliwala, Topic-sensitive PageRank, 11th international conference on World Wide Web, May 07-11, 2002, Honolulu, Hawaii, USA.
- M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in PageRank. NIPS 14. MIT Press, 2002.
- Lan Nie , Brian D. Davison , Xiaoguang Qi, Topical link analysis for web search, 29th ACM SIGIR conference on Research and development in information retrieval, August 06-11, 2006, Seattle, Washington, USA.

Mixture models

- **mixture model** is a model in which independent variables are fractions of a total. Discrete random variable *X* is a mixture of *n* component discrete random variables *Yi*.
- a **probability mixture model** is a probability distribution that is a convex combination of other probability distributions

$$f_X(x) = \sum_{i=1}^{n} a_i f_{Y_i}(x)$$

- In a **parametric mixture model**, the component distributions are from a parametric family, with unknown parameters θi

$$f_X(x) = \sum_{i=1}^{n} a_i f_Y(x;\theta_i)$$

- and **continuous mixture**

$$f_X(x) = \int_\Theta h(\theta) f_Y(x;\theta) d\theta, \quad h(\theta) \geq 0, \forall \theta \in \Theta \quad \text{and} \quad \int_\Theta h(\theta) d\theta = 1$$

Estimation in mixture models

- Known distribution *Y* and sample from *X*, would like to determine the α_i and θ_i values.
- Expectation-maximization algorithm is an iterative
- The expectation step: with guessed parameters, for each data point x_j and distribution Y_i

$$y_{i,j} = \frac{a_i f_Y(x_j;\theta_i)}{f_X(x_j)}$$

- The maximization step:

$$a_i = \frac{1}{N}\sum_{i=1}^{N} y_{i,j} \quad \text{and} \quad \mu_i = \frac{\sum_j y_{i,j} x_j}{\sum_j y_{i,j}}$$

Distributions

- **Binomial distribution:** discrete probability distribution of # successes in *n* Bernoulli trials (success/failure outcomes)

$$f(k;n,p) = P(X = k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

  - **n=10, #red=2, #blue=3, p = 0.4, P(red = 4) = …**

- **Multinomial distribution**: discrete probability distribution of # occurrences of each outcome (x_i) among *k* outcomes in *n* independent trials (outcomes probabilities p_1, …, p_k)

$$f(x_1,\ldots,x_k;n,p_1,\ldots,p_k) = \begin{cases} \frac{n!}{x_1!\ldots x_k!}p_1^{x_1}\ldots p_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \\ 0 & \text{otherwise} \end{cases}$$

$$= P(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k)$$

  - **n=10, #red=2, #blue=2, #black=1, p = 0.4, P(red=5, blue=2, black=3)=…**

Distributions

- **Beta distribution**: Used to model random variables that vary between two finite limits, characterized by two parameters $\alpha>0$ and $\beta>0$. Beta distribution is quite useful for modeling proportions.

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \qquad \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

- Examples:
  - The percentage of impurities in a certain manufactured product
  - The proportion of flat (by weight) in a piece of meat
- **Dirichlet distribution Dir(α)**: the multivariate generalization of the beta distribution, and conjugate prior of the multinomial distribution in Bayesian statistics.

$$f(x_1,\ldots,x_{k-1};\alpha_1,\ldots,\alpha_k) = p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}x_1^{\alpha_1-1}\cdots x_k^{\alpha_k-1}$$

Distributions

Discrete variables

- **Binomial distribution**: Discretely distributing the unit to two outcomes in n experiments.
- **Multinomial distribution**: Discretely distributing the unit to k outcomes in n experiments

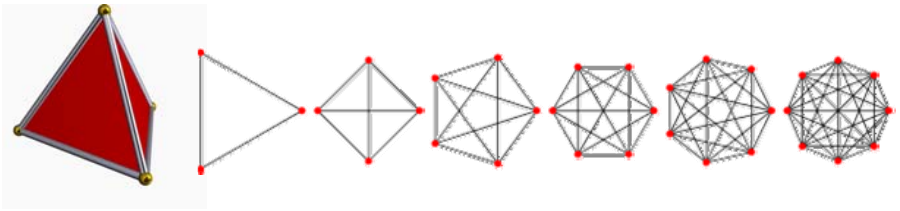
Continuous variables

- **Beta distribution:** Continuously distributing the unit between two limits (2-simplex).
- **Dirichlet distribution**: Continuously distributing the unit between k limits (k-simplex).

Poisson distribution $f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$

Simplex

- A simplex, sometimes called a hyper tetrahedron is the generalization of a tetrahedral region of space to *n* dimensions. The boundary of a k-simplex has k+1 0-faces (polytope vertices), k(k+1)/2 1-faces (polytope edges), and $\binom{k+1}{i+1}$ i-faces,



graphs for the n-simplexes with n=2 to 7

- Nd = 4, (w1 = ?, w2 = ?, w3 = ?, w4 = ?) word token
- P(w1=band, w2=music, w3=song, w4=rock|topic=music)

LDA and exchangeability

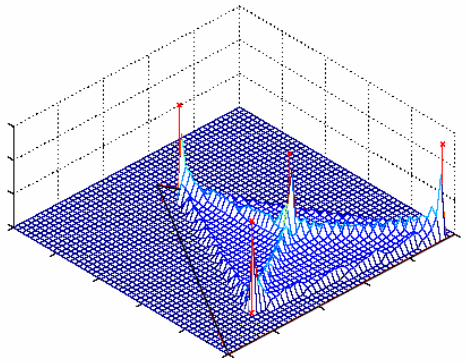
- We assume that words are generated by topics and that those topics are infinitely exchangeable within a document.
- By de Finetti's theorem:

$$p(\mathbf{w},\mathbf{z}) = \int p(\theta)\left(\prod_{n=1}^{N} p(z_n|\theta)\,p(w_n|z_n)\right)d\theta$$

- By marginalizing out the topic variables, we get eq. 3 in the previous slide.

- Density on unigram distributions $p(w|\theta;\beta)$ under LDA for three words and four topics.
- The triangle is the 2-D simplex representing all possible multinomial distributions over three words.
  - Each vertex corresponds to a distribution that assigns probability one to one of the words;
  - the midpoint of an edge gives probability 0.5 to two of the words; and the centroid of the triangle is the uniform distribution over all three words.
- The four points marked with an x are the locations of the multinomial distributions $p(w|z)$ for each of the four topics, and the surface shown on top of the simplex is an example of a density over the (V - 1)-simplex (multinomial distributions of words) given by LDA.

$$p(w|\theta,\beta) = \sum_{z} p(w|z,\beta)\,p(z|\theta)$$



Notations

- P(z): distribution over topics z in a particular document
- P(w|z): prob. distribution over words w given topic z
- $P(z_i = j)$: probability that the jth topic was sampled for the ith word token
- $P(w_i|z_i = j)$ as the probability of word w_i under topic j.
- Distribution over words within a document

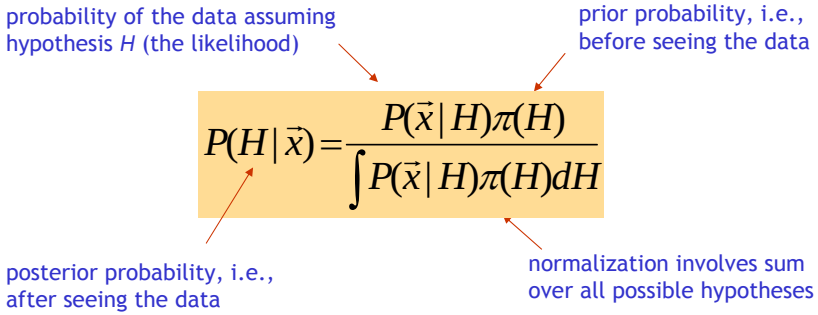
$$p(w_i) = \sum_{j=1}^{T} p(w_i \mid z_i = j)\,p(z_i - j)$$

- $\phi^{(j)}$= P(w|z=j): multinomial dist. over words for topic j
- $\theta^{(d)}$= P(z): multinomial dist. over topics for document d

ϕ and θ: which words are important for which topic and which topics are important for a particular document

Bayesian statistics: general philosophy

- Interpretation of probability extended to degree of belief (subjective probability). Use this for hypotheses:



- Bayesian methods can provide more natural treatment of non-repeatable phenomena: probability that Kitajima wins gold medal in Olympic 2008, ...