Topic Models

Pawan Goyal

CSE, IITKGP

September 7-9, 2016

Why Topic Modeling?

Information Overload

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

Why Topic Modeling?

Information Overload

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

Main Tools: Search and Links

- We type keywords into a search engine and find a set of related documents
- We look at these documents and possibly navigate to other documents

Why Topic Modeling?

Search Based-on themes

- Imagine searching and exploring documents based on themes that run through them
- We might “zoom-in” or “zoom-out” to find specific or broader themes
- We might look at how themes change through time, how they are connected to each other
- Find the theme first and then examine the documents pertaining to that theme

Why Topic Modeling?

Topic Modeling

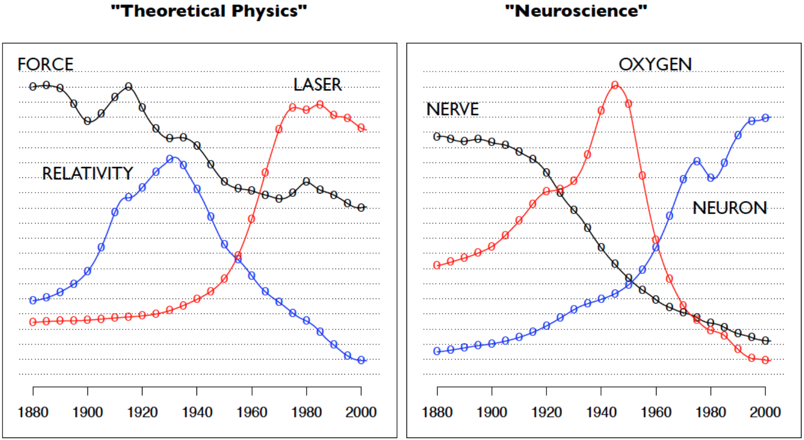
Provides methods for automatically organizing, understanding, searching and summarizing large electronic archives without any prior annotation or labeling

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

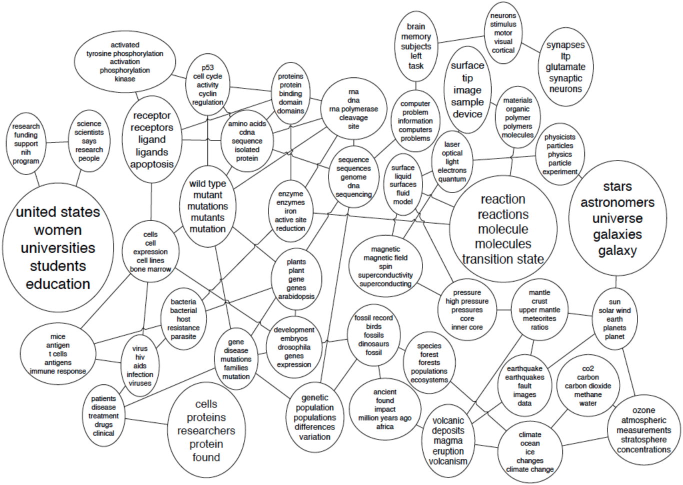
Applications: Discover Topics from a corpus

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

Applications: Model the evolution of topics over time



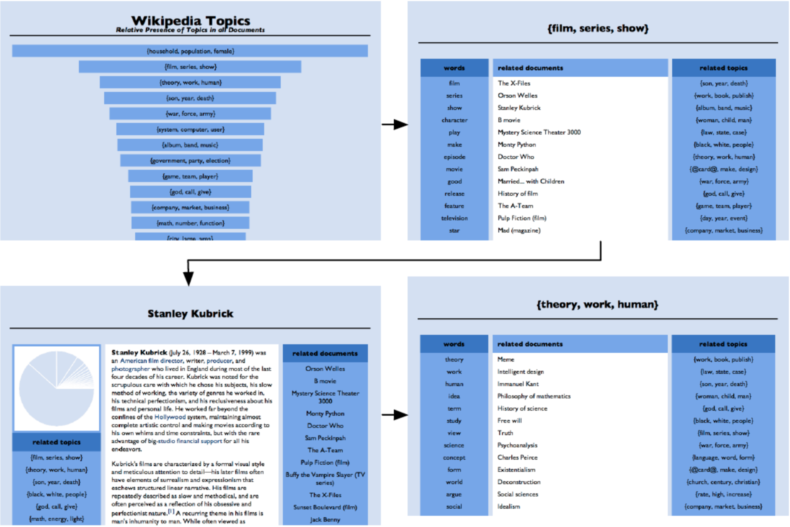
Applications: Model connections between topics



Link Prediction using Relational Topic Models

Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo	RTM (ψ_e)
Rates of convergence of the Hastings and Metropolis algorithms	
Possible biases induced by MCMC convergence diagnostics	
Bounding convergence time of the Gibbs sampler in Bayesian image restoration	
Self regenerative Markov chain Monte Carlo	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Rate of Convergence of the Gibbs Sampler by Gaussian Approximation	
Diagnosing convergence of Markov chain Monte Carlo algorithms	
Exact Bound for the Convergence of Metropolis Chains	LDA + Regression
Self regenerative Markov chain Monte Carlo	
Minorization conditions and convergence rates for Markov chain Monte Carlo	
Gibbs-markov models	
Auxiliary variable methods for Markov chain Monte Carlo with applications	
Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models	
Mediating instrumental variables	
A qualitative framework for probabilistic inference	
Adaptation for Self Regenerative MCMC	

Applications: Organize and browse large corpora



{film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}
feature	The A-Team	{game, team, player}
television	Pulp Fiction (film)	{day, year, event}
star	Mad (magazine)	{company, market, business}

Stanley Kubrick

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick's films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature.[1] A recurring theme in his films is man's inhumanity to man. While often viewed as

related topics
{film, series, show}
{theory, work, human}
{son, year, death}
{black, white, people}
{god, call, give}
{math, energy, light}

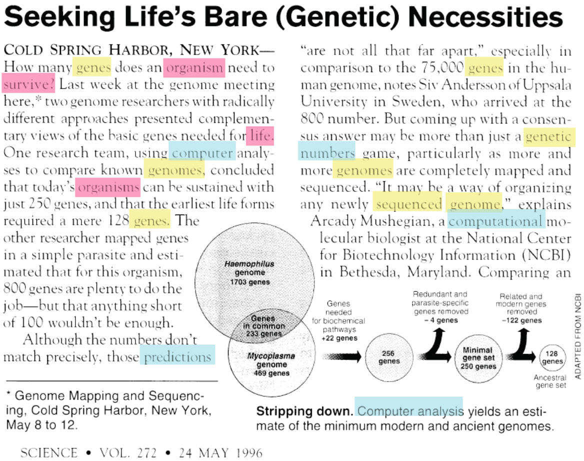
related documents
Orson Welles
B movie
Mystery Science Theater 3000
Monty Python
Doctor Who
Sam Peckinpah
The A-Team
Pulp Fiction (film)
Buffy the Vampire Slayer (TV series)
The X-Files
Sunset Boulevard (film)
Jack Benny

{theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}
world	Deconstruction	{church, century, christian}
argue	Social sciences	{rate, high, increase}
social	Idealism	{company, market, business}

https://www.princeton.edu/ achaney/tmve/wiki100k/browse/topic-presence.html

Intuition: Documents exhibit multiple topics



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.

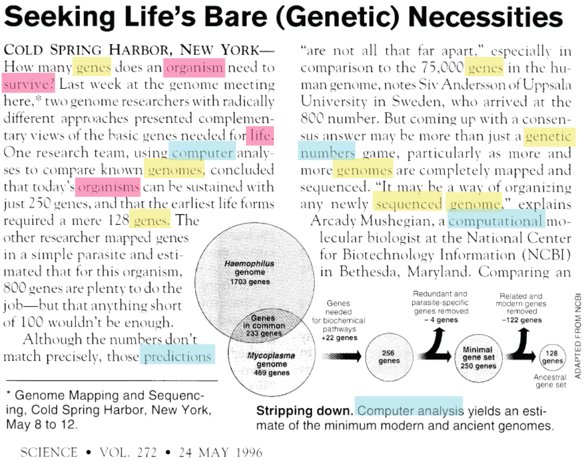
---

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

SCIENCE • VOL. 272 • 24 MAY 1996

This article is about using data analysis to determine the number of genes an organism needs to survive

Intuition: Documents exhibit multiple topics



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.

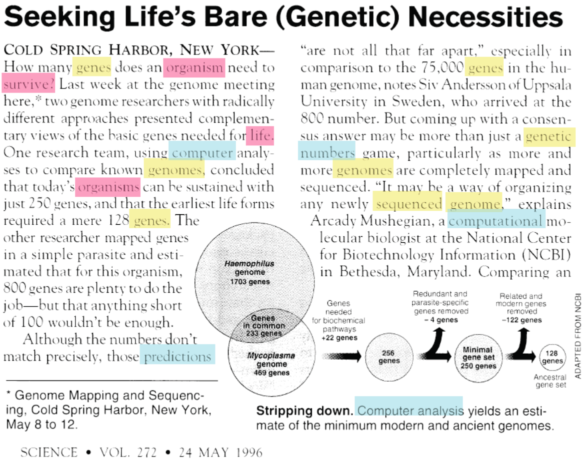
---

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

SCIENCE • VOL. 272 • 24 MAY 1996

Highlighted words: 'blue': data analysis, 'pink': evolutionary biology, 'yellow': genetics

Intuition: Documents exhibit multiple topics



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.

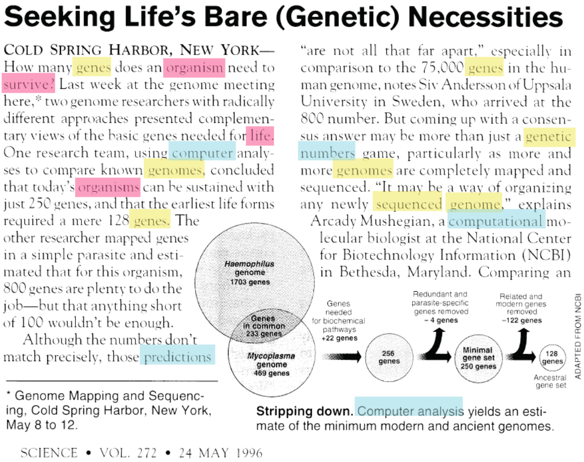
---

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

SCIENCE • VOL. 272 • 24 MAY 1996

The article blends genetics, data analysis and evolutionary biology in different proportions

Intuition: Documents exhibit multiple topics



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

Knowing that this article blends those topics would help situate it in a collection of scientific articles

Topic Model: Basic Idea

A generative statistical model that captures this intuition.

Generative Model

Documents are mixture of topics, where a topic is a probability distribution over words.

Topic Model: Basic Idea

A generative statistical model that captures this intuition.

Generative Model

Documents are mixture of topics, where a topic is a probability distribution over words.

genetics topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability.

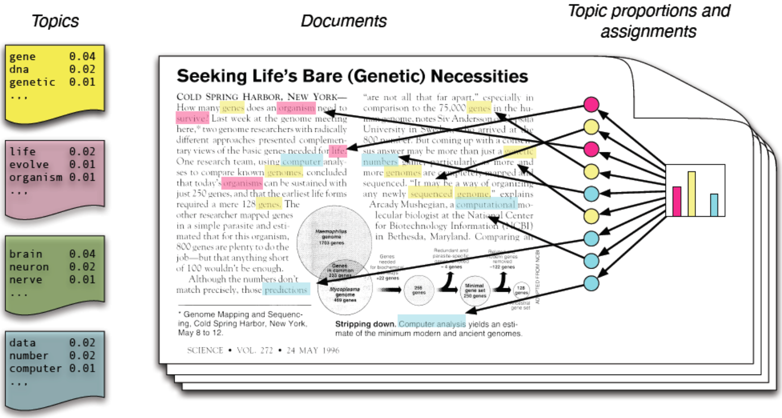
Topic Model: Basic Idea

A generative statistical model that captures this intuition.

Generative Model

Documents are mixture of topics, where a topic is a probability distribution over words.
genetics topic has words about genetics with high probability and the *evolutionary biology* topic has words about evolutionary biology with high probability.
Technically, the generative model assumes that the topics are generated first, before the documents.

Generative Model for 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.

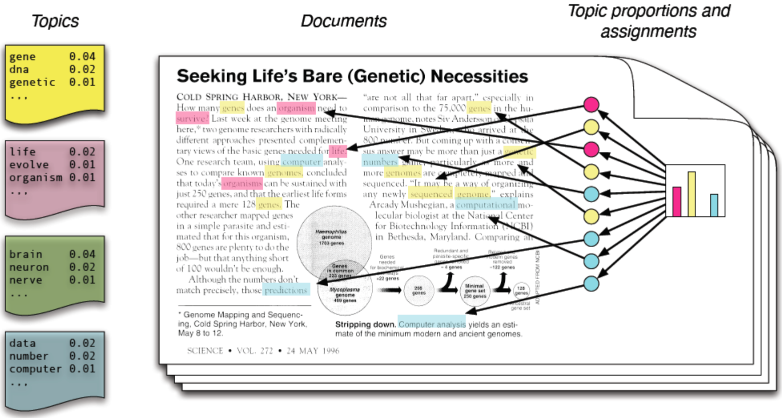
---

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

SCIENCE • VOL. 272 • 24 MAY 1996

- 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

Generative Model for LDA



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
...	

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.

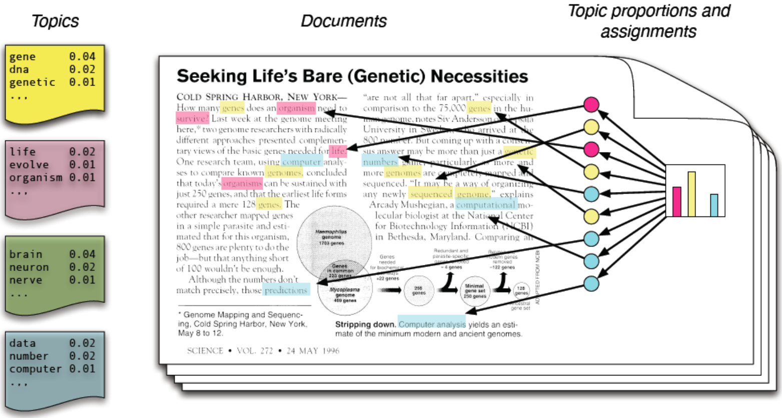
---

* 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

Generative Model for 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

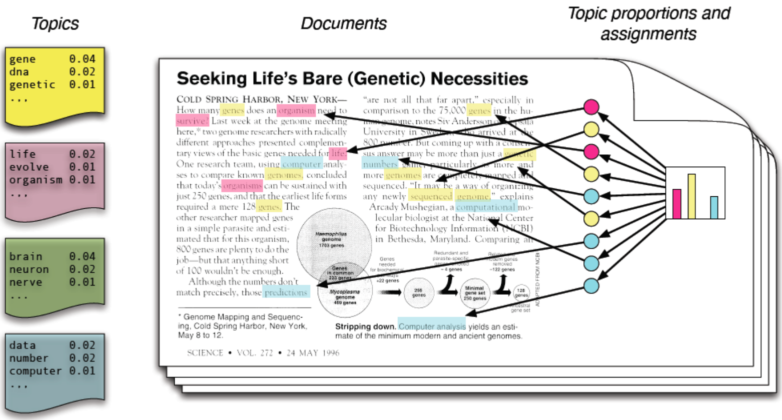
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

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

SCIENCE • VOL. 272 • 24 MAY 1996

- 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

Generative Model for LDA



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
...	

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

What does the statistical model reflect?

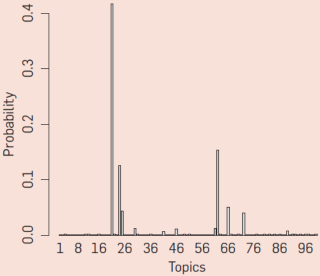
- All the document in the collection share the same set of topics, but each document exhibits those topics in different proportions
- Each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics

What does the statistical model reflect?

- All the document in the collection share the same set of topics, but each document exhibits those topics in different proportions
- Each word in each document is drawn from one of the topics, where the selected topic is chosen from the per-document distribution over topics

In the example article, the distribution over topics would place probability on *genetics*, *data analytics* and *evolutionary biology*, and each word is drawn from one of those three topics.

Real Inference with LDA for the example article

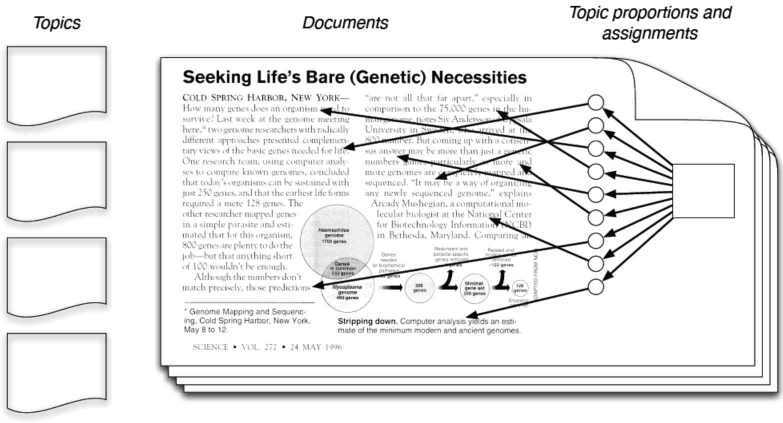


"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Central Problem of LDA

- The documents themselves are observed, while the topic structure - the topics, per-document topic distributions, and the per-document per-word topic assignments - is *hidden structure*.
- The central computational problem is to use the observed documents to infer the hidden topic structure, i.e. *reversing* the generative process.

Goal: The posterior distribution



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

> *Infer the hidden variables*
>
> *Compute their distribution conditioned on the documents*

Topics from TASA corpus

37,000 text passages from educational materials (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

Topics from TASA corpus

Documents with different content can be generated by choosing different distributions over topics.

- Equal probability to first two topics:

Topics from TASA corpus

Documents with different content can be generated by choosing different distributions over topics.

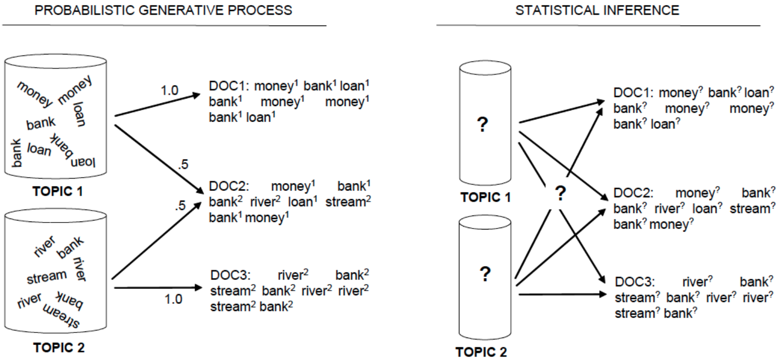
- Equal probability to first two topics: about a person who has taken too many drugs and how that affected color perceptions.
- Equal probability to the last two topics:

Topics from TASA corpus

Documents with different content can be generated by choosing different distributions over topics.

- Equal probability to first two topics: about a person who has taken too many drugs and how that affected color perceptions.
- Equal probability to the last two topics: about a person who experienced a loss of memory, which required a visit to the doctor.

Generative model and statistical inference



PROBABILISTIC GENERATIVE PROCESS

DOC1: money^1 bank^1 loan^1 bank^1 money^1 money^1 bank^1 loan^1

DOC2: money^1 bank^1 bank^2 river^2 loan^1 stream^2 bank^1 money^1

DOC3: river^2 bank^2 stream^2 bank^2 river^2 river^2 stream^2 bank^2

STATISTICAL INFERENCE

DOC1: money$^?$ bank$^?$ loan$^?$ bank$^?$ money$^?$ money$^?$ bank$^?$ loan$^?$

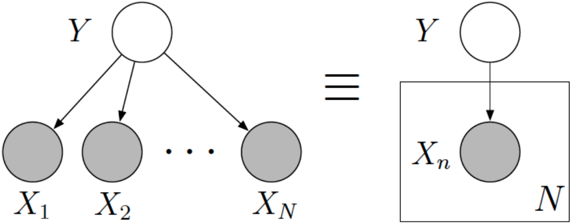
DOC2: money$^?$ bank$^?$ bank$^?$ river$^?$ loan$^?$ stream$^?$ bank$^?$ money$^?$

DOC3: river$^?$ bank$^?$ stream$^?$ bank$^?$ river$^?$ river$^?$ stream$^?$ bank$^?$

Important points

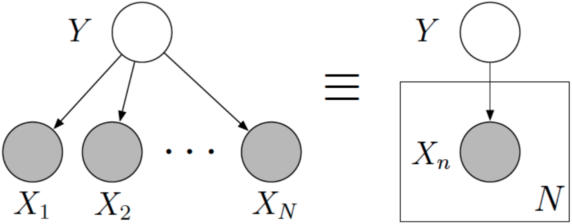
- *bag-of-words assumption:* The generative process does not make any assumptions about the order of words in the documents.
- *capturing polysemy:* The way that the model is defined, there is no notion of mutual exclusivity that restricts words to be part of one topic only. Ex: both ‘money’ and ‘river’ topics can give high probability to the word ‘bank’.

Graphical Model (Notation)



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

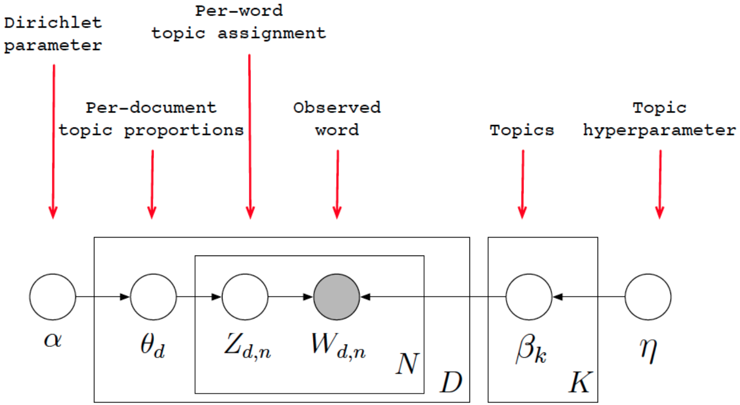
Graphical Model (Notation)



- Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables
- E.g., this graph corresponds to

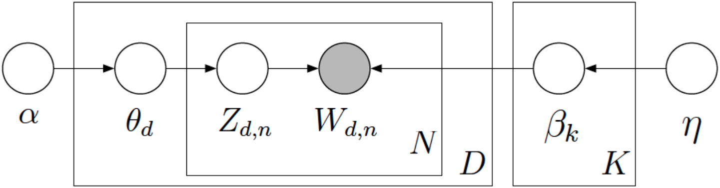
$$p(y,x_1,\ldots,x_N) = p(y)\prod_{n=1}^{N} p(x_n|y)$$

LDA: Graphical Model



Each piece of the structure is a random variable.

Latent Dirichlet Allocation: Generative Model



1. Draw each topic $\beta_i \sim \mathrm{Dir}(\eta)$, for $i \in \{1, \ldots, K\}$.
2. For each document:
   1. Draw topic proportions $\theta_d \sim \mathrm{Dir}(\alpha)$.
   2. For each word:
      1. Draw $Z_{d,n} \sim \mathrm{Mult}(\theta_d)$.
      2. Draw $W_{d,n} \sim \mathrm{Mult}(\beta_{z_{d,n}})$.

What is Latent Dirichlet Allocation (LDA)?

- ‘Latent’ has the same sense in LDA as in Latent semantic indexing, i.e. capturing topics as latent variables
- The distribution that is used to draw the per-document topic distributions is called a *Dirichlet distribution*. This result is used to allocate the words of the documents to different topics.

What is Latent Dirichlet Allocation (LDA)?

- ‘Latent’ has the same sense in LDA as in Latent semantic indexing, i.e. capturing topics as latent variables
- The distribution that is used to draw the per-document topic distributions is called a *Dirichlet distribution*. This result is used to allocate the words of the documents to different topics.

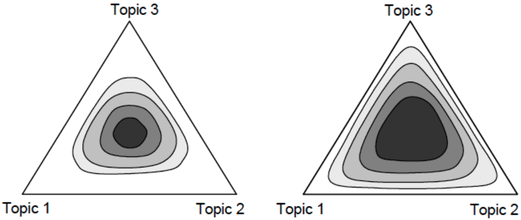
Dirichlet Distribution

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

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

Dirichlet Distribution

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

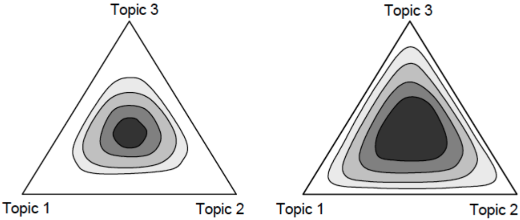


α_is: **hyper-parameters of the model:**

α_j can be interpreted as a prior observation count for the number of times topic j is sampled in a document

Dirichlet Distribution

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

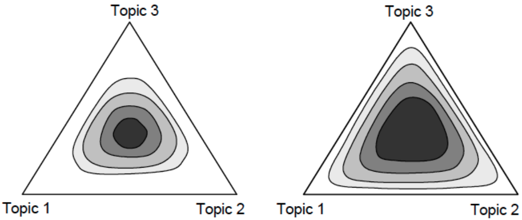


α_is: **hyper-parameters of the model:**

These priors can be interpreted as forces in the topic distributions with higher α moving the topics away from the corners of the simplex

Dirichlet Distribution

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

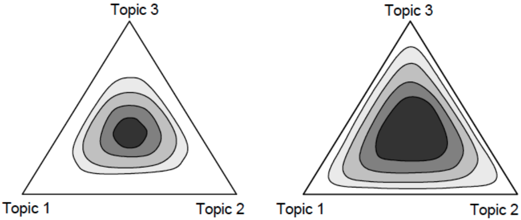


α_is: **hyper-parameters of the model:**

When $\alpha < 1$, there is a bias to pick topic distributions favoring just a few topics

Dirichlet Distribution

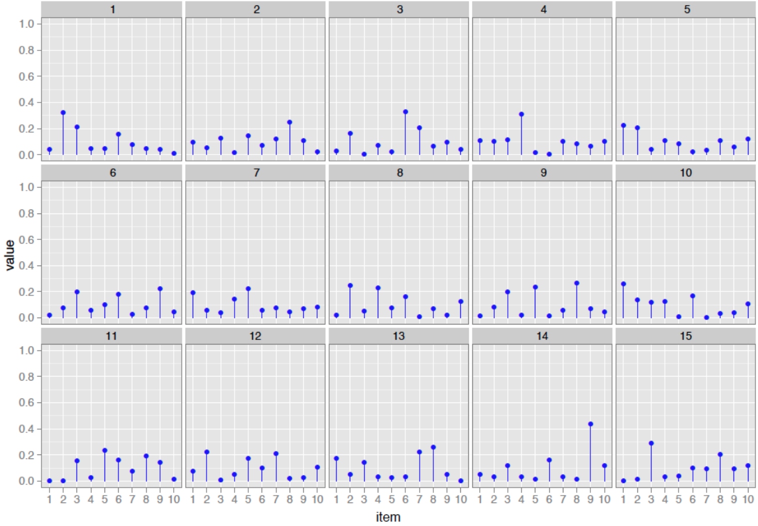
$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}$$



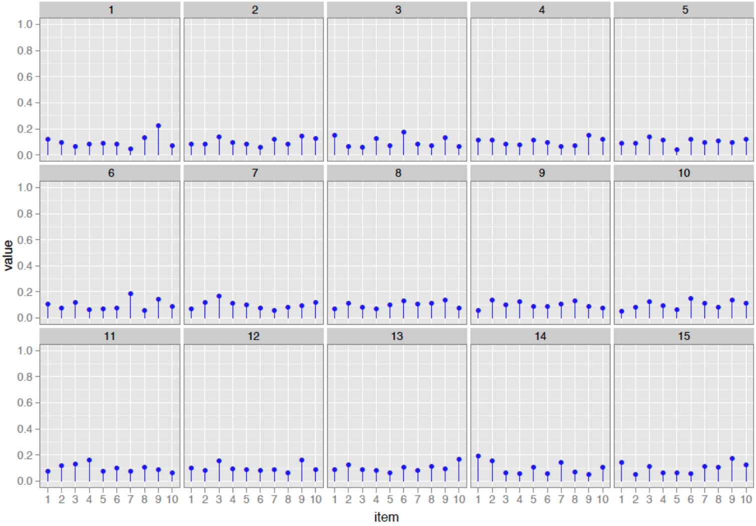
α_is: **hyper-parameters of the model:**

It is convenient to use a symmetric Dirichlet distribution with a single hyper-parameter $\alpha_1 = \alpha_2 \ldots = \alpha$

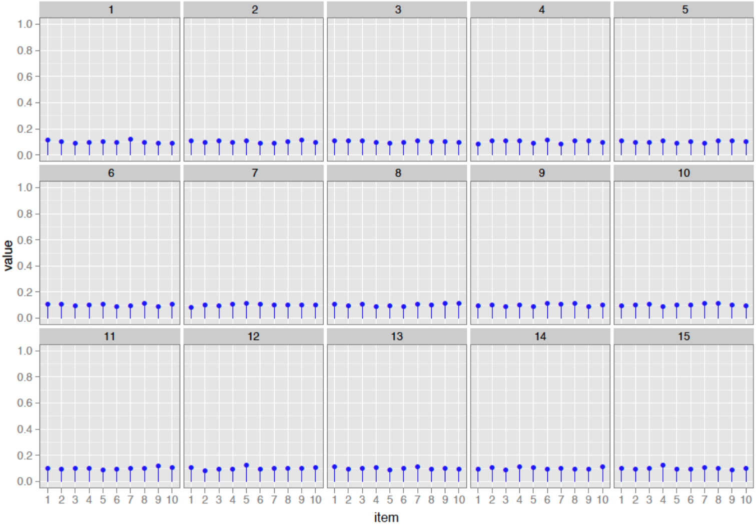
Effect of α: $\alpha = 1$



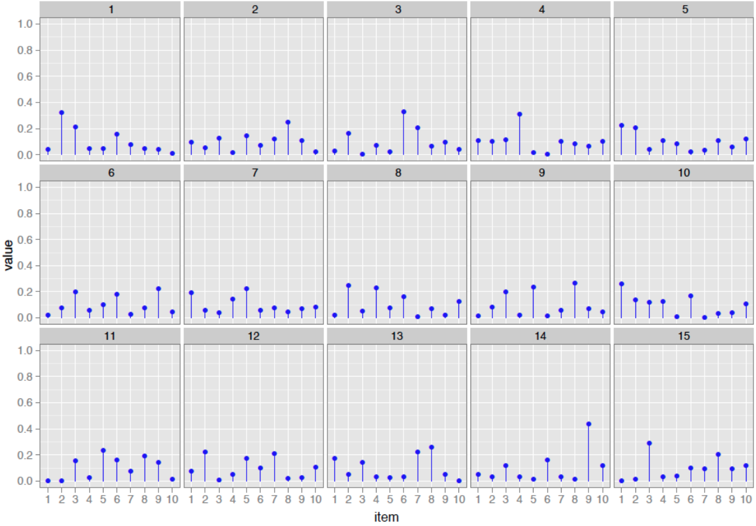
Effect of α: α = 10



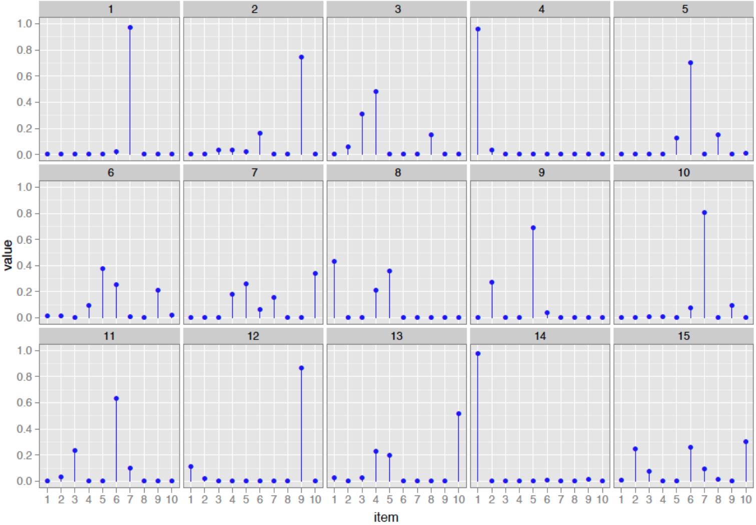
Effect of α: $\alpha = 100$



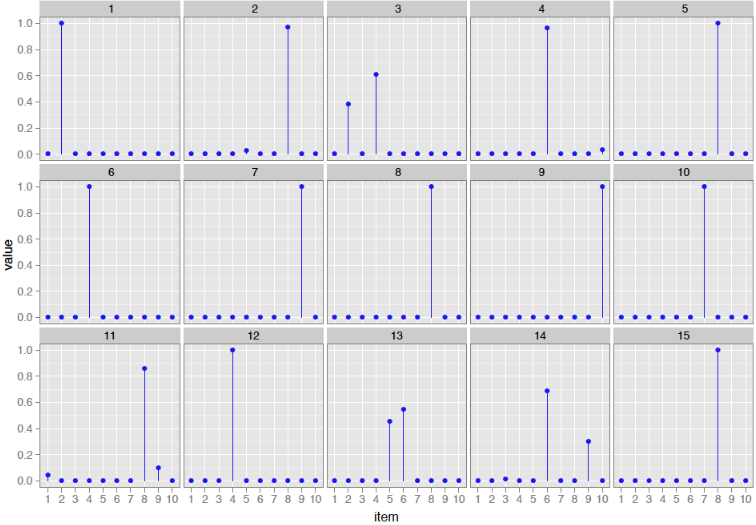
Effect of α: α = 1



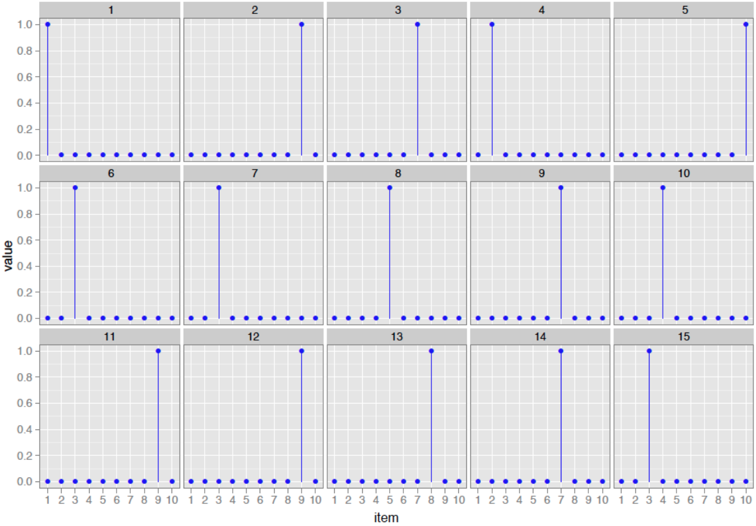
Effect of α: α = 0.1



Effect of α: $\alpha = 0.01$



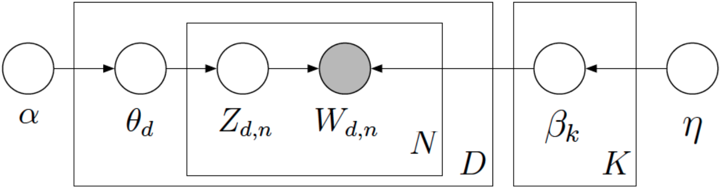
Effect of α: $\alpha = 0.001$



Online Implementations

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

Latent Dirichlet Allocation: Statistical Inference



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

Approximating the posterior

Algorithms to approximate it fall in two categories:

Sampling-based Algorithms

Collect samples from the posterior to approximate it with an empirical distribution

Approximating the posterior

Algorithms to approximate it fall in two categories:

Sampling-based Algorithms

Collect samples from the posterior to approximate it with an empirical distribution

Variational Methods

- Deterministic alternative to sampling-based algorithms
- The inference problem is transformed to an optimization problem

Gibbs Sampling

- A form of Markov chain Monte Carlo (MCMC), which simulates a high-dimensional distribution by sampling on lower-dimensional subset of variables where each subset is conditioned on the value of all others
- Sampling is done sequentially and proceeds until the sampled values approximate the target distribution
- It directly estimates the posterior distribution over z, and uses this to provide estimates for β and θ

Gibbs Sampling

- Suppose we have a word token i for which we want to find the topic assignment probability : $p(z_i = j)$
- Represent the collection of documents by a set of word indices w_i and document indices d_i for this token i
- Gibbs sampling considers each word token in turn and estimates the probability of assigning the current word token to each topic, conditioned on the topic assignment to all other word tokens
- From this conditional distribution, a topic is sampled and stored as the new topic assignment for this word token
- This conditional is written as $P(z_i = j | z_{-i}, w_i, d_i, .)$

Gibbs Sampling

- Let us define two matrices C^{WT} and C^{DT} of dimensions $W \times T$ and $D \times T$ respectively.
- ${C_{wj}}^{WT}$ contains the number of times word w is assigned to topic j, not including the current instance
- ${C_{dj}}^{WT}$ contains the number of times topic j is assigned to some word token in document d, not including the current instance

Gibbs Sampling

- Let us define two matrices C^{WT} and C^{DT} of dimensions $W \times T$ and $D \times T$ respectively.
- ${C_{wj}}^{WT}$ contains the number of times word w is assigned to topic j, not including the current instance
- ${C_{dj}}^{WT}$ contains the number of times topic j is assigned to some word token in document d, not including the current instance

$$P(z_i = j | z_{-i}, w_i, d_i, .) \propto \frac{{C_{w_{i}j}}^{WT} + \eta}{\sum_{w=1}^{W} {C_{wj}}^{WT} + W\eta} \frac{{C_{d_{i}j}}^{DT} + \alpha}{\sum_{j=1}^{T} {C_{dj}}^{DT} + T\alpha}$$

- The left part is the probability of word w under topic j (How likely a word is for a topic) whereas
- The right part is the probability of topic j under the current topic distribution for document d (How dominant a topic is in a document)

Algorithm

- Start: Each word token is assigned to a random topic in $[1 \ldots T]$
- For each word token, a new topic is sampled as per $P(z_i = j | z_{-i}, w_i, d_i, .)$, adjusting the matrices C^{WT} and C^{DT}
- A single pass through all word tokens in the document is one *Gibbs sample*
- After the burnin period, these samples are saved at regularly spaced intervals, to prevent correlations between samples

Estimating θ and β

$$\beta_i^{(j)} = \frac{C_{ij}^{WT} + \eta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\eta}$$

$$\theta_j^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

These values correspond to predictive distributions of

- sampling a new token of word i from topic j, and
- sampling a new token in document d from topic j

An Example

The algorithm can be illustrated by generating artificial data from a known topic model and applying the algorithm to check whether it is able to infer the original generative structure.

Example

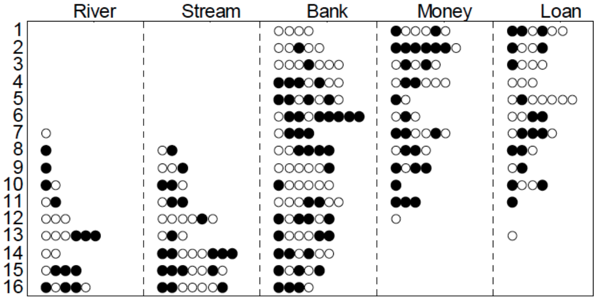
- Let topic 1 give equal probability to MONEY, LOAN, BANK and topic 2 give equal probability to words RIVER, STREAM, and BANK

$$\beta_{MONEY}{}^{(1)} = \beta_{LOAN}{}^{(1)} = \beta_{BANK}{}^{(1)} = 1/3$$

$$\beta_{RIVER}{}^{(2)} = \beta_{STREAM}{}^{(2)} = \beta_{BANK}{}^{(2)} = 1/3$$

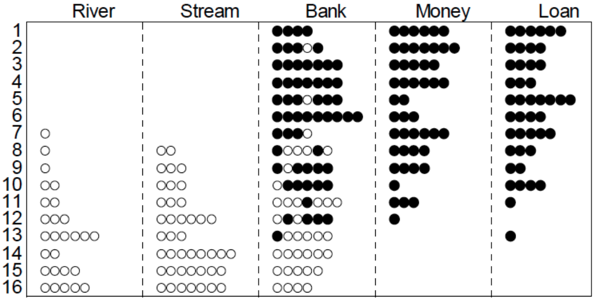
- We generate 16 documents by arbitrarily mixing two topics.

Initial Structure



Colors reflect initial random assignment, black = topic 1, white = topic 2

After 64 iterations of Gibbs Sampling



$$\beta_{MONEY}^{(1)} = 0.32, \beta_{LOAN}^{(1)} = 0.29, \beta_{BANK}^{(1)} = 0.39$$

$$\beta_{RIVER}^{(2)} = 0.25, \beta_{STREAM}^{(2)} = 0.4, \beta_{BANK}^{(2)} = 0.35$$

Computing Similarities

Document Similarity

Similarity between documents d_i and d_2 can be measured by the similarity between their topic distributions $\theta^{(d_1)}$ and $\theta^{(d_2)}$

KL divergence : $D(p,q) = \sum_{j=1}^{T} p_j log_2 \frac{p_j}{q_j}$

Symmetrized KL divergence: $\frac{1}{2}[D(p,q) + D(q,p)]$ seems to work well

Similarity with respect to query q

Maximize the conditional probability of query given the document:

$$p(q|d_i) = \prod_{w_k \in q} p(w_k|d_i)$$

$$= \prod_{w_k \in q} \sum_{j=1}^{T} P(w_k|z=j)P(z=j|d_i)$$

Computing Similarities

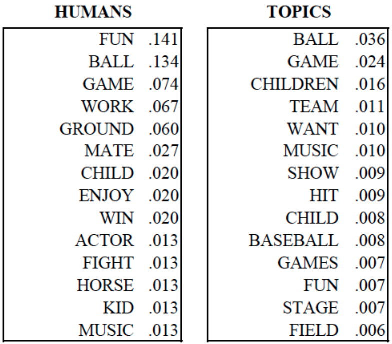
Similarity between two words

Having observed a single word in a new context, what are the other words that might appear in the same context, based on the topic interpretation for the observed word?

$$p(w_2|w_1) = \sum_{j=1}^{T} p(w_2|z=j)p(z=j|w_1)$$

Example

Observed and predicted responses for the word ‘PLAY’



HUMANS	
FUN	.141
BALL	.134
GAME	.074
WORK	.067
GROUND	.060
MATE	.027
CHILD	.020
ENJOY	.020
WIN	.020
ACTOR	.013
FIGHT	.013
HORSE	.013
KID	.013
MUSIC	.013

TOPICS	
BALL	.036
GAME	.024
CHILDREN	.016
TEAM	.011
WANT	.010
MUSIC	.010
SHOW	.009
HIT	.009
CHILD	.008
BASEBALL	.008
GAMES	.007
FUN	.007
STAGE	.007
FIELD	.006

Modeling Science

Data

The OCR'ed collection of *Science* from 1990-2000

- 17K documents
- 11M words
- 20K unique terms (stop words and rare words removed)

Modeling Science

Data

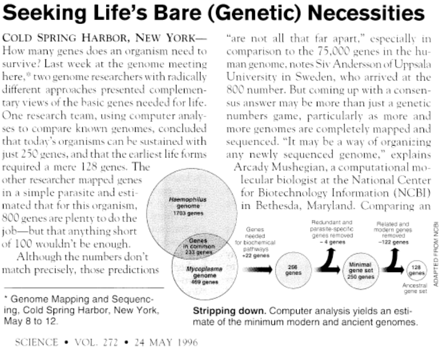
The OCR'ed collection of *Science* from 1990-2000

- 17K documents
- 11M words
- 20K unique terms (stop words and rare words removed)

Model

100-topic 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.

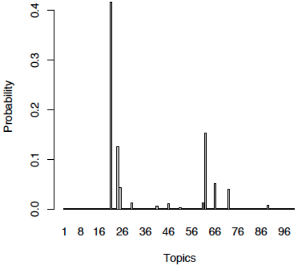
Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just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

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

Modeling Richer Assumptions in Topic Models

- Correlated topic models
- Dynamic topic models
- Measuring scholarly impact

Correlated Topic Models

- The Dirichlet is an exponential family distribution on the simplex, positive vectors that sum to one
- However, the near independence of components makes it a poor choice for modeling topic proportions
- An article about *fossil fuels* is more likely to also be about *geology* than about *genetics*

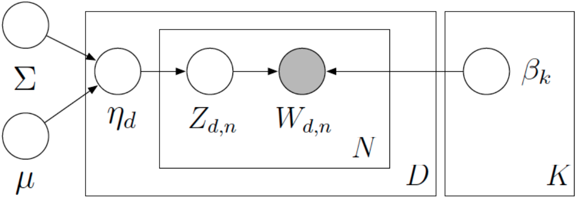
Using logistic normal distribution

A multivariate normal distribution of a k-dimensional vector $x = [X_1, X_2, \ldots, X_k]$ can be written as

$$x \sim \; N_k(\mu, \Sigma)$$

with k-dimensional mean vector μ and $k \times k$ covariance matrix Σ

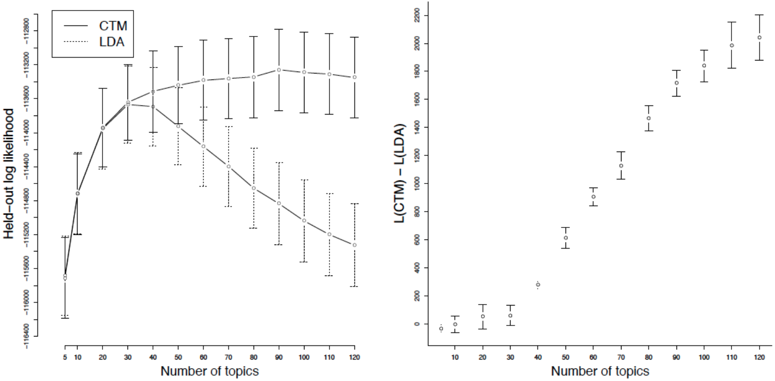
Correlated Topic Model (CTM)



Noconjugate prior
on topic proportions

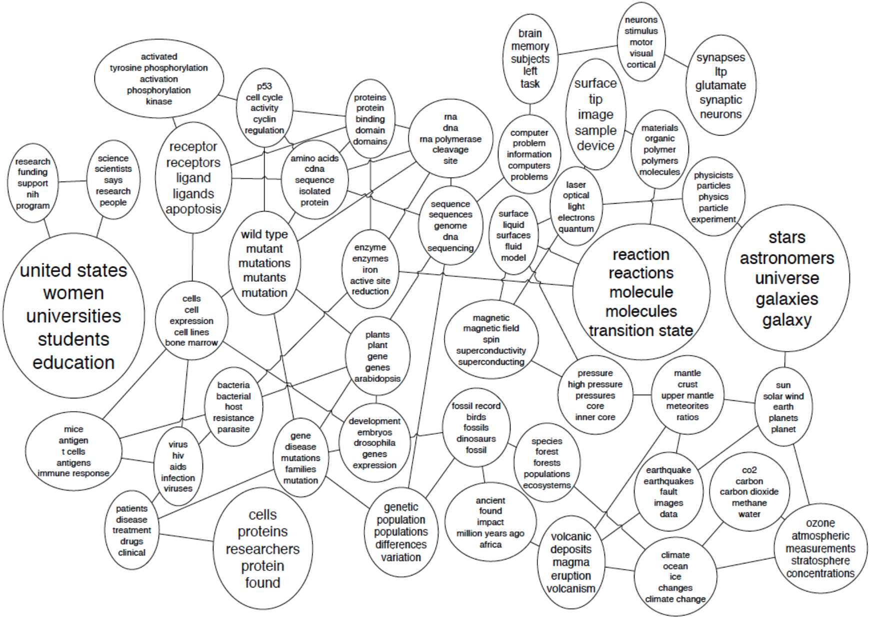
- Draw topic proportions from a logistic normal, where topic occurrences can exhibit correlation.
- Use for:
  - Providing a “map” of topics and how they are related
  - Better prediction via correlated topics

CTM supports more topics and provides a better fit than LDA



Held-out log probability on *Science*

Correlated Topics



activated
tyrosine phosphorylation
activation
phosphorylation
kinase

p53
cell cycle
activity
cyclin
regulation

proteins
protein
binding
domain
domains

rna
dna
rna polymerase
cleavage
site

brain
memory
subjects
left
task

neurons
stimulus
motor
visual
cortical

synapses
ltp
glutamate
synaptic
neurons

surface
tip
image
sample
device

computer
problem
information
computers
problems

materials
organic
polymer
polymers
molecules

research
funding
support
nih
program

science
scientists
says
research
people

receptor
receptors
ligand
ligands
apoptosis

amino acids
cdna
sequence
isolated
protein

physicists
particles
physics
particle
experiment

laser
optical
light
electrons
quantum

sequence
sequences
genome
dna
sequencing

surface
liquid
surfaces
fluid
model

wild type
mutant
mutations
mutants
mutation

enzyme
enzymes
iron
active site
reduction

united states
women
universities
students
education

stars
astronomers
universe
galaxies
galaxy

reaction
reactions
molecule
molecules
transition state

cells
cell
expression
cell lines
bone marrow

plants
plant
gene
genes
arabidopsis

magnetic
magnetic field
spin
superconductivity
superconducting

bacteria
bacterial
host
resistance
parasite

pressure
high pressure
pressures
core
inner core

mantle
crust
upper mantle
meteorites
ratios

sun
solar wind
earth
planets
planet

mice
antigen
t cells
antigens
immune response

virus
hiv
aids
infection
viruses

gene
disease
mutations
families
mutation

development
embryos
drosophila
genes
expression

fossil record
birds
fossils
dinosaurs
fossil

species
forest
forests
populations
ecosystems

earthquake
earthquakes
fault
images
data

co2
carbon
carbon dioxide
methane
water

patients
disease
treatment
drugs
clinical

cells
proteins
researchers
protein
found

genetic
population
populations
differences
variation

ancient
found
impact
million years ago
africa

volcanic
deposits
magma
eruption
volcanism

climate
ocean
ice
changes
climate change

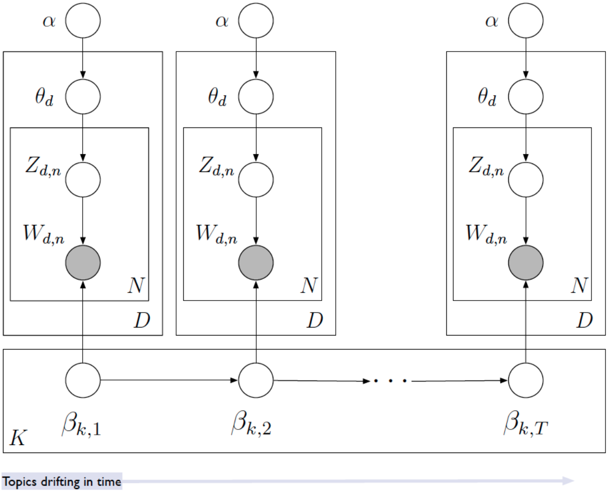
ozone
atmospheric
measurements
stratosphere
concentrations

Dynamic Topic Models

LDA assumption

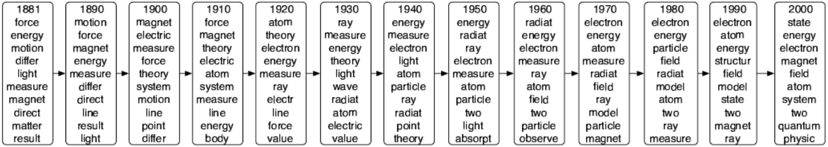
- LDA assumes that the order of documents does not matter
- Not appropriate for corpora that spans hundreds of years
- We might want to track how language changes over time

Dynamic Topic Models



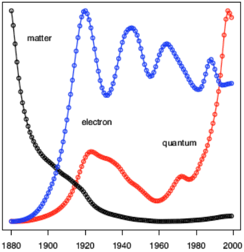
$$\beta_{k,t}|\beta_{k,t-1} \sim N(\beta_{k,t-1}, \sigma^2 I)$$

Dynamic Topic Models



1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
force	motion	magnet	force	atom	ray	energy	energy	radiat	electron	electron	electron	state
energy	force	electric	magnet	theory	measure	measure	radiat	energy	energy	energy	atom	energy
motion	magnet	measure	theory	electron	energy	electron	ray	electron	atom	particle	energy	electron
differ	energy	force	electric	energy	theory	light	electron	measure	measure	field	structur	magnet
light	measure	theory	atom	measure	light	atom	measure	ray	radiat	radiat	field	field
measure	differ	system	system	ray	wave	particle	atom	atom	field	model	model	atom
magnet	direct	motion	measure	electr	radiat	ray	particle	field	ray	atom	state	system
direct	line	line	line	line	atom	radiat	two	two	model	two	two	two
matter	result	point	energy	force	electric	point	light	particle	particle	ray	magnet	quantum
result	light	differ	body	value	value	theory	absorpt	observe	magnet	measure	ray	physic

"Atomic Physics"

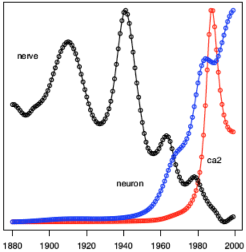


1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

Dynamic Topic Models

1881	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
brain	movement	brain	movement	movement	stimulate	record	respons	response	respons	cell	cell	neuron
movement	eye	eye	brain	sound	muscle	nerve	record	stimulate	cell	neuron	channel	active
action	right	movement	sound	muscle	sound	stimulate	stimulate	record	potential	response	neuron	brain
right	hand	right	nerve	active	movement	response	nerve	condition	stimul	active	ca2	cell
eye	brain	left	active	nerve	response	muscle	muscle	active	neuron	brain	active	fig
hand	left	hand	muscle	stimulate	nerve	electrode	active	potential	active	stimul	brain	response
left	action	nerve	left	fiber	frequency	active	frequency	stimulus	nerve	muscle	receptor	channel
muscle	muscle	vision	eye	reaction	fiber	brain	electrode	nerve	eye	system	muscle	receptor
nerve	sound	sound	right	brain	active	fiber	potential	subject	record	nerve	respons	synapse
sound	experiment	muscle	nervous	response	brain	potential	study	eye	abstract	receptor	current	signal

"Neuroscience"



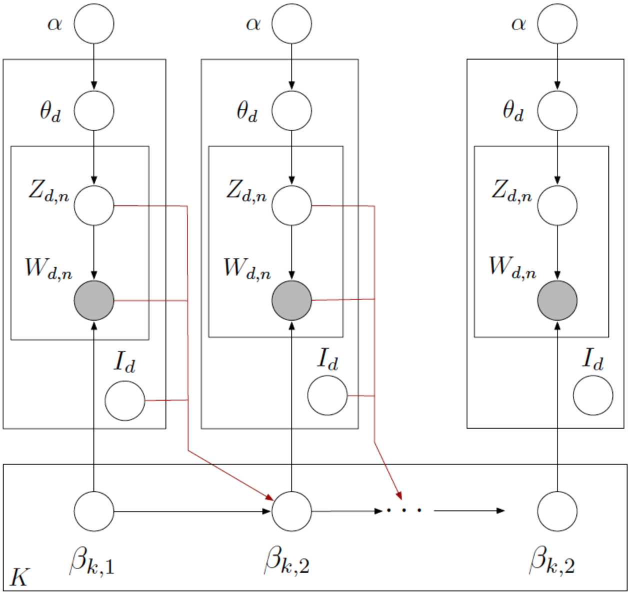
1887 Mental Science
1900 Hemianopsia in Migraine
1912 A Defence of the ``New Phrenology"
1921 The Synchronal Flashing of Fireflies
1932 Myoesthesis and Imageless Thought
1943 Acetylcholine and the Physiology of the Nervous System
1952 Brain Waves and Unit Discharge in Cerebral Cortex
1963 Errorless Discrimination Learning in the Pigeon
1974 Temporal Summation of Light by a Vertebrate Visual Receptor
1983 Hysteresis in the Force-Calcium Relation in Muscle
1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

Measuring Scholarly Impact

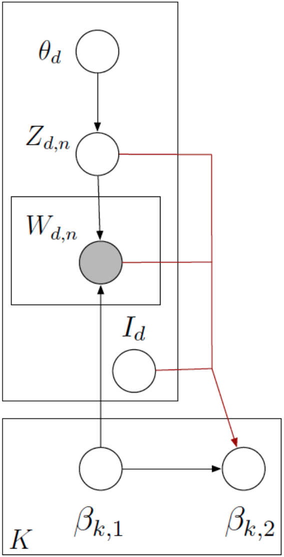
How to model influence?

- *Idea from Dynamic Topic Models*, influential articles reflect future changes in language use
- The *influence* of an article is a latent variable
- Influential articles affect the drift of the topics that they discuss
- The posterior gives a retrospective estimate of influential article

Measuring Scholarly Impact



Measuring Scholarly Impact



- Each document has an influence score I_d.
- Each topic drifts in a way that is biased towards the documents with high influence.
- The posterior of $I_{1:D}$ can be examined to retrospectively find articles that best explain future changes in language.

Example Problem

Suppose you are using Gibbs sampling to estimate the distributions, θ and β for topic models. The underlying corpus has 5 documents and 5 words, $\{$*River, Stream, Bank, Money, Loan*$\}$ and the number of topics is 2. At certain point, the structure of the documents looks like the following Table. For instance, the first row indicates that the document 1 contains 4 instances of word 'Bank', 6 instances of word 'Money' and 6 instances of word 'Loan'. Black and white circles denote whether the word is currently assigned to topics t_1 and t_2 respectively.

Use this structure to estimate $\beta_{MONEY}{}^{(2)}$ and $\beta_{BANK}{}^{(1)}$ at this point. You can take the values of η and α to be 0.1 each.

Doc. Id	River	Stream	Bank	Money	Loan
1			●●●●	●●●●●●	●●●●●●
2			●●●○●	●●●●●●●	●●●●
3	○	○○○	●○○○●○	●●●●	●●●
4	○○○○○○	○○○	●○○○○○		
5	○○	○○○○○○	○○○○○○		