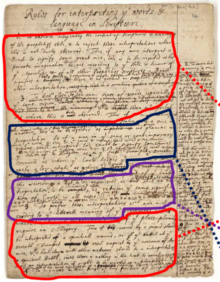
IFI 9000 Analytics Methods
Topic Models

by **Houping Xiao**

Spring 2021

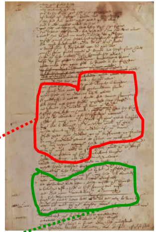
Introduction

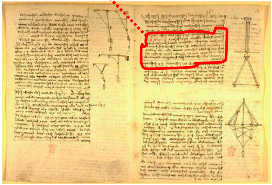
What is a "topic"?



Representation: a probabilistic distribution over words.

retrieval	0.2
information	0.15
model	0.08
query	0.07
language	0.06
feedback	0.03
……	





Topic: A broad concept/theme, semantically coherent, which is *hidden* in documents

e.g., politics; sports; technology; entertainment; education etc.

Consider a document as a mixture of topics

[Criticism of government response to the hurricane primarily consisted of criticism of its response to the approach of the storm and its aftermath, specifically in the delayed response] to the [flooding of New Orleans. … 80% of the 1.3 million residents of the greater New Orleans metropolitan area evacuated] …[Over seventy countries pledged monetary donations or other assistance]. …

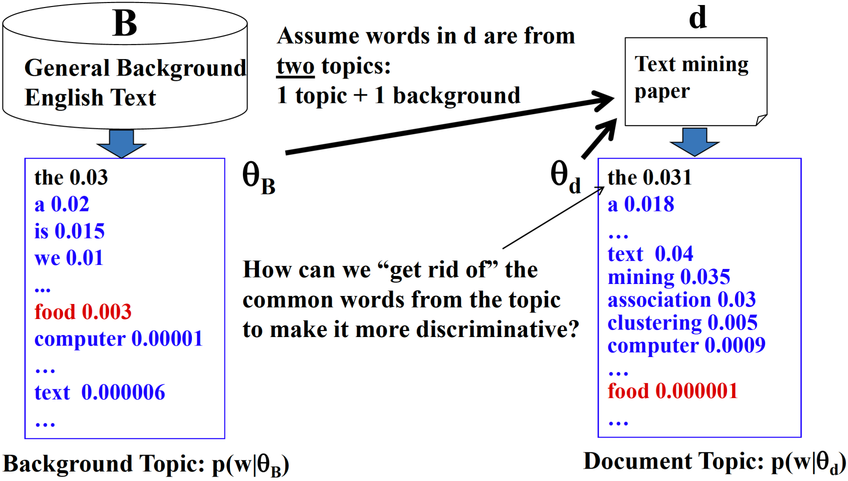
- Topic θ_1: [government 0.3, response 0.2, ...]
- Topic θ_2: [city 0.2, new 0.1, orleans 0.05, ...]
- ...
- Topic θ_k: [donate 0.1, relief 0.05, help 0.02, ...]
- Background θ_0: [is 0.05, the 0.04, 1 0.03, ...]

- How can we discover $\theta_0, \cdots, \theta_k$
- Many applications would be enabled by discovering such topics
  - summarize themes, retrieve documents, segment documents, etc

Basic ideas of topic models

- A topic is a multinomial distribution over words
- A document is a mixture of topics (How a document is "generated"?)
  - sampling topics from a prior
  - sampling a word at a time from the distribution given the topic
- Topic modeling
  - Fitting the topic model to the text
  - Answering topic-related questions by computing various kinds of posterior distributions, e.g., $p(topic|time)$, $p(sentiment|topic)$

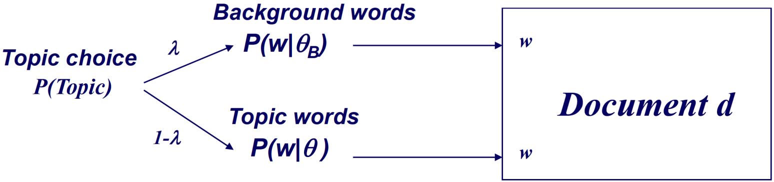
Topic modeling: An example with 1 topic + 1 "background"



Topic modeling: An example with 1 topic + 1 "background"

Assume $p(w|\theta_B)$ and λ are *known*

λ = mixing proportion of background topic in d



$$p(w) = \lambda p(w \mid \theta_B) + (1-\lambda) p(w \mid \theta)$$

$$\log p(d \mid \theta) = \sum_{w \in V} c(w,d) \log[\lambda p(w \mid \theta_B) + (1-\lambda) p(w \mid \theta)]$$

Expectation Maximization $\hat{\theta} = \arg\max_{\theta} \log p(d \mid \theta)$

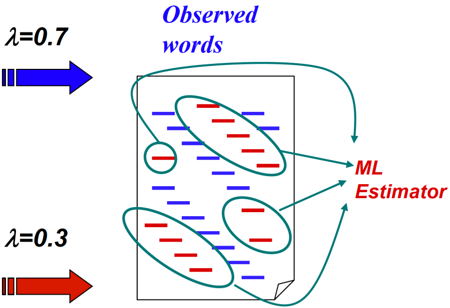
How to estimate topic-word distributions θ?

Known Background $p(w|\theta_B)$

the 0.2
a 0.1
we 0.01
to 0.02
…
text 0.0001
mining 0.00005
…

Unknown topic $p(w|\theta)$ for "Text mining"

…
text =?
mining =?
association =?
word =?
…



Suppose we know the identity/label of each word …

But we don't!

But, we can make a guess!

We guess the topic assignments

- Assignment a hidden variable $z_i \in \{1(background), 0(topic)\}$

	z_i
the	1
paper	1
presents	1
a	1
text	0
mining	0
algorithm	0
the	1
paper	0
...	...

Suppose the parameters are all known, what's a reasonable guess of z_i?

- depends on λ

- depends on $p(w|\theta_B)$ and $p(w|\theta)$

$$p(z_i=1|w_i) = \frac{p(z_i=1)p(w|z_i=1)}{p(z_i=1)p(w|z_i=1)+p(z_i=0)p(w|z_i=0)}$$

$$= \frac{\lambda p(w|\theta_B)}{\lambda p(w|\theta_B)+(1-\lambda)p^{current}(w|\theta)}$$ ***E-step***

$$p^{new}(w_i|\theta) = \frac{c(w_i,d)(1-p(z_i=1|w_i))}{\sum_{w'\in V} c(w',d)(1-p(z_i=1|w'))}$$ ***M-step***

θ_B and θ are competing for explaining words in document d!

- Initialization: $p(w|\theta)$ is set randomly
- EM iteration

An example of EM algorithm

$$p^{(n)}(z_i = 1 | w_i) = \frac{\lambda p(w_i | \theta_B)}{\lambda p(w_i | \theta_B) + (1-\lambda) p^{(n)}(w_i | \theta)}$$

Expectation-Step:
Augmenting data by guessing hidden variables

$$p^{(n+1)}(w_i | \theta) = \frac{c(w_i, d)(1 - p^{(n)}(z_i = 1 | w_i))}{\sum_{w_j \in vocabulary} c(w_j, d)(1 - p^{(n)}(z_j = 1 | w_j))}$$

Maximization-Step
With the "augmented data", estimate parameters using maximum likelihood

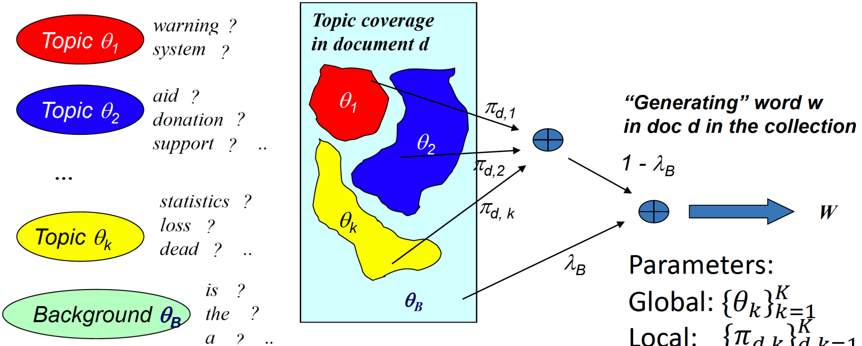
Assume λ=0.5

Word	#	$P(w\|\theta_B)$	Iteration 1		Iteration 2		Iteration 3	
			$P(w\|\theta)$	$P(z=1)$	$P(w\|\theta)$	$P(z=1)$	$P(w\|\theta)$	$P(z=1)$
The	4	0.5	**0.25**	0.67	**0.20**	0.71	**0.18**	0.74
Paper	2	0.3	**0.25**	0.55	**0.14**	0.68	**0.10**	0.75
Text	4	0.1	**0.25**	0.29	**0.44**	0.19	**0.50**	0.17
Mining	2	0.1	**0.25**	0.29	**0.22**	0.31	**0.22**	0.31
Log-Likelihood			-16.96		-16.13		-16.02	

Models

- Probabilistic Latent Semantic Analysis (pLSA)
- Latent Dirichlet Allocation (LDA)
- Correlation Explanation (CorEx) *(not covered in this course)*

Generalize to $k \geq 2$ topics



Parameters:

Global: $\{\theta_k\}_{k=1}^{K}$

Local: $\{\pi_{d,k}\}_{d,k=1}^{K}$

Manual: λ_B

pLSA

- T. Hofmann, Probabilistic latent semantic indexing, 1999
- Topic: a multinomial distribution over words
- Document
  - a mixture of k topics
  - mixing weights reflect the topic coverage
- Topic modeling
  - word distribution under topic: $p(w|\theta)$
  - topic coverage: $p(\pi|d)$

EM for estimating multiple topics

Known Background $p(w \mid \theta_B)$

the 0.2
a 0.1
we 0.01
to 0.02
…

Unknown topic model $p(w|\theta_1)$=?

"Text mining"

…
text =?
mining =?
association =?
word =?
…

…

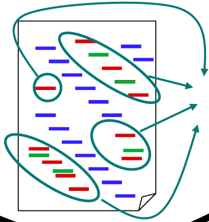
Unknown topic model $p(w|\theta_2)$=?

"information retrieval"

…
information =?
retrieval =?
query =?
document =?
…

E-Step:
Predict topic labels using Bayes Rule

Observed Words



M-Step:
ML Estimator based on "fractional counts"

Model parameter estimation

- E-step: word w in doc d is generated
  - from topic j

$$p(z_{d,w} = j) = \frac{\pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}{\sum_{j'=1}^{k} \pi_{d,j'}^{(n)} p^{(n)}(w|\theta_{j'})}$$

  - from background

$$p(z_{d,w} = B) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1-\lambda_B) \sum_{j=1}^{k} \pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}$$

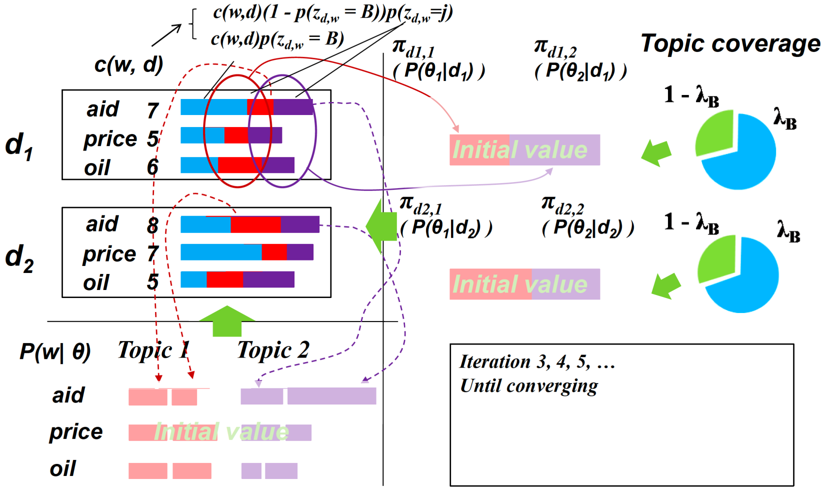
- M-step: re-estimate
  - mixing weights

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w \in V} c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)}{\sum_{j'} \sum_{w \in V} c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j')}$$

  - word-topic distribution

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d \in C} c(w,d)(1-p(z_{d,w}=B))p(z_{d,w}=j)}{\sum_{w' \in V} \sum_{d \in C} c(w',d)(1-p(z_{d,w'}=B))p(z_{d,w'}=j)}$$

How the algorithm works?



Sample pLSA topics from TDT corpus

"plane"	"space shuttle"	"family"	"Hollywood"
plane	space	home	film
airport	shuttle	family	movie
crash	mission	like	music
flight	astronauts	love	new
safety	launch	kids	best
aircraft	station	mother	hollywood
air	crew	life	love
passenger	nasa	happy	actor
board	satellite	friends	entertainment
airline	earth	cnn	star

pLSA with prior knowledge

- What if we have some domain knowledge in mind
  - we want to see topics such as "battery" and "memory" for opinions about a laptop
  - we want words like "aple" and "orange" co-occur in a topic
  - one topic should be fixed to model background words
- We can easily incorporate such knowledge as priors of pLSA model

Deficiency of pLSA

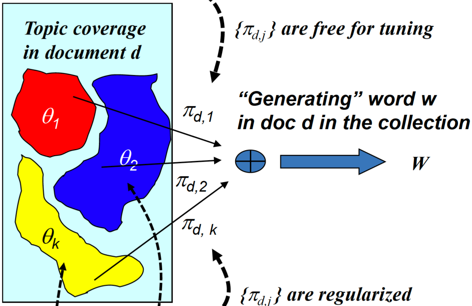
- Not a fully generative mdoel
  - can't compute probability of a new document
  - heuristic wordaround is possible
- Many parameters to estimate, high complexity of mdoels
  - many local maxima
  - prone to overfitting

LDA

- Make pLSA a fully generative model by imposing Dirichlet priors
  - Dirichlet priors over $p(\pi|d)$
  - Dirichlet priors over $p(w|\theta)$
  - a Bayesian version of pLSA
- Provide mechanism to deal with new documents
  - flexible to model many other observations in a document

LDA = pLSA with Dirichlet priors

pLSA:
Topic coverage $\pi_{d,j}$ is specific to each "training document", thus can't be used to generate a new document



$\{\pi_{d,j}\}$ are regularized

LDA:
Topic coverage distribution $\{\pi_{d,j}\}$ for any document is sampled from a Dirichlet distribution, allowing for generating a new doc

$$p(\vec{\pi}_d) = Dirichlet(\vec{\alpha})$$

Magnitudes of α and β determine the variances of the prior, thus also the concentration of prior (larger α and β ➔ stronger prior)

In addition, the topic word distributions $\{\theta_j\}$ are also drawn from another Dirichlet prior

$$p(\vec{\theta}_i) = Dirichlet(\vec{\beta})$$

pLSA v.s. LDA

- pLSA

$$p_d(w|\{\theta_j\},\{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j} p(w|\theta_j)$$

$$\log p(d|\{\theta_j\},\{\pi_{d,j}\}) = \sum_{w\in V} c(w,d) \log[\sum_{j=1}^{k} \pi_{d,j} p(w|\theta_j)]$$

$$\log p(C|\{\theta_j\},\{\pi_{d,j}\}) = \sum_{d\in C} \log p(d|\{\theta_j\},\{\pi_{d,j}\})$$

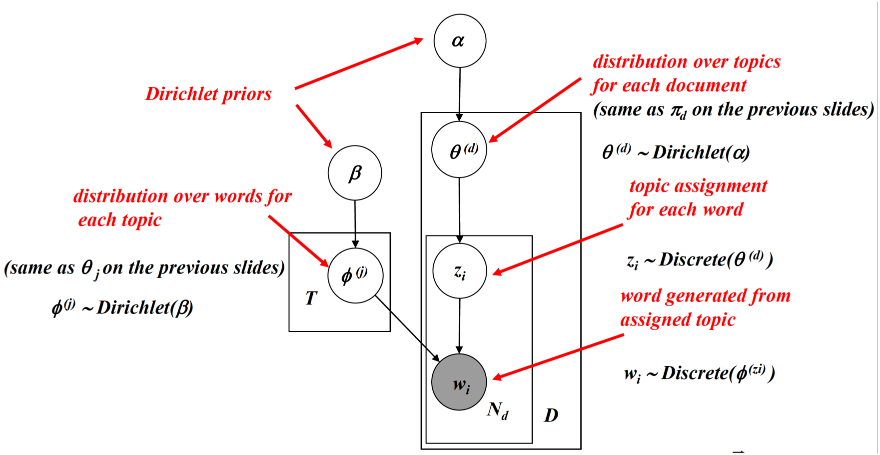
- LDA

$$p_d(w|\{\theta_j\},\{\pi_{d,j}\}) = \sum_{j=1}^{k} \pi_{d,j} p(w|\theta_j)$$

$$\log p(d|\{\theta_j\},\boldsymbol{\alpha}) = \int \sum_{w\in V} c(w,d) \log[\sum_{j=1}^{k} \pi_{d,j} p(w|\theta_j)] p(\boldsymbol{\pi}_d|\boldsymbol{\alpha}) d\boldsymbol{\pi}_d$$

$$\log p(C|\boldsymbol{\alpha},\boldsymbol{\beta}) = \int \sum_{d\in C} \log p(d|\{\theta_j\},\boldsymbol{\alpha}) \prod_{j=1}^{k} p(\theta_j|\boldsymbol{\beta}) d\theta_1 \cdots d\theta_k$$

LDA as a graphical model



- Most approximate inference algoritms aim to infer $p(z_i|\mathbf{w}, \alpha, \beta)$ from which other interesting variables can be easily computed

Approximate inferences for LDA

- Deterministic approximation
  - variational inference
  - expectation propagation
- Markov chain Monte Carlo
  - full Gibbs sampler
  - collapsed Gibbs sampler: most efficient and popular, but can only work with conjugate prior

Topics learned by LDA

"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

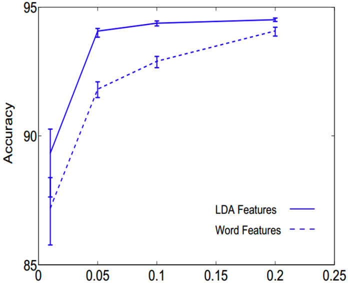
Topic assignments in document

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.

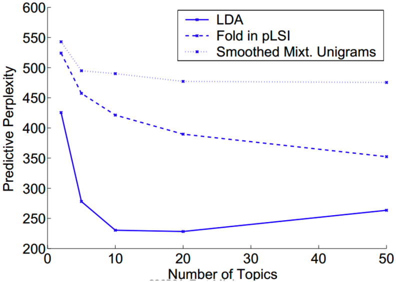
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY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

How to use the topics?

- document classification
  - a new type of feature representation

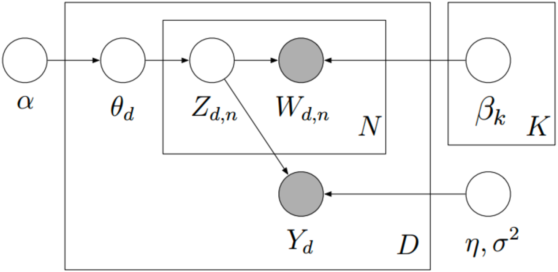


- Collaborative filtering
  - a new type of user profile

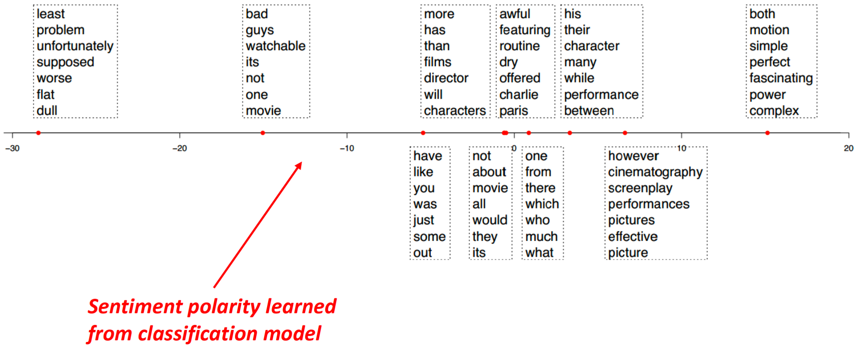


Supervised topic model

- A generative model for classification
  - topic generates both words and labels



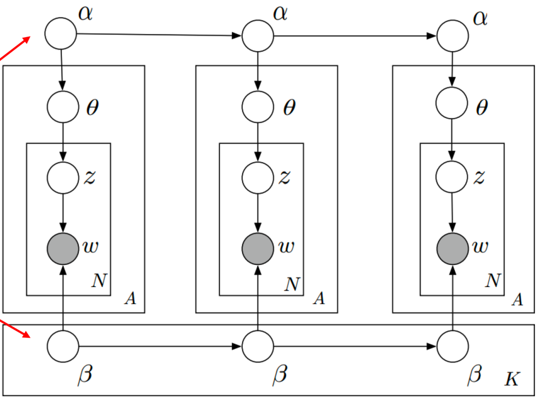
sentiment polarity of topics



Dynamic topic model

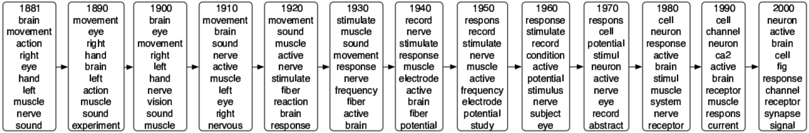
- Capture the evolving topics over time

Markov assumption about the topic dynamics

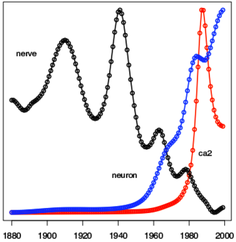


Dynamic topic model

- Evolution of topics



"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

Summary

- Topic models are a new family of document modeling approaches, especially useful for
  - discovering latent topics in text
  - analyzing latent structures and patterns of topics
  - extensible for joint modeling and analysis of text and associated non-textual data
- pLSA and LDA are two basic topic models (more variants or models) that tend to function similarly, with LDA better as a generative model
- However, all topic models suffer from the problem of multiple local maxima
  - make it hard and impossible to reproduce research results
  - make it hard and impossible to interpret results in real applications
- Complex models can't scale up to handle large amounts of text data
  - collapsed Gibbs sampling is efficient, but only working for conjugate priors
  - parallel alorithms are promising
  - ...

The End