Topic Models and Applications to Short Documents

Dieu-Thu Le

Email: dieuthu.le@unitn.it
Trento University

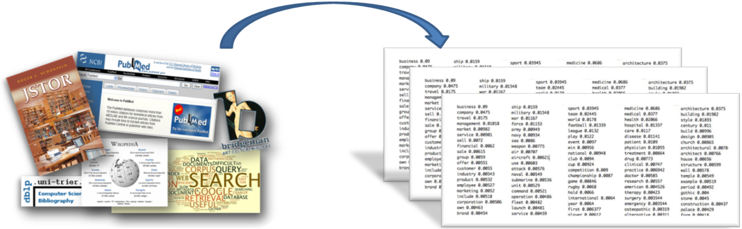
April 6, 2011

Outline

Problems with data collections

- With the availability of large document collections online, it becomes more difficult to represent and extract knowledge from them
- We need new tools to organize and understand these vast collections

Topic Models



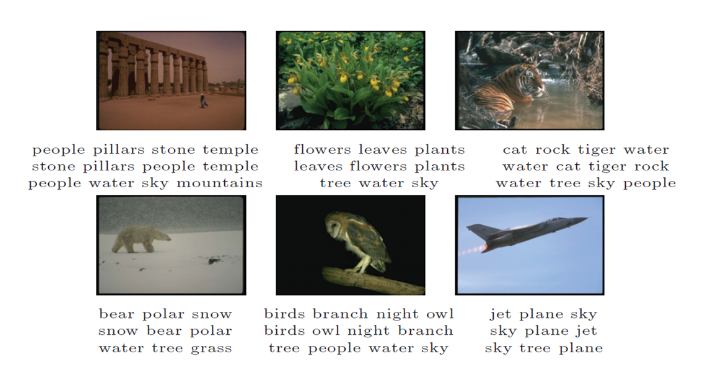
Topic Models provide methods for statistical analysis of document collections & other discrete data

- Uncover the hidden topical patterns in the collection
- Discover patterns of word-use and connect documents that exhibit similar patterns

Discover Topics from a Document Collection

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

Image Annotation with Topic Models



[1]

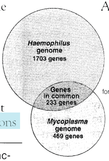
[1] Source: Y.Shao et al. Semi-supervised topic modeling for image annotation, 2009

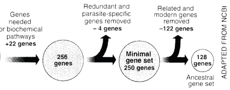
Intuition behind LDA (Latent Dirichlet Allocation)

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.

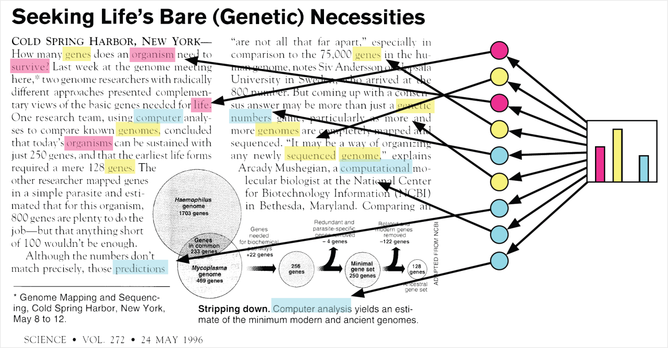
SCIENCE • VOL. 272 • 24 MAY 1996

2

Simple intuition: Documents exhibit multiple topics

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

Generative Process


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.

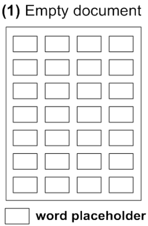
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

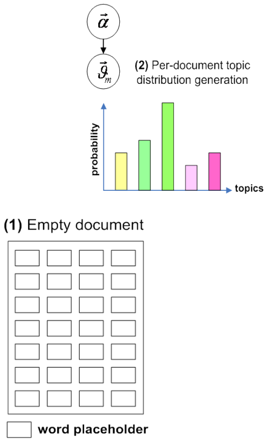
Cast this intuition into a probabilistic procedure by which documents can be generated:

- Choose a distribution over topics for a document
- For each word, choose a topic according to the distribution

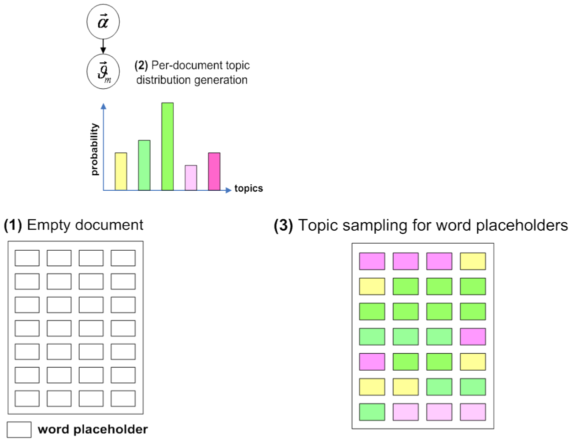
Generative Process (2)



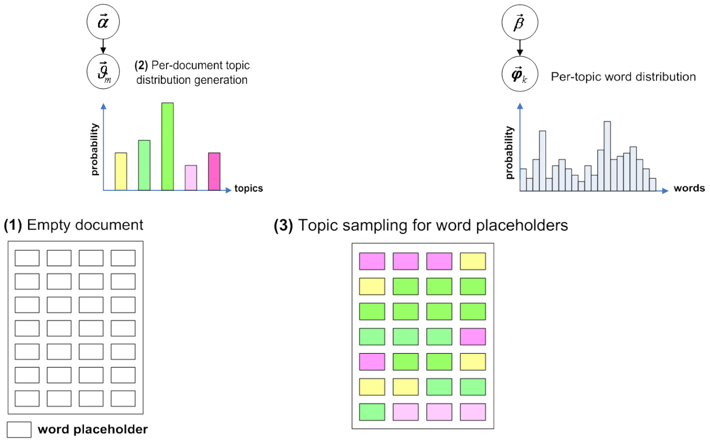
Generative Process (2)



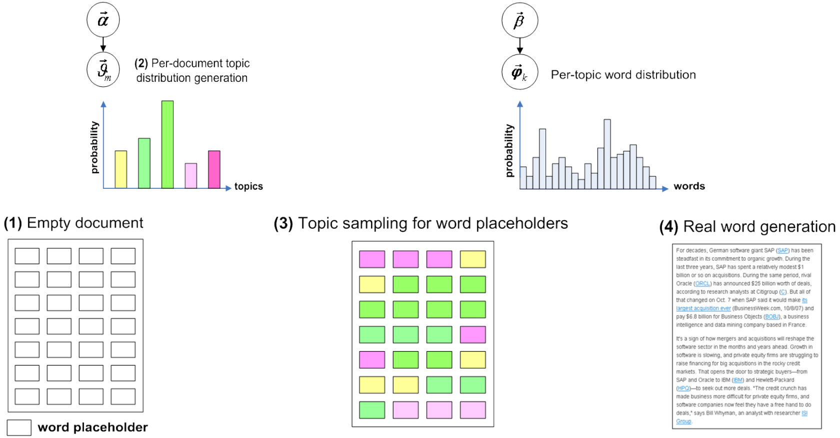
Generative Process (2)



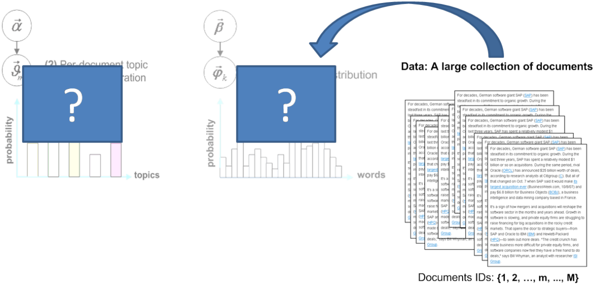
Generative Process (2)



Generative Process (2)



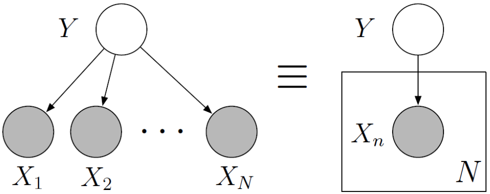
Statistical Inference: a Reverse Process



In reality, what we observe are only documents. Given these documents, our goal is to know what topic model is most likely to have generated the data:

- What are the words for each topic?
- What are the topics for each document?

Graphical Models Notation



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

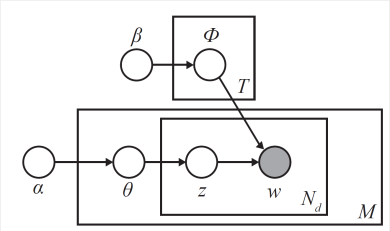
E.g, this graph is:

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

Notations

- Word: $1 \ldots V$
- Document: $w = (w_1, w_2, \ldots, w_{Nd})$ sequence of N words
- Corpus: $D = (w_1, w_2, \ldots, w_M)$ collection of M documents

LDA: Graphical Model

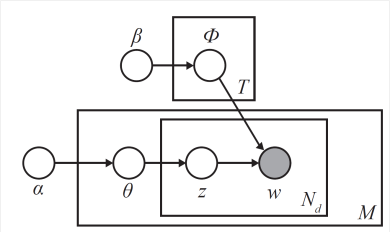


- α, β: Dirichlet prior
- M: number of doc
- N_d: number of words in d
- z: latent topic
- w: observed word
- θ: distribution of topic in doc
- ϕ: distribution of words generated from topic z

Using plate notation:

- Sampling of distribution over topics for each document d
- Sampling of word distributions for each topic z until T topics have been generated

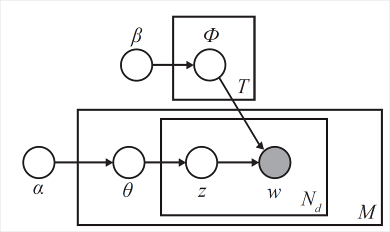
LDA: Graphical Model



Key Problem

Compute posterior distribution of the hidden variables given a document

Algorithm for Extracting Topics



- How to estimate posterior distribution of hidden variables given a collection of documents?
  - Direct: e.g., via expectation-maximization (EM) [Hofmann, 1999]
  - Indirect: estimate the posterior distribution over z. E.g., Gibbs Sampling [Griffiths & Steyvers, 2004]

Gibbs Sampling for LDA

- Random start
- Iterative
- For each word, we compute:
  - How dominate is a topic z in doc d? How often was topic z already used in doc d?
  - How likely is a word for a topic z? How often was the word w already assigned to topic z?

Gibbs Sampling for LDA

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

- Topic of each word will be sampled from this distribution
- #times word $w_i \Rightarrow$ topic j (except the current)
- total words $\Rightarrow$ topic k
- #words in doc $d \Rightarrow$ topic j (except the current)
- #words in doc m

Gibbs Sampling for LDA

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

- Topic of each word will be sampled from this distribution
- #times word $w_i \Rightarrow$ topic j (except the current)
- total words $\Rightarrow$ topic k
- #words in doc $d \Rightarrow$ topic j (except the current)
- #words in doc m

Gibbs Sampling for LDA

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

- Topic of each word will be sampled from this distribution
- #times word $w_i \Rightarrow$ topic j (except the current)
- total words $\Rightarrow$ topic k
- #words in doc $d \Rightarrow$ topic j (except the current)
- #words in doc m

Gibbs Sampling for LDA

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

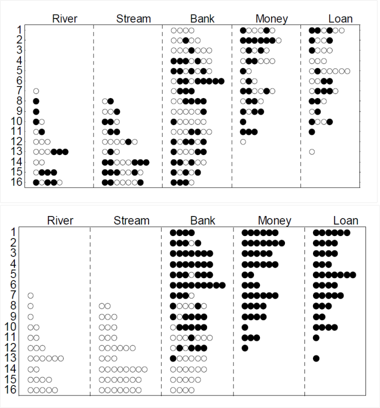
- Topic of each word will be sampled from this distribution
- #times word $w_i \Rightarrow$ topic j (except the current)
- total words $\Rightarrow$ topic k
- #words in doc $d \Rightarrow$ topic j (except the current)
- #words in doc m

Gibbs Sampling for LDA

$$P(z_i = j | z_{-i}, w_i, d_i, \cdot) \propto \frac{C^{WT}_{w_i j} + \beta}{\sum_{w=1}^{W} C^{WT}_{wj} + W\beta} \frac{C^{DT}_{d_i j} + \alpha}{\sum_{t=1}^{T} C^{DT}_{d_i t} + T\alpha}$$

- Topic of each word will be sampled from this distribution
- #times word w_i $\Rightarrow$ topic j (except the current)
- total words $\Rightarrow$ topic k
- #words in doc d $\Rightarrow$ topic j (except the current)
- #words in doc m

Gibbs Sampling Convergence



- Random Start
- N iterations
- Each iteration updates count-matrices

Convergence:

- count-matrices stop changing

Estimating θ and ϕ

$$\phi_i'^{(j)} = \frac{C_{ij}^{WT} + \beta}{\sum_{k=1}^{W} C_{kj}^{WT} + W\beta}$$

$$\theta_j'^{(d)} = \frac{C_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} C_{dk}^{DT} + T\alpha}$$

Short & Sparse Text Segments

- The explosion of
  - e-commerce
  - online communication, and
  - online publishing
- Typical examples
  - Web search snippets
  - Forum & chat messages
  - Blog and news feeds/summaries
  - Book & movie summaries
  - Product descriptions
  - Customer reviews
  - Short descriptions of entities, such as people, company, hotel, etc.

Challenges

- Very short
  - From a dozen of words to several sentences
  - Noisier
  - Less topic-focused
- Sparse
  - Not enough common words or shared context among them
- Consequences
  - Difficult in similarity measure
  - Hard to classify and clustering correctly

Synonym & Polysemy with Topics

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

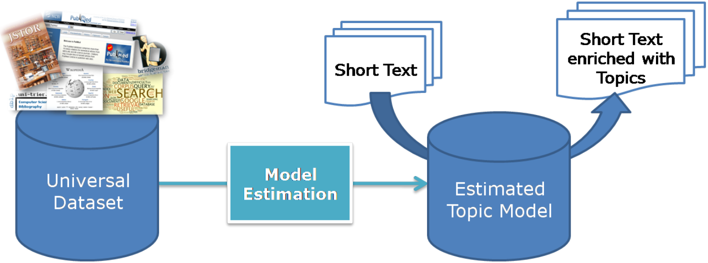
Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

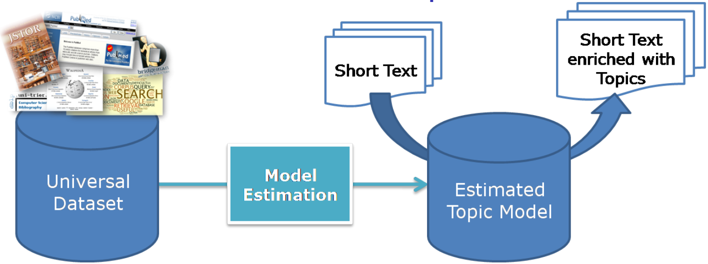
word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Short Text Enrichment with Topic Models



- Take advantage of available large collections, learn a topic model
- Use this model to analyze topics for short text documents
- Enrich short text documents with topics that have high probability

Short Text Enrichment with Topic Models



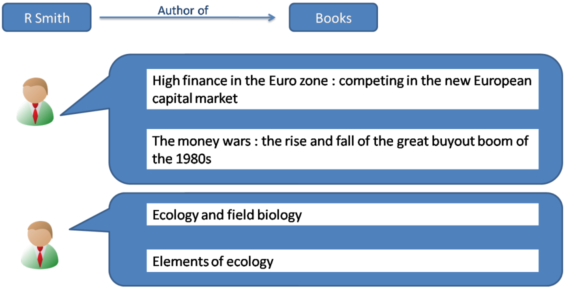
- Deal with problems of sparse and short texts: word choice, synonym, polysemy
- Increase the co-occurrence phenomenon among them
- Expand and enrich the shared context of data
- General and flexible: can be applied for different tasks, domains, languages

Applications

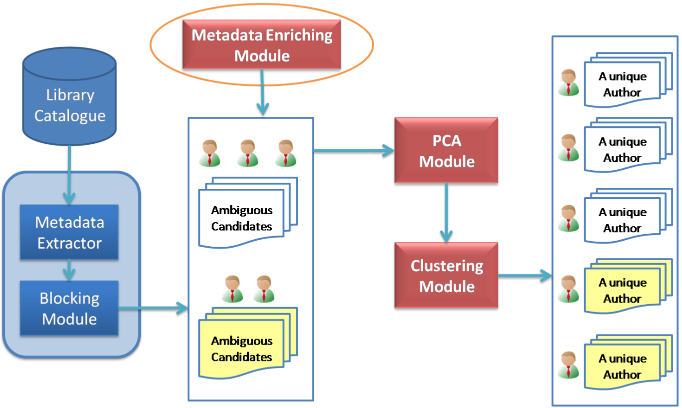
- **Author Name Disambiguation**
  Enrich books' titles, scientific/general domain, in English
- **Online Contextual Advertising**
  Enrich webpages and advertisements, general domain, in Vietnamese
- **Query Classification**
  Enrich queries, art domain, in English

Author Name Disambiguation

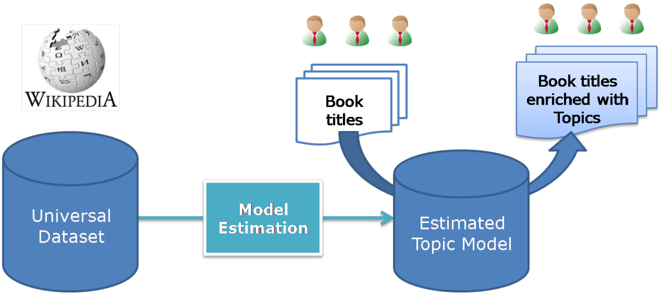
- Ambiguous author name: Different authors having the same name
- Author Name Disambiguation: a crucial service in catalogue searching & data integration



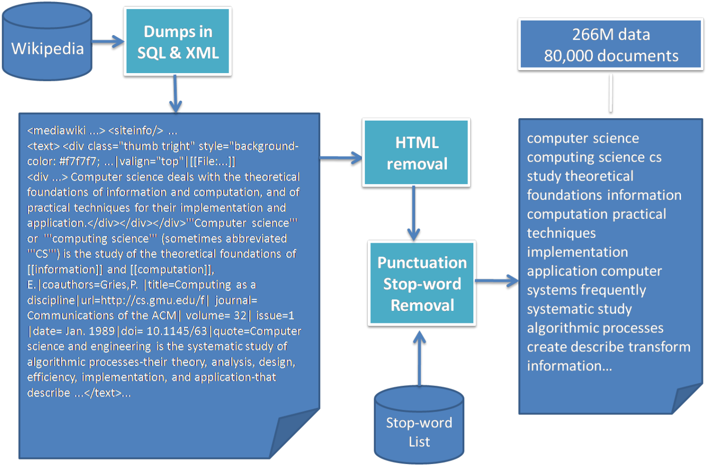
Author Name Disambiguation: A Framework



Metadata enriching module with Topics



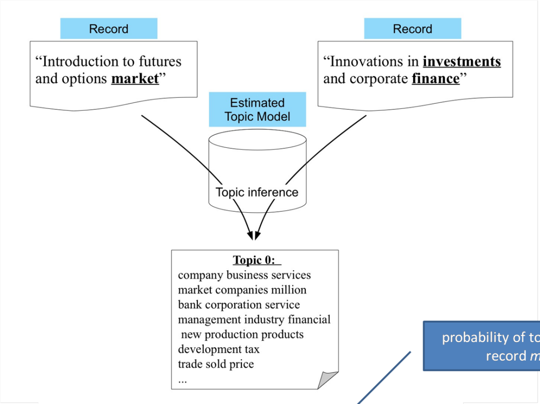
Wikipedia Preprocessing



Sample topics extracted from the estimated model

Topic 0	Topic 8	Topic 23	Topic 39	Topic 68	Topic 86	Topic 96
company	album	cells	law	storm	war	school
business	music	disease	court	tropical	army	university
services	band	medical	police	damage	force	college
market	song	patients	legal	winds	battle	high
companies	released	treatment	rights	typhoon	military	students
million	singer	cell	public	cyclone	air	schools
bank	rock	blood	justice	storms	navy	education
service	guitar	health	laws	caused	ship	institute
industry	live	medicine	judge	landfall	command	year
financial	records	brain	criminal	season	attack	program
tax	vocals	protein	supreme	pacific	fire	campus

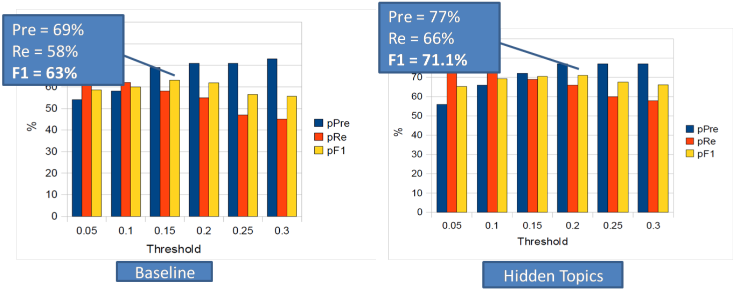
Hidden Topic Inference for Metadata



probability of topic k in record m

$$Frequency_{m,k} = \begin{cases} round\left(scale \times \vartheta_{m,k}\right), \text{ if } \vartheta_{m,k} \geq cut\text{-}off \\ 0, \text{ if } \vartheta_{m,k} < cut\text{-}off \end{cases}$$

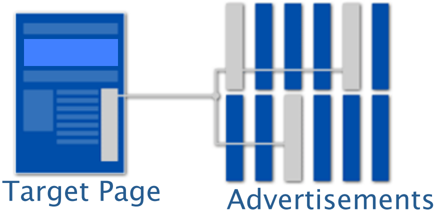
Results


Pre = 69%
Re = 58%
F1 = 63%
%
Threshold
0.05
0.1
0.15
0.2
0.25
0.3
pPre
pRe
pF1
Baseline
Pre = 77%
Re = 66%
F1 = 71.1%
Hidden Topics

Online Contextual Advertising

A solution for "reaching the **right person** with the **right message** at the **right time**".

Contextual Matching & Ranking

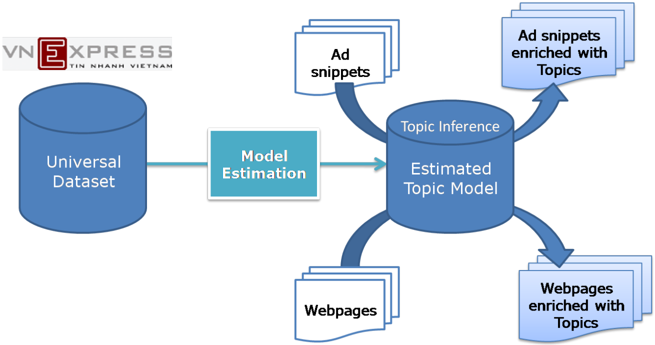


- A set of Web pages $P = p_1, p_2, , p_n$
- A set of ads: $A = \{a_1, a_2, , a_m\}$

Matching & Ranking:

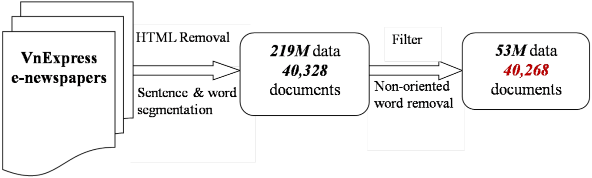
- For each $p \in P$ (p is called "target page")
- Match & rank all ads in A w.r.t p such that k-top ads $A* = \{a_{p1}, , a_{pk}\} \subset A$ are most relevant to the content of p

Webpage & Advertisement Enriching with Topics



Topic Analysis of Large News Collections

http://vnexpress.net



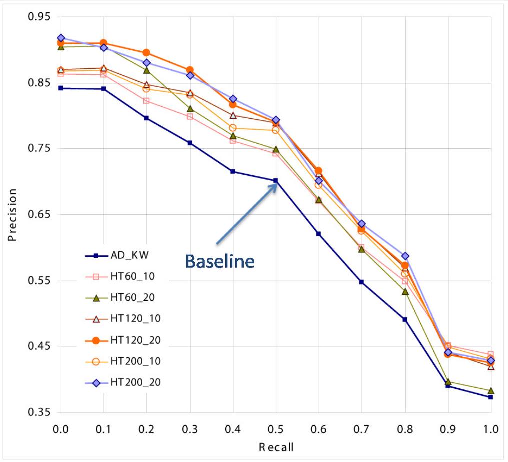
Using Latent Dirichlet Allocation (LDA) [Blei et al. 2003] & Gibbs Sampling [Griffiths & Steyvers 2004]

Sample topics extracted from the estimated model

Topic 1	Topic 3	Topic 15	Topic 44
phòng (room)	**bác_sĩ** (doctor)	**thời_trang** (fashion)	**thiết_bị** (equipment)
không_gian (space)	**bệnh_viện** (hospital)	**người_mẫu** (model)	**sản_phẩm** (product)
thiết_kế (design)	**thuốc** (medicine)	**mặc** (wear)	**máy** (machine)
ngôi_nhà (house)	**bệnh** (disease)	**trang_phục** (clothes)	**màn_hình** (screen)
tầng (floor)	**phẫu_thuật** (surgery)	**thiết_kế** (design)	**công_nghệ** (technology)
trang_trí (decorate)	**điều_trị** (treatment)	**đẹp** (beautiful)	**điện_thoại** (telephone)
nội_thất (interior)	**bệnh_nhân** (patient)	**váy** (dress)	**hãng** (company)
tường (wall)	**y_tế** (medical)	**sưu_tập** (collection)	**sử_dụng** (use)
ánh_sáng (light)	**ung_thư** (cancer)	**mang** (wear)	**thị_trường** (market)
đèn (lamp)	**tình_trạng** (condition)	**phong_cách** (style)	**usd** (USD)
phòng_ngủ (bedroom)	**cơ_thể** (body)	**quần_áo** (costume)	**pin** (battery)
rộng (wide)	**sức_khoẻ** (health)	**nổi_tiếng** (famous)	**cho_phép** (allow)
bố_trí (arrange)	**đau** (hurt)	**quần** (trousers)	**samsung** (samsung)
vườn (garden)	**gây** (cause)	**trình_diễn** (perform)	**di_động** (mobile)
kính (glass)	**khám** (health check)	**thích** (like)	**sony** (sony)
cảm_giác (feel)	**kết_quả** (result)	**quyến_rũ** (charming)	**nhạc** (music)
diện_tích (square)	**căn_bệnh** (illness)	**sang_trọng** (luxurious)	**máy_tính** (computer)
căn_phòng (apartment)	**nặng** (serious)	**vẻ_đẹp** (beauty)	**hỗ_trợ** (support)
khu (area)	**cho_biết** (inform)	**gái** (girl)	**điện_tử** (electronic)
hiện_đại (modern)	**máu** (blood)	**gương_mặt** (figure)	**tính_năng** (feature)

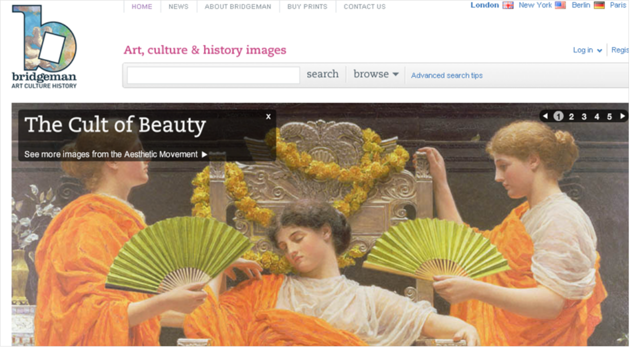
Full results at `http://gibbslda.sourceforge.net/vnexpress-200topics.txt`

Result

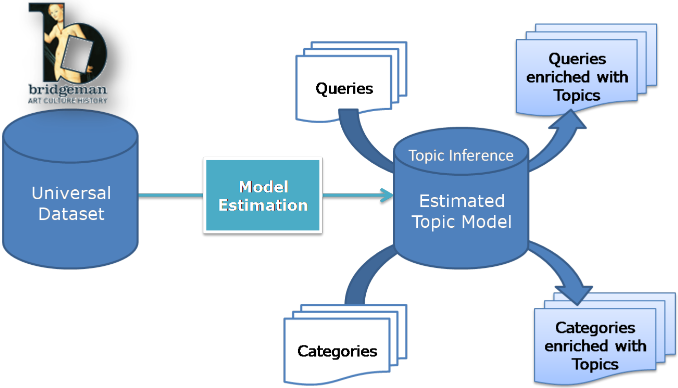

0.95
0.85
0.75
0.65
0.55
0.45
0.35
Precision
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
AD_KW
HT60_10
HT60_20
HT120_10
HT120_20
HT200_10
HT200_20
Baseline

Query Classification Task

- Classifying queries to a target taxonomy
- Domain: Art, Culture & History images



Query enriching with Topics



Result

Setting	Hits				$\%_{Top\_3}$
	# 1	# 2	# 3	$\sum_{Top\_3}$	
Baseline 1	13	17	5	33	**60%**
Baseline 2	15	14	7	35	63.6%
TM 1	14	15	5	32	58.2%
TM 2a	22	14	6	40	72.7%
TM 2b	31	9	6	44	**80%**

Table: Results of Query Classification: with Click Through Information

Conclusions

- Topic Models can be useful tools for statistical analysis of document collections
- These models make explicit assumptions about the process responsible for generating a document
- Topic Models estimated from large corpora can be exploited to deal with the problem of short and sparse text, experimented in different tasks with promising results

Bibliography

D.M. Blei and J.D. Lafferty, *A correlated topic model of science*, The Annals of Applied Statistics **1** (2007), no. 1, 17–35.

D.M. Blei, A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation*, The Journal of Machine Learning Research **3** (2003), 993–1022.

T.L. Griffiths and M. Steyvers, *Finding scientific topics*, Proceedings of the National Academy of Sciences of the United States of America **101** (2004), no. Suppl 1, 5228.

D.T. Le, C.T. Nguyen, Q.T. Ha, X.H. Phan, and S. Horiguchi, *Matching and ranking with hidden topics towards online contextual advertising*, 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, IEEE, 2008, pp. 888–891.

X. Phan, C. Nguyen, D. Le, L. Nguyen, S. Horiguchi, Q. Ha, E. Iosif, A. Potamianos, P. Velardi, A. Cucchiarelli, et al., *A Hidden Topic-Based Framework Towards Building Applications with Short Web Documents*, Knowledge and Data Engineering, IEEE Transactions on, 1–1.

Dieu-Thu Le Raffaella Bernardi, *Metadata enrichment via topic models for author name disambiguation*, Advanced Language Technologies for Digital Libraries, Hot Topic series, Springer (2011).