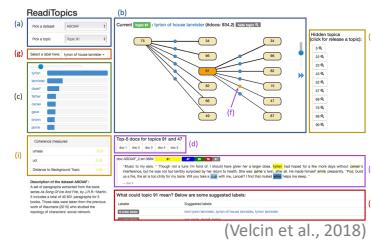


Topic learning

Julien Velcin
Master MALIA-MIASHS
2024-2025



Outline

- Why topic learning
- Topic learning with matrix factorization
- Probabilistic graphical models
- Latent Dirichlet Allocation
- Illustration on several case studies
- More graphical models

(part 3/4)

2

Foreword

- Several slides are directly taken from the talk given by David M. Blei for KDD (2011)
<https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf>
- Some of them related to graphical model come from the HP team (2006)
http://home.in.tum.de/~xiao/h/pub/3G_talk.pdf
- The derivation for the Gibbs sampling is taken from the technical report of Y. Wang (2008)
<https://cxwangyi.files.wordpress.com/2012/01/lit.pdf>

Outline

- Why topic learning
- Topic learning with matrix factorization
- Probabilistic graphical models
- Latent Dirichlet Allocation
- Illustration on several case studies
- More graphical models

Siginin Subscribe Search

the guardian

UK world sport football opinion culture business lifestyle fashion environment tech travel

headlines Thursday 28 January 2016

Zika virus spreading 'explosively', says World Health Organisation

Director general convenes emergency committee saying it is deeply concerning virus linked to birth defects has now been detected in more than 20 countries

Now 14°C Lyon

17.00 20.00 23.00 02.00 8°C 8°C 8°C

Brazil Recife, city at centre of Zika epidemic Video What you need to know Should I cancel my holiday? Latest advice for travellers

Apples, berries, peppers: Natural compound in fruit and veg could help prevent weight gain - study

US Marco Rubio, from 'Republican saviour' to prophet of gloom ... and back again

4 December 2015

THE HUFFINGTON POST

UNITED KINGDOM

Edition: UK Search The Huffington Post

FRONT PAGE NEWS POLITICS BUSINESS TECH YOUNG VOICES COMEDY ENTERTAINMENT CELEBRITY LIFESTYLE PARENTS BLOGS

Politics • COP21 • Building Modern Men • What's Working • Environment • Media • Women • Impact • Entrepreneurs • Young Talent • Christmas • Smart Living

5

Accueil Notifications Messages #malavita

Top Direct Comptes Photos Vidéos Autres options 424k

Suggestions Actualiser Tout afficher

- Kaplan International @K... Suivre Sponsorié
- Aras BOZKURT @arasbozkurt Suivre
- Stéphane Pouilly @spouyl... Suivre

Trouver des amis

Tendances Modifier

- #Enviroiture
- Sponsorié par Allianz France
- #ASSEPSG
- #JacquelineSauvage
- #Camping
- #TheVoice
- #Malavita
- Milan
- Bordeaux
- Ronaldo
- Florian Thauvin
- Benoit Violier

11 nouveaux résultats

- Marion @Marion_LeJct · 2 min Trop cool ce film de LucBesson #Malavita #TF1
- Gabi 🌟 @11gabi_01 · 2 min Et putain... 🎬 #Malavita
- Mouna Camara @mouna_camara · 2 min Tres bon film #Malavita
- Stephanie L@SLidouren · 2 min Après #Malavita place à #LOLUSA
- Black Mamba @SmallHawkeye · 2 min Bon ce film était bof. Rien d'exceptionnel, des petits passages marrant. Dommage avec un tel casting... #Malavita
- Ree @HirtRee · 2 min Film d'action américain en normandie 😎 #Malavita

7

Google News

Google

News U.K. edition

Top Stories

- Kolkata
- Jacob Zuma
- Nuclear Security Summit
- Manchester United F.C.
- FC Barcelona
- Arsenal F.C.
- Refugees
- Google
- Apple
- Quantum Break
- Lyon, Rhône-Alpes, F...

See realtime coverage

Suggested for you

- World
- U.K.
- Business
- Technology
- Entertainment
- Sports
- Science
- Health

Top Stories

- Cameron defends blocking steel tariffs as Javid faces workers' anger
- Denmark PM's tough stance criticised by international media but has popular support at home
- Immigration Sweden sends sharp signal with plan to expel up to 80,000 asylum seekers
- US Marco Rubio, from 'Republican saviour' to prophet of gloom ... and back again
- Brazil Recife, city at centre of Zika epidemic Video What you need to know Should I cancel my holiday? Latest advice for travellers
- Apples, berries, peppers: Natural compound in fruit and veg could help prevent weight gain - study
- US Marco Rubio, from 'Republican saviour' to prophet of gloom ... and back again
- British amateur sailor dies in Clipper Round The World Yacht Race tragedy
- Isil offered British soldier details for fanatics to attack here as delivery driver guilty of plot to kill US troops
- Dilma Rousseff, Brazil's warrior president

The Guardian - 19 minutes ago David Cameron defended Britain's decision to reject higher EU tariffs on Chinese steel yesterday as the business secretary faced the anger of Port Talbot workers whose livelihoods have been undermined by cut-price imports.

Financial Times There are ways to help the British steel industry, but a government bail-out isn't the solution Telegraph.co.uk Opinion: Fitch downgrades Tata Steel and UK arm Business Standard

YouTube Financial T... Daily Mail The Indep... Daily Mail Mirror.co.uk WalesOnline Mirror

6

Comparing news media

- Joint work with a sociologist (J.C. Soulages, Max Weber lab), in a project close to data journalism (Velcin et al., workshop @EGC 2017)
- Usual preprocessing (tokenisation, stopwords...)

Version	langue	#articles	longueur	#mots
US	anglais	12 067	454.4	5 482 661
FR	français	4 133	369.6	1 527 416
BR	portugais	2 355	429.5	1 011 373

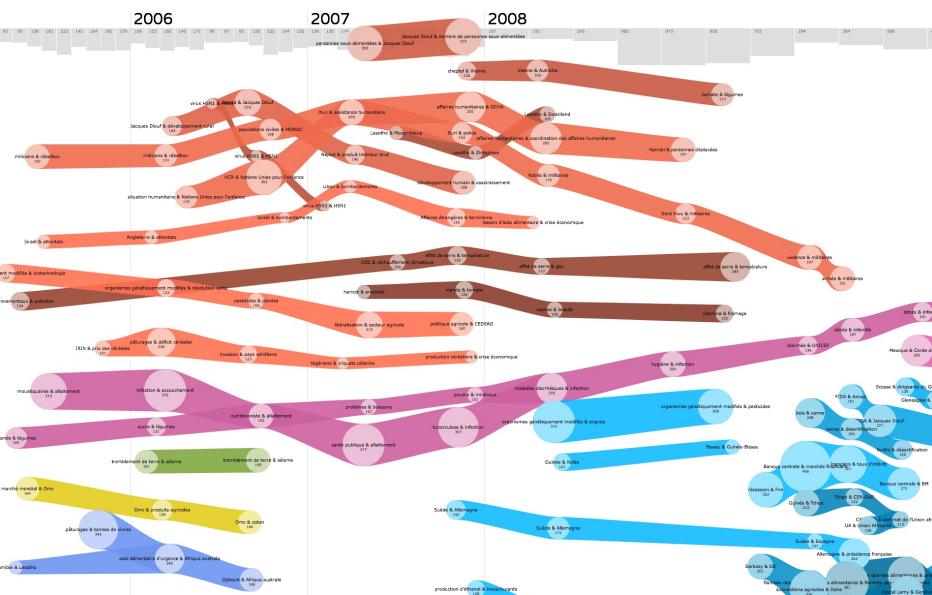
- How to compare those three versions by using LDA?
- Attribute one category for each topic manually (e.g., sport or media)
- Estimate the importance of each category

Some topics extracted with LDA

topic	#doc	cat.	mots les plus probables
en français (sur 4133 articles) :			
z18	28	1	manifestation, paris, police, travail, loi, contre, syndicats, place, bastille, 2016
z19	36	1	loi, travail, gouvernement, l'état, texte, l'assemblée, d'urgence, mois, projet, conseil
z25	39	2	jeux, rio, olympiques, olympique, août, jo, athlètes, 2016, brésil, cérémonie
z47	18	3	morandini, jean-marc, inrocks, catherine, l'animateur, lui, qu'il, europe, comédiens, plainte
z73	47	4	nice, 14, l'attentat, anglais, promenade, camion, attentat, police, soir, christian
en anglais (sur 12067 articles) :			
z14	92	5	refugees, children, refugee, people, countries, world, syrian, rights, million, year
z21	74	2	gymnastics, biles, olympic, team, simone, olympics, gymnast, gold, rio, hernandez
z3	46	6	pokemon, game, pokémon, playing, players, catch, «pokemon, go», pizza, play
z50	56	7	muslim, religious, muslims, faith, church, god, christian, religion, hate, american
z27	140	8	clinton, voters, trump, poll, polls, americans, election, support, vote, relationships
en portugais (sur 2355 articles) :			
z44	52	8	dilma, presidente, impeachment, senado, senadores, processo, senador, rousseff, julgamento, defesa
z58	7	9	sexo, menstruação, durante, rao, mcccane, comédia, realmente, corpo, riso, menstruada
z71	11	7	negros, brancos, negras, pessoas, racial, negra, racismo, país, movimento, black
z37	57	2	brasil, vôlei, jogo, medalha, vitória, ouro, seleção, set, brasileiras, torcida
z99	20	7	lgbt, gay, preconceito, violência, sexual, direitos, família, orgulho, estupro, aborto

Les catégories attribuées ici (cat.) correspondent à : 1- Economie / Social, 2- Sport / JO, 3- Show business / people, 4- Sécurité / attentats, 5- Politique étrangère, 6- Technologie / science, 7- Société, 8- Politique nationale, 9- Santé.

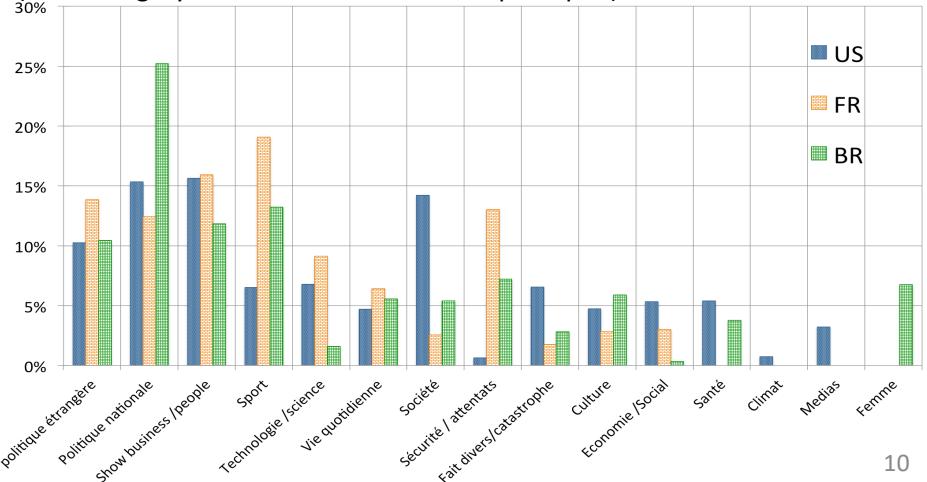
9



Pulseweb project: <http://pulseweb.context.net>

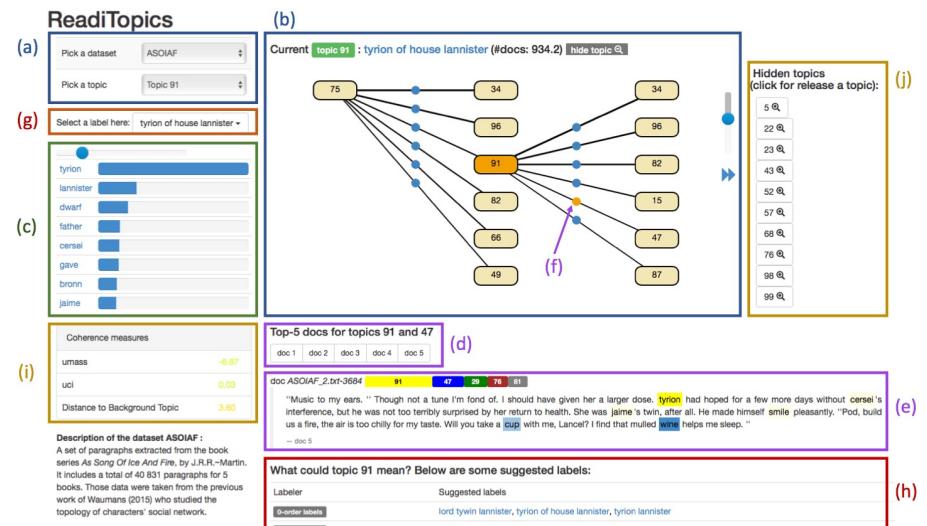
Compared results

Normalized distribution of 15 categories
(each category can be associated to multiple topics)



10

Readitopics (Velcin et al., 2018)



12

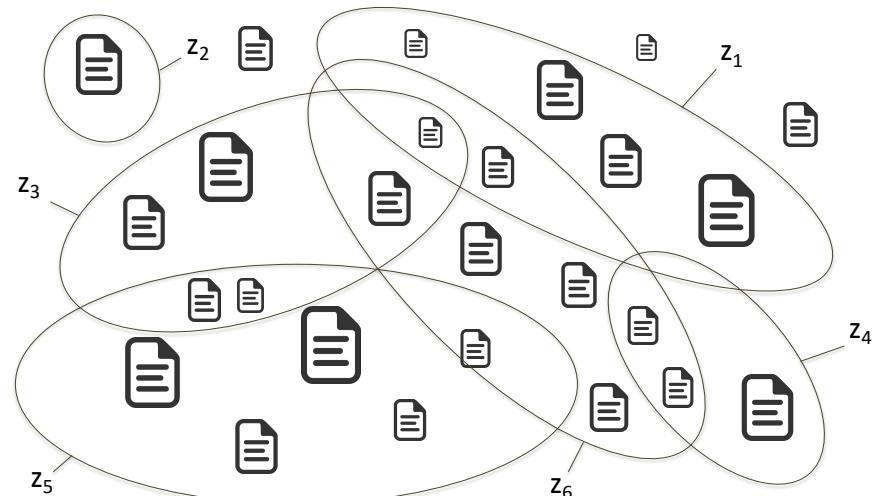
Why topic learning?

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

- discover the **hidden themes** that pervade the collection
- annotate the documents according to those themes
- use annotations to organize, summarize, and search the texts

13

Difference with clustering (Xie and Xing, 2013)



14

Another interesting view

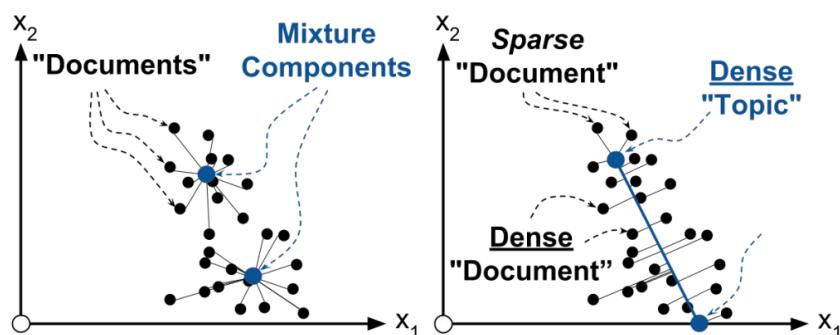


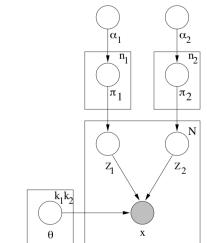
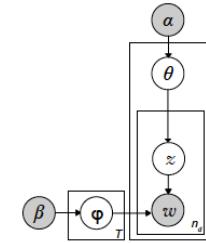
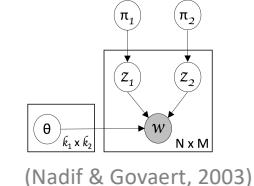
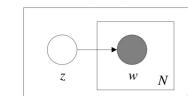
Figure 1: (Left) In mixtures, documents are drawn from exactly one component distribution. (Right) In admixtures, documents are drawn from a distribution whose parameters are a convex combination of component parameters.

(taken from: <http://bigdata.ices.utexas.edu/project/topic-models-with-word-dependencies>)

15

Different probabilistic models

Remember the simple mixture model:



16

Various approaches for topic learning

- Geometrical approaches
 - TDT (Allan et al., 1998) (Pons-Porrata et al., 2003)
 - AGAPE (Velcin and Ganascia, 2007)
- Algebraic approaches
 - LSA (Deerwester et al., 1990)
 - NMF (Paatero et Tapper, 1994)
 - Dictionary learning (Jenatton et al., 2010)
- Neural approaches
 - AVITM (Srivastava and Sutton, 2017)
 - and many others... (BerTopics, topicGPT...)
- Probabilistic approaches
 - pLSA, LDA... (see the following)

17

Some background on text mining

- Bag-of-words assumption
- Usual preprocessing
 - removing numbers and punctuation
 - removing stopwords
- Classic input:

Terms	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
data	1	1	0	0	2	0	0	0	0	1	2	1	1	1	0	1	0	0	0	
examples	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
introduction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
mining	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	
network	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	
package	0	0	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	



18

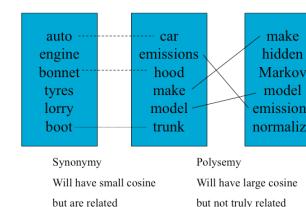
Outline

- Why topic learning
- **Topic learning with matrix factorization**
- Probabilistic graphical models
- Latent Dirichlet Allocation
- Illustration on several case studies
- More graphical models

19

Latent Semantic Analysis / Indexing

- LSI (LSA) is a statistical technique that attempts to estimate the hidden content structure within documents
- Projects queries and documents into a space with “latent” semantic dimensions
- Overcoming the problem of polysemy and synonymy



20

Latent Semantic Analysis

- Terms co-occurring are projected onto the same dimensions
- In this new space, queries and documents can be related even if they don't share any common word
- Closely related to PCA:
 - Singular Value Decomposition (SVD)
 - Keeping only the **k** highest singular values

21

Singular Value Decomposition (SVD)

- SVD decomposes the Term x Document matrix X as $X = T.S.D^T$, such as:
 - T = left singular vector matrix
 - D = right singular vector matrix
 - S = diagonal matrix of singular values
- Similar to the eigenvector-eigenvalue (spectral) decomposition $Y = V.L.V^T$:
 - T : matrix of eigenvectors of $Y=X.X^T$
 - D : matrix of eigenvectors of $Y=X^T.X$
 - S^2 : diagonal matrix L of eigenvalues

22

Latent Semantic Analysis (LSA)

$$\begin{array}{c} \text{documents} \\ | \\ \text{terms} \end{array} X = \begin{array}{c} T \\ 0 \\ | \\ m \times m \end{array} \quad \begin{array}{c} S \\ 0 \\ | \\ m \times d \end{array}$$

$$X = \begin{array}{c} T \\ 0 \\ | \\ k \times k \end{array} \quad \begin{array}{c} S \\ 0 \\ | \\ k \times d \end{array}$$

Reducing the matrix S
by setting SV to 0

$$X = T S D^T$$

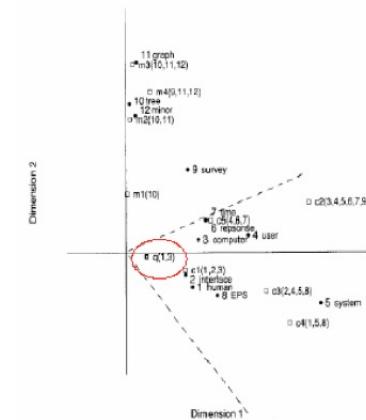
23

Classic LSI Example [Deerwester et al., 90]

Technical Memo Example

Titles:	Documents								
	c1	c2	c3	c4	c5	m1	m2	m3	m4
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

2-D Plot of Terms and Docs from Example



24

Computing Similarity in LSA

- Based on X :
 - Pairwise **term** comparison: $X \cdot X^T$
 - Pairwise **doc** comparison: $X^T \cdot X$
- Based on \hat{X} :
 - Pairwise **term** comparison: $\hat{X} \cdot \hat{X}^T = (TS) \cdot (TS)^T$
 - Pairwise **doc** comparison: $\hat{X}^T \cdot \hat{X} = (DS) \cdot (DS)^T$
- Comparing a query and a doc: ad-hoc computation based on \hat{X}

25

Some issues on LSI/LSA

- SVD assumes normally distributed data
- Time complexity in $O(n^2k^3)$
 - n = number of terms
 - k = number of “topics”
- Finding the optimal k is still an open issue (just as for KMeans and LDA-based models)
- Has proved to be a valuable tool in many areas of NLP as well as IR

26

Some issues on LSI/LSA (con't)

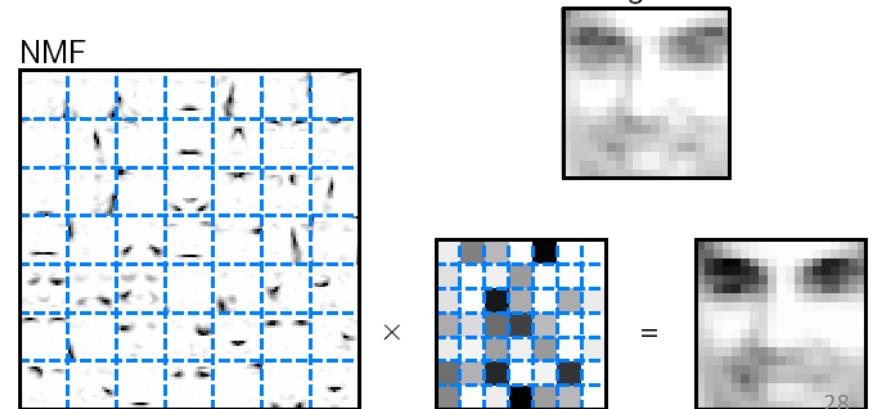
- LSA or PCA involves adding up some basis images then subtracting others
- Basis images (vectors) are not physically intuitive
- Subtracting doesn't make sense in context of some applications

27

Non-negative Matrix Factorization

[Lee et al., 1999, 2001]

- Like PCA, except the coefficients in the linear combination cannot be negative



PCA vs NMF

- PCA
 - Designed for producing optimal (in some sense) basis images
 - Just because it's optimal doesn't mean it's good for your application
- NMF
 - Designed for producing coefficients with a specific property
 - Forcing coefficients to behave induces "nice" basis images
 - No SI unit for "nice"

29

Objective function

- NMF = factorization into 2 non-negative matrices:
$$\mathbf{X} \approx \mathbf{WH}$$
 with $\mathbf{W} \geq 0, \mathbf{H} \geq 0$
- If we use the Euclidean distance, the objective function is the following: $\|\mathbf{X} - \mathbf{WH}\|_F^2$
- Note that we can use other differences, such as the Kullback-Leibler (KL) divergence or the Itakura-Saito divergence, see (Févotte and Idier, 2011)
- Using gradient descent, we can calculate alternating updating rules

Févotte, C., & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β -divergence. *Neural computation*, 23(9), 2421-2456.

30

Sketch of the derivation

- First, let's set \mathbf{W} constant and calculate the gradient on \mathbf{H}

$$D(\mathbf{X}, \mathbf{WH}) = \sum_m \sum_n (x_{mn} - \mathbf{WH}|_{mn})^2$$

$$\nabla_H D(\mathbf{X}, \mathbf{WH}) = \nabla_H \text{tr}[(\mathbf{X} - \mathbf{WH})^T (\mathbf{X} - \mathbf{WH})]$$

$$\nabla_H D(\mathbf{X}, \mathbf{WH}) = -2\mathbf{W}^T \mathbf{X} + 2\mathbf{W}^T \mathbf{WH}$$

- Then use the gradient to minimize the cost function

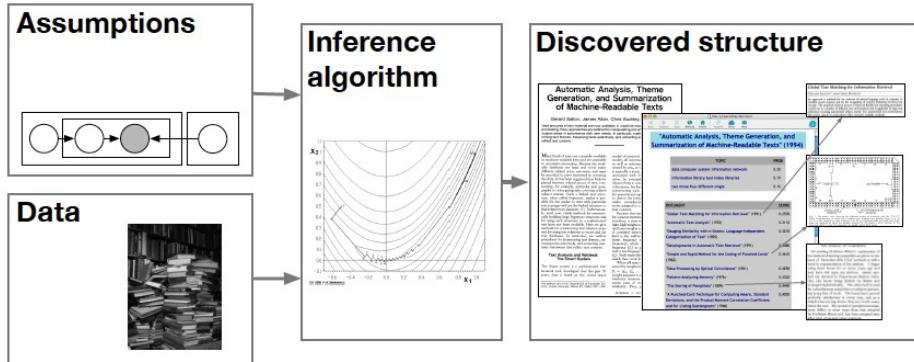
$$\mathbf{W} \leftarrow \mathbf{W} - \eta_H \nabla_H D(\mathbf{X}, \mathbf{WH})$$

Outline

- Why topic learning
- Topic learning with matrix factorization
- **Probabilistic graphical models**
- Latent Dirichlet Allocation
- Illustration on several case studies
- More graphical models

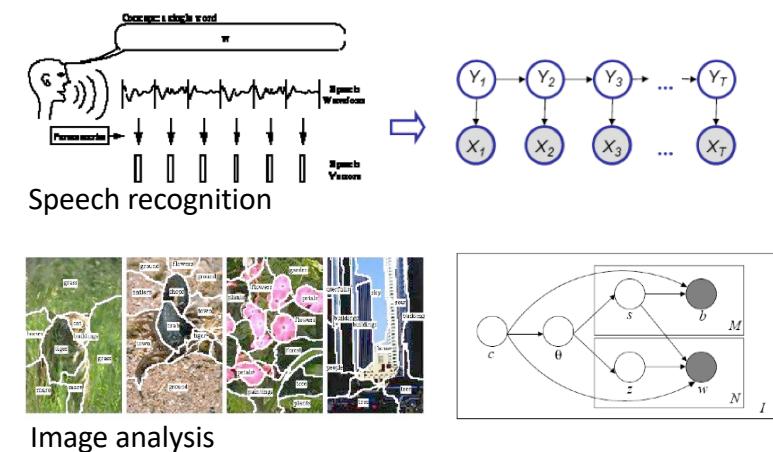
Probabilistic graphical models

« If you remember one picture »:



33

Examples given by A. McCallum



34

And for topic learning

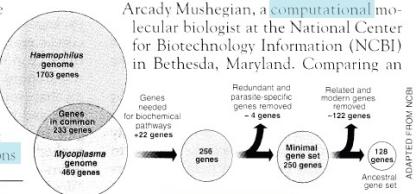
Seeking Life's Bare (Genetic) Necessities

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

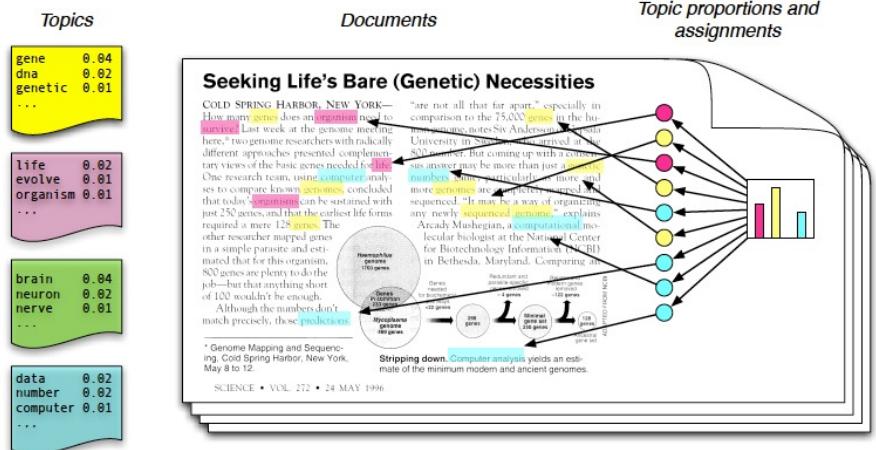
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes 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



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

Discovering latent structures



36

Image				
Ground Truth	field, foals, horses, mare	beach, horizon, people, water	waved, albatross, flight, sky	coast, sky, water, waved
PLSA-WORDS Annotation	grass, foals, horses, garden, trees	water, trees, beach, flowers, garden	city, flight, ceremony, pond, swallow-tailed	trees, sky, snow, clouds, coast
GM-PLSA Annotation	horses, foals, mare, field, grass	beach, water, sky, trees, horizon	albatross, sky, flight, bird, waved	sky, coast, water, clouds, waved

continuous pLSA (Li et al., 2010)

Topic learning goes beyond words!



True caption
birds tree
Corr-LDA
birds nest leaves branch tree
GM-LDA
water birds nest tree sky
GM-Mixture
tree ocean fungus mushrooms coral



True caption
fish reefs water
Corr-LDA
fish water ocean tree coral
GM-LDA
water sky vegetables tree people
GM-Mixture
fungus mushrooms tree flowers leaves



True caption
mountain sky tree water
Corr-LDA
sky water tree mountain people
GM-LDA
sky tree water people buildings
GM-Mixture
buildings sky water tree people



True caption
clouds jet plane
Corr-LDA
sky plane jet mountain clouds
GM-LDA
sky water people tree clouds
GM-Mixture
sky plane jet clouds pattern

37

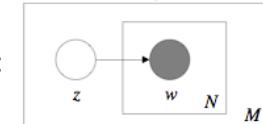
(Blei et al., 2003)

Graphical models

Definition given by Michael I. Jordan :

« It is a family of probability distributions defined in terms of a directed or undirected graph. The nodes in the graph are identified with random variables, and joint probability distributions are defined by taking products over functions defined on connected subsets of nodes. »

For instance:



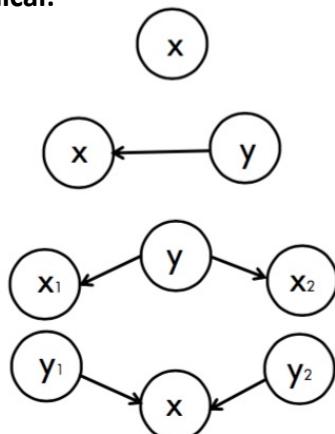
38

Algebraic vs. graphical

Algebraic :

- $p(x)$
- $p(x | y)$
- $p(x_1, x_2 | y)$
- $p(x | y_1, y_2)$

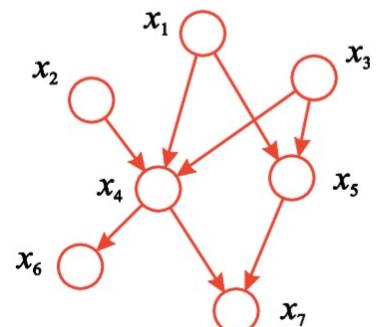
Graphical:



Joint distribution

$$p(x_1, \dots, x_D) = \prod_{i=1}^D p(x_i / pa_i)$$

where pa_i stands for the parents of node i



39

40

Benefits of graphical representation

- A graphical model gives all the (conditional) dependencies between variables.
- It describes a *generative* process.
- The joint probability is simplified by using the independency between variables:

$$p(x_1, \dots, x_D) = \prod_{i=1}^D p(x_i / pa_i)$$

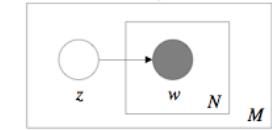
41

Example of a generative process

Really simple model = associate **one** topic to each document

For each document M:

- draw a topic z
 $z \sim Mult(p)$
- for each token N :
 - draw a word w given z
 $w \sim Mult_z(p)$



42

Outline

- Why topic learning
- Topic learning with matrix factorization
- Probabilistic graphical models
- **Latent Dirichlet Allocation**
- Illustration on several case studies
- More graphical models

43

Reminder

Data are assumed to be observed from a generative probabilistic process that includes **hidden** variables.

In text, the hidden variables are the thematic structure.

Infer the hidden structure using **posterior inference**

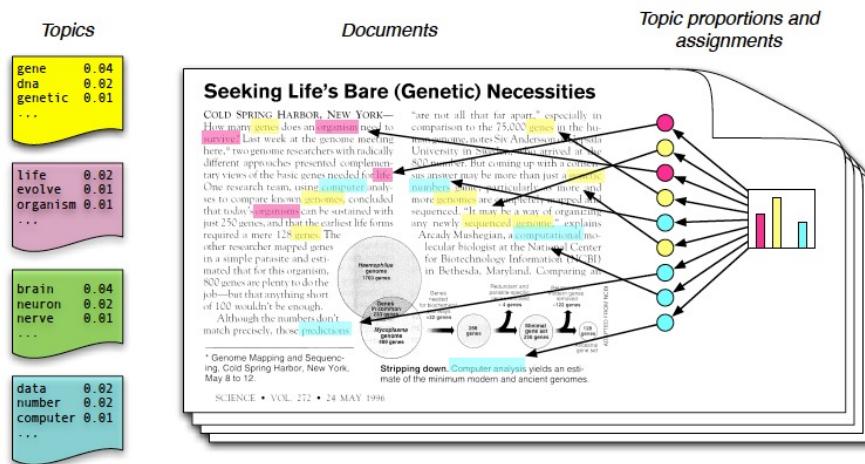
What are the topics that describe this collection?

Situate **new data** into the estimated model.

How does a new document fit into the topic structure?

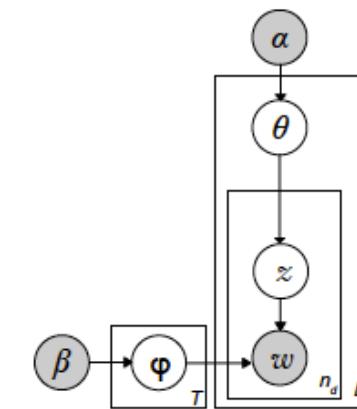
44

Generative model for LDA



45

LDA as a graphical model



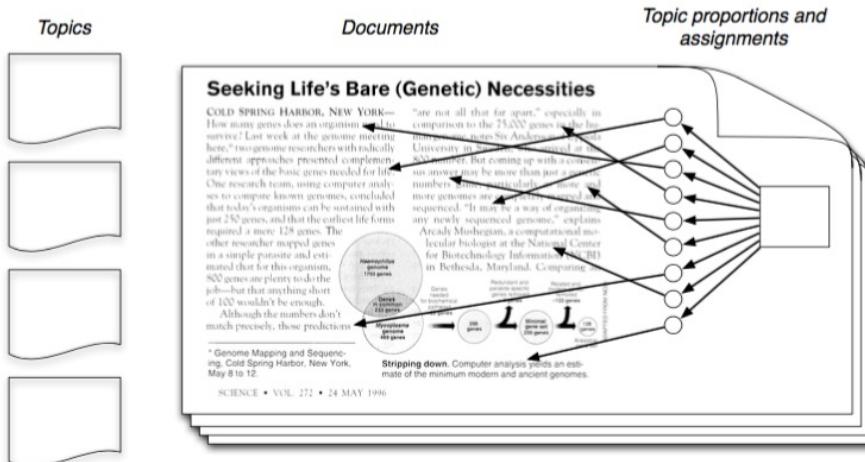
$$\phi : p(w_i/z_j)$$

$$\theta : p(z_j/d_m)$$

$$\alpha : \text{prior } p(\theta)$$

$$\beta : \text{prior } p(\phi)$$

The posterior distribution



47

Infering the latent variables

$$p(\theta, z | w, \alpha, \beta) = \frac{p(\theta, z, w / \alpha, \beta)}{p(w / \alpha, \beta)}$$

Intractable!

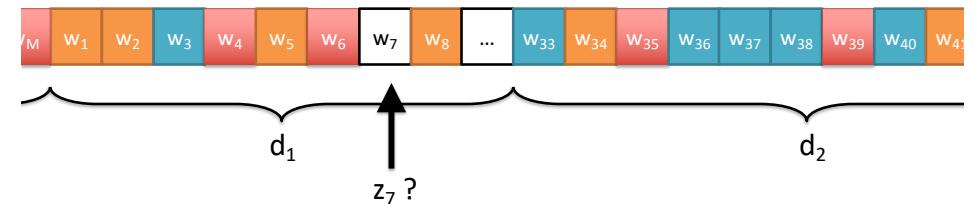
$$p(\mathbf{w} | \alpha, \beta) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \int \left(\prod_{i=1}^k \theta_i^{\alpha_i - 1} \right) \left(\prod_{n=1}^N \sum_{i=1}^k \prod_{j=1}^V (\theta_i \beta_{ij})^{w_n^j} \right) d\theta$$

48

Then how to compute the posterior?

- Variational inference
 - lower bounds the likelihood
- MCMC methods
 - (collapsed) Gibbs sampling
 - easier and faster to implement
- Expectation propagation...

Estimation with MCMC (1)



$$p(z_i | Z_{-i}, W, \alpha, \beta) \propto \frac{p(Z, W/\alpha, \beta)}{p(Z_{-i}, W_{-i}/\alpha, \beta)}$$

49

50

Estimation with MCMC (2)

- Joint distribution for z and w :

$$p(z, w | \alpha, \beta) = p(w | z, \beta) p(z | \alpha)$$

- First term:

$$p(w | z, \beta) = \int p(w | z, \phi) p(\phi | \beta) d\phi$$

$$\prod_{k=1}^K \prod_{v=1}^V \phi_{k,v}^{\psi_{k,v}}$$

$$\prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_{k,v}^{\beta_v - 1}$$

conjugacy

Estimation with MCMC (3)

$$p(w | z, \beta) = \int \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_{k,v}^{\psi_{k,v} + \beta_v - 1} d\phi_k$$

count of word v to topic k

prior on word v

Beta function

51

52

Estimation with MCMC (4)

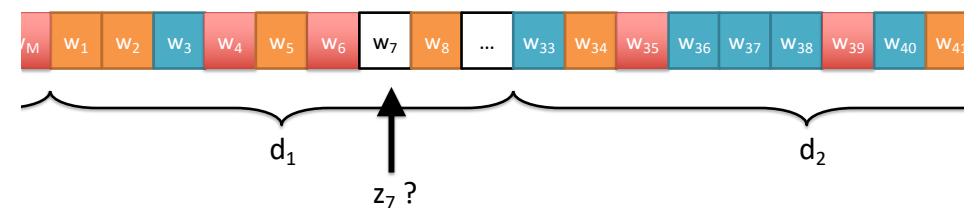
Now it is easy to show that the formula can be simplified as (Φ has been “integrated out”):

$$p(w/z, \beta) = \prod_{k=1}^K \frac{B(\psi_k + \beta)}{B(\beta)}$$

The same reasoning holds for $p(z/\alpha)$
So that we can calculate $p(z, w|\alpha, \beta)$
...and drives the Gibbs sampling procedure

53

Estimation with MCMC (5)



updating
rules

$$\phi_{k,v} = \frac{\psi_{k,v} + \beta_v}{\sum_{v'=1}^V \psi_{k,v'} + \beta_{v'}}$$

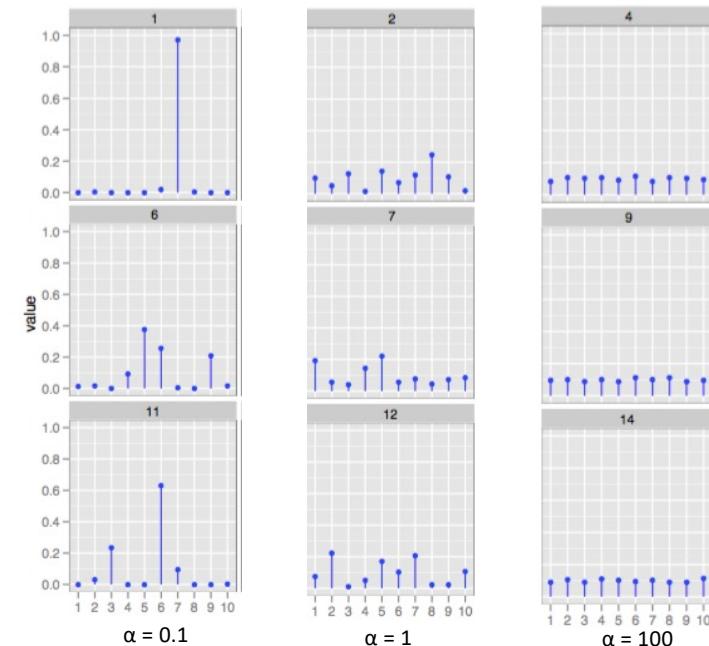
$$\theta_{d,k} = \frac{\Omega_{d,k} + \alpha_k}{\sum_{k'=1}^K \Omega_{d,k'} + \alpha_{k'}}$$

54

Zoom on the Dirichlet prior

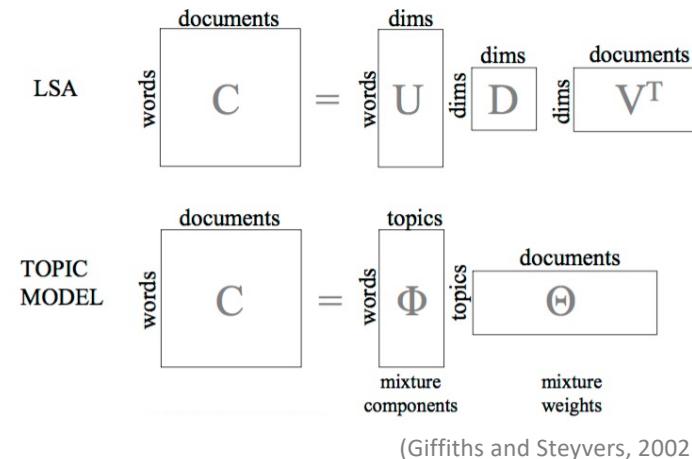
- The Dirichlet distribution is an exponential family distribution over the simplex, i.e., positive vectors that sum to one
- It is conjugate to the multinomial. Given a multinomial observation, the posterior distribution of θ is a Dirichlet.
- The parameter α controls the mean shape and sparsity of θ .
- The topic proportions are a K dimensional Dirichlet. The topics are a V dimensional Dirichlet.

55



56

Not so far from LSA



Outline

- Why topic learning
- Topic learning with matrix factorization
- Probabilistic graphical models
- Latent Dirichlet Allocation
- **Illustration on several case studies**
- More graphical models

57

58

Different types of datasets

- Scientific articles
- 20 Newsgroups
- Discharge summaries

} in collaboration with
P. Poncelet, M. Roche
and J.A. Lossio (LIRMM)

in collaboration with S. Chevret, R. Flicoteau
and M. Dermouche (INSERM – APHP)

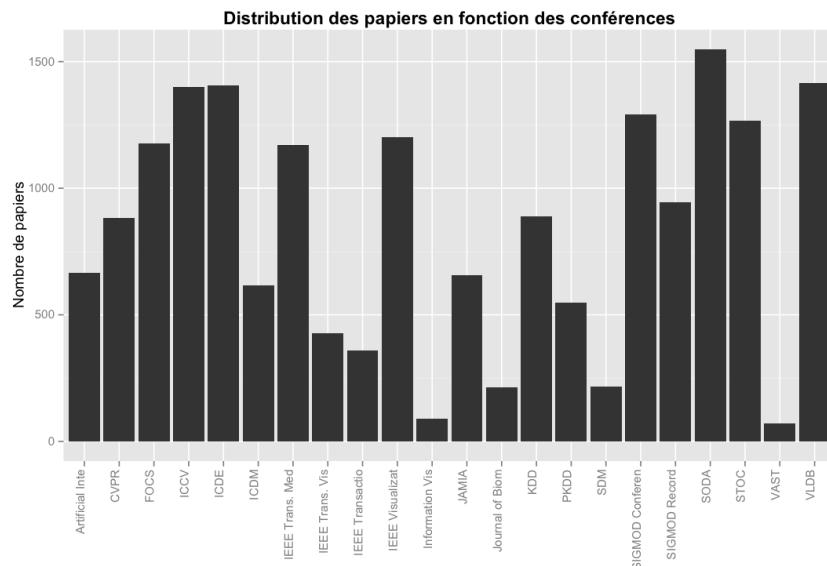
Scientific articles

- + de 18,000 titles with or w/o abstracts published between 1990 and 2005 (Tang et al., 2012)
 - database: ICDE, VLDB, SIGMOD...
 - data mining (after 1994) : KDD, ICDM...
 - visualization: CVPR, InfoViz, ICCV...
 - theoretical computer science: FOCS, SODA...
 - medical informatics: JAMIA, AIME...

59

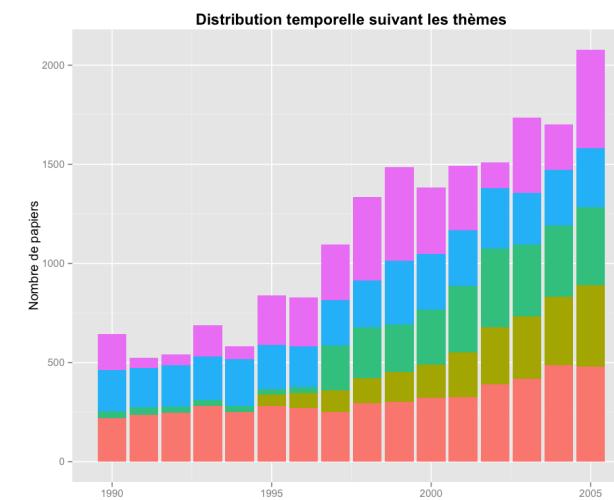
60

Distribution over venues



61

Temporal evolution of topics



62

Topics extracted with LDA (vocabulary of 5000 words)

Ida 0 : data - query - queries - database - performance - xml - system - processing - systems - relational - paper - efficient - databases - algorithms - memory - techniques - access - results - storage - time - optimization - present - index - distributed - show - structure - operations - approach - model - join...

Ida 1 : algorithm - problem - time - algorithms - graph - show - number - problems - approximation - graphs - bound - lower - bounds - complexity - set - optimal - case - polynomial - random - log - constant - linear - results - size - result - network - general - present - tree - model...

Ida 2 : image - images - method - surface - motion - model - object - algorithm - visualization - paper - volume - approach - data - rendering - objects - shape - points - present - models - flow - results - methods - technique - segmentation - reconstruction - surfaces - recognition - point - tracking - structure...

Ida 3 : data - mining - algorithm - clustering - learning - paper - approach - classification - results - method - algorithms - methods - problem - patterns - large - set - model - sets - analysis - show - number - time - models - present - search - performance - detection - association - pattern - efficient...

Ida 4 : data - information - system - systems - research - database - paper - web - visualization - user - model - application - management - knowledge - applications - design - users - databases - medical - analysis - integration - semantic - support - technology - development - network - environment - issues - language - process...

Topics extracted with LDA (vocabulary of 5000 ngrams, n>1)

Ida 0 : volume rendering - case study - research paper - vector fields - decision support - information technology - visualization techniques - a case study - volume data - vector field...

Ida 1 : data mining - time series - experimental results - knowledge discovery - machine learning - nearest neighbor - support vector - feature selection - decision tree - association rule...

Ida 2 : lower bound - lower bounds - polynomial time - approximation algorithms - extended abstract - approximation algorithm - running time - upper bound - competitive ratio - high probability...

Ida 3 : database systems - query processing - database system - query optimization - xml data - data management - query language - database management - management systems - relational database...

Ida 4 : experimental results - object recognition - computer vision - image sequences - optical flow - extended abstract - image segmentation - pattern matching - real images - motion estimation (...) a new approach...

63

64

20 NewsGroups

- 20,000 texts distributed in 20 categories

<http://qwone.com/~jason/20Newsgroups/>

comp.graphics	rec.autos	sci.crypt
comp.os.ms-windows.misc	rec.motorcycles	sci.electronics
comp.sys.ibm.pc.hardware	rec.sport.baseball	sci.med
comp.sys.mac.hardware	rec.sport.hockey	sci.space
comp.windows.x		
misc.forsale	talk.politics.misc	talk.religion.misc
	talk.politics.guns	alt.atheism
	talk.politics.mideast	soc.religion.christian

65

From: ahlenius@rtsg.mot.com (Mark Ahlenius)
Subject: converting color gif to X pixmap

I have looked through the FAQ sections and have not seen a answer for this.

I have an X/Motif application that I have written. I have a couple of gif files (or pict) that I have scanned in with a color scanner. Now I would like to be able to convert the gif files into a format that could be read into my application and displayed on the background of its main window. Preferably with pixmaps, or perhaps as an XImage.

I have found functions in the pbmplus program suite to convert gif to xbm, but that is monochrome, and I really do need color.

I have looked at xv, which reads in gif, and writes out several formats, but have not found a way to write out a file which can be read in as a pixmap.

Is there an easy way to do this?

66

From: leech@cs.unc.edu (Jon Leech)
Subject: Space FAQ 15/15 - Orbital and Planetary Launch Services

Archive-name: space/launchers
Last-modified: \$Date: 93/04/01 14:39:11 \$

ORBITAL AND PLANETARY LAUNCH SERVICES

category:
sci.space

The following data comes from _International Reference Guide to Space Launch Systems_ by Steven J. Isakowitz, 1991 edition.

Notes:

- * Unless otherwise specified, LEO and polar paylaods are for a 100 nm orbit.
- * Reliability data includes launches through Dec, 1990. Reliability for a familiy of vehicles includes launches by types no longer built when applicable
- * Prices are in millions of 1990 \$US and are subject to change.
- * Only operational vehicle families are included. Individual vehicles which have not yet flown are marked by an asterisk (*) If a vehicle had first launch after publication of my data, it may still be marked with an asterisk.

67

Vehicle (nation)	Payload kg (lbs)	GTO	Reliability	Price	Launch Site (Lat. & Long.)
	LEO	Polar			

Ariane (ESA)	35/40	87.5%	Kourou		
			(5.2 N, 52.8 W)		
AR40	4,900	3,900	1,900	1/1	\$65m
	(10,800)	(8,580)	(4,190)		
AR42P	6,100	4,800	2,600	1/1	\$67m
	(13,400)	(10,600)	(5,730)		
AR44P	6,900	5,500	3,000	0/0 ?	\$70m
	(15,200)	(12,100)	(6,610)		
AR42L	7,400	5,900	3,200	0/0 ?	\$90m
	(16,300)	(13,000)	(7,050)		
AR44LP	8,300	6,600	3,700	6/6	\$95m
	(18,300)	(14,500)	(8,160)		
AR44L	9,600	7,700	4,200	3/4	\$115m
	(21,100)	(16,900)	(9,260)		
* AR5	18,000	???	6,800	0/0	\$105m
	(39,600)		(15,000)		
		[300nm]			

category:
sci.space
(con't)

68

Topics extracted with LDA (vocabulary of 10,000 words)

Excerpt of the 20 extracted topics:

Ida 5 : window - file - program - server - set - motif - widget - application - problem - entry - display - code - sun - error - xterm - manager - running - work - subject - open - make - line - openwindows - number - size - x11r5 - function - run - version - client...

Ida 6 : image - file - jpeg - images - format - color - files - gif - program - display - version - bit - printer - convert - quality - programs - software - screen - formats - xv - good - colors - print - graphics - free - article - windows - postscript - tiff - fonts...

Ida 18 : god - jesus - church - bible - christ - christian - people - christians - sin - lord - faith - love - life - man - paul - word - law - time - article - good - heaven - hell - father - christianity - john - homosexuality - spirit - scripture - holy - things...

Ida 19 : space - nasa - launch - earth - article - orbit - shuttle - moon - mission - system - satellite - solar - time - spacecraft - data - years - lunar - station - flight - sky - cost - mars - project - venus - high - pat - surface - planet - program - henry...

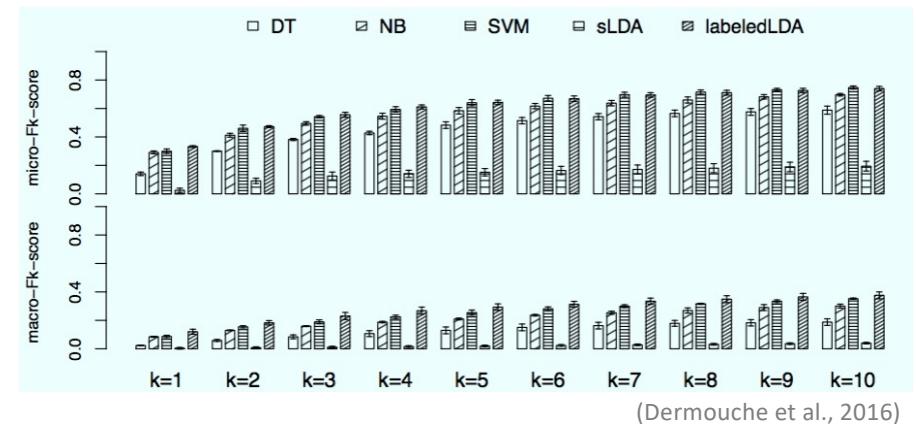
69

Discharge summaries

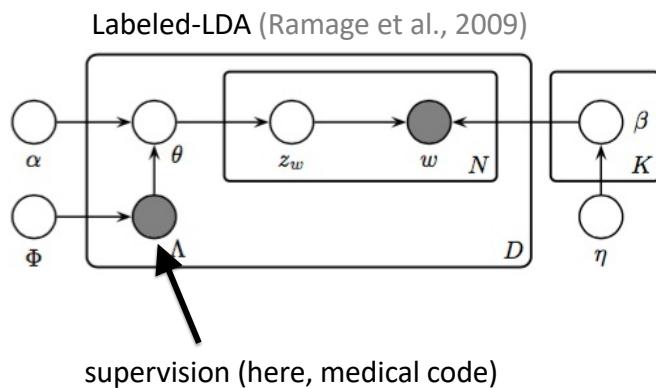
Dataset	ICD version	Lang.	#docs.	#unique words	#codes	Avg. #words /doc.	Avg. #docs. /code
URO-FR	CIM10	French	4 690	11 143	60	46	78
HEMATO-FR	CIM10	French	3 720	13 371	30	76	124
MIMIC-EN	ICD9	English	7 956	12 951	252	59	32

N40 masculin Antécédents médicaux Bloc de branche Arthrose Glaucome
Consultations Consultation urologie le 18 01 2010 Tentative d ablation de sonde vésicale Echec d ablation de sonde Programmer RTUP Examens complémentaires urologie Consultation urologie le 18 01 2010 Echographie Prostate 68cc Sonde vésicale en place Intervention urologie Cr opératoire urologie le 28 01 2010 Date d intervention s 28 01 2010Type RESECTION ENDOSCOPIQUE DE PROSTATE Histoire de la maladie Patient de 72 ans suivi pour adénome de la prostate Episode de rétention aigüe d urine en janvier 2010 nécessitant la mise en place d une sonde à demeure en urgence Echographie prostate de 68 gr Echec de tentative de l ablation de la sonde vésicale Indication à un traitement endoscopique pour RESECTION ENDOSCOPIQUE DE LA PROSTATE U3 Cr opératoire urologie le 28 01 2010 Date d intervention s 28 01 2010Type RESECTION ENDOSCOPIQUE DE PROSTATE Synthèse de l évolution Les suites opératoires ont été simples Arrêt des

Some comparative results



Supervised topic modeling



71

72

C61: Tumeur maligne de la prostate (Prostate cancer)	N39.3: Incontinence urinaire d'effort (Stress urinary incontinence)	Z52.4: Donneur de rein (Kidney donor)	N30.0: Cystite aiguë (Acute cystitis)	S30.2: Contusion des organes génitaux externes (Congestion of the external genitalia)
prostatectomie ⁵	incontinent	prélèvement (sample)	pontage (bypass)	observer (watch)
radical	bandelette (band)	faveur (favour)	arterielle (arterial)	hospitalisé(inpatient)
laparotomie (laparotomy)	effort (stress)	manuel (hand-operated)	Ditropan	med
score	trans-obturatrice ⁵	artère (artery)	post-mictionnel ⁵	ext
lobe (lobus)	urodynamique ⁵	assisté (assisted)	Kardégic	motif (cause)
mini	toux (cough)	DFG (GFR)	diurne (diurnal)	chir (surgery)
capsulaire (capsular)	bud (urodynam. test)	laparoscopique ⁵	surtout (especially)	ATCD (med. history)
élévé (high)	rééducation ⁵	contre (against)	fonctionnel(functional)	clinique-uro
extension	urgenterie ⁵	apparenté (related)	impériosité (urge)	fam (familial)
curatif (curative)	position	min	hypertension ⁵	suggérer (suggest)
#documents=356	#documents=47	#documents=39	#documents=16	#documents=18
F ₁ -score=0.68	F ₁ -score=0.83	F ₁ -score=0.96	F ₁ -score=0.00	F ₁ -score=0.22

C81.9: Lymphome de Hodgkin (Hodgkin's lymphoma)	C88.0: Macroglobulinémie de Waldenström (Waldenström's macroglobulinemia)	D46.2: Anémie réfractaire avec excès de blastes (refractory anemia with excess of blasts)	C83.0: Lymphome de blastes (refractory lymphoma)	E85.3: Amylose généralisée secondaire (secondary generalized amyloidosis)
Hodgkin	Waldenström	senior	critère (criterion)	amylose
ABVD	IgM	multirésistant(resistant)	participer(participate)	troponine (troponin)
IVOX	lymphoplasmocytaire	remise (redelivery)	accepter (accept)	formule (formula)
classique (classical)	macroglobulinémie ⁵	monoclonal	consentement(consent)	BNP
panoramique(panoramic)		béta (beta)	aborder (approach)	VCD
escalade (escalation)		AREB (RAEB)	attendu (expected)	évolution (evolution)
étoposide (etoposide)		leuco	logistique (logistics)	dosage (dose)
BEAM		Vidaza	version	pro
SPI (IPS)		myélotypasique ⁵	objectif (goal)	arriver (reach)
nodulaire (nodular)		BHC	contrainte(constraint)	immunochimique ⁵
		mgX (m.g.)		
#documents=168	#documents=72	#documents=37	#documents=38	#documents=85
F ₁ -score=0.75	F ₁ -score=0.74	F ₁ -score=0.78	F ₁ -score=0.38	F ₁ -score=0.34

3

...and in Harry Potter

excerpt from
20 topics:

School houses

house 0.04657586
gryffindor 0.04424846
points 0.03416309
slytherin 0.03261149
hundred 0.02252612
hat 0.02175032
will 0.02097452
cup 0.01554393
hufflepuff 0.01399234
taken 0.01321654

year 0.01958022
weasley 0.04050274
percy 0.03038466
fred 0.02869831
george 0.02448244
twins 0.01942340
cupboard 0.012015328

Professors of Hogwarts

professor 0.141749006
mcgonagall 0.074869763
dumbledore 0.035060689
quirrell 0.022321786
flitwick 0.015952334
turban 0.011175245
reached 0.007990520
teacher 0.007990520
dumbledore's 0.007194338
talking 0.007194338

4, Private Drive

uncle 0.065995844
dudley 0.062179040
vernon 0.057271720
aunt 0.035461411
petunia 0.031099349
letter 0.018013164
dudley's 0.012560586
room 0.012015328
street 74

But also...

looked 0.03222237	hagrid 0.06322251
like 0.02515218	yeh 0.06136366
eyes 0.02122431	ter 0.04835173
long 0.02122431	yer 0.03719865
know 0.02420258	said 0.02170826
back 0.02317602	dragon 0.01861018
see 0.01877648	fer 0.01799056
something 0.01862983	gringotts 0.01737095
think 0.01804323	got 0.01675133
now 0.01730997	don 0.01613172
going 0.01716332	
well 0.01511021	
harry 0.07567148	
one 0.03453437	door 0.03846168
first 0.02718846	open 0.02367537
time 0.01739391	cloak 0.02121098
much 0.01616959	looking 0.01874660
next 0.01543500	two 0.01874660
never 0.01494527	floor 0.01677509
day 0.01298636	forward 0.01677509

75

Preliminary tests of topic labeling

in collaboration with C. Gravier (LHC), M. Roche and P. Poncelet (LIRMM)

topic 4	topic 8	topic 19	topic 22	topic 35
snitch	parchment	fred	sirius	street
broom	quill	george	place	little
crowd	piece	said	order	alley
one	read	percy	dumbledore	way
bludger	writing	weasley	hogwarts	house
pitch	letter	ron	black	garden
team	ink	prefect	knew	diagon
wood	words	got	twelve	village
two	written	joke	also	side
stands	back	lee	secret	windows
My own label				
Golden snitch	Writing on a piece of parchment with a black quill	Fred and George (the twins)	Order of the phoenix, hidden number 12 Grimauld place	Diagon alley
0-order labeling (see Mei et al., 2007)				
bludger	piece of parchment	fred and george	order	diagon alley
bludger ,äi	quill	george and lee	order of the phoenix	diagon alley
stands	quick-quotes quill	prefect	order of merlin	alley
1-order labeling (see Mei et al., 2007)				
bludger	parchment	george and lee	grimmauld place	shop windows
snitch	piece of parchment	fred and george	number twelve	dark street
pitch	bottle of ink	fred	sirius black	street
Likelihood based on p(d/z)				
team mascots	black quill	brother percy	grimmauld place	main street
comet two sixty	roll of parchment	prefect badge	order of the phoenix	high street
golden snitch	piece of parchment	george and lee	number twelve	side street
76				

Outline

- Why topic learning
- Topic learning with matrix factorization
- Probabilistic graphical models
- Latent Dirichlet Allocation
- Illustration on several case studies
- **More graphical models**

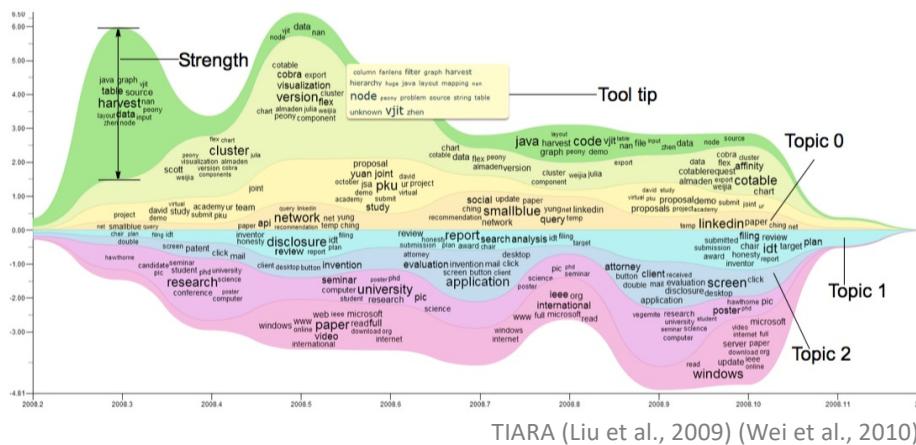
77

Some interesting issues

- Finding the “best” number of topics (Teh et al., 2004)
- Automatic topic summarization (Mei et al., 2007)
- Finding the hidden structure between topics CTM (Blei and Lafferty, 2006)
- Combining topics with other information: authors, opinion, structure, etc.
 - Author-Topic (Rosen-Zvi et al., 2006)
 - Topic-Opinion (Mei et al., 2007) (Dermouche et al., 2014)
- Link with word embedding (Das et al., 2015)
- Temporal extensions (see the following)

78

Topic dynamics

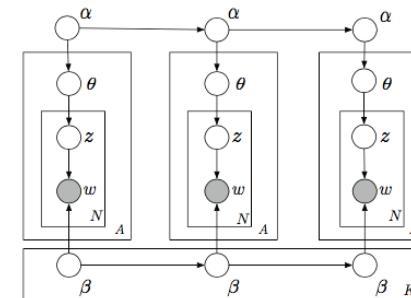


79

Dynamic Topic Model

(Blei and Lafferty, 2006)

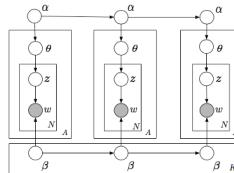
- Graphical model:



- Links inspired by Brownian motion

80

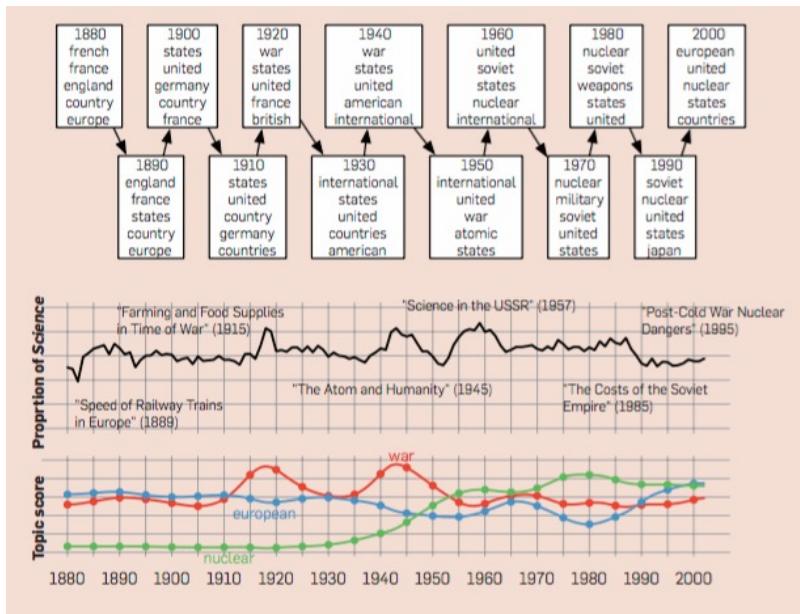
Generative process



1. Draw topics $\beta_t | \beta_{t-1} \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$.
2. Draw $\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$.
3. For each document:
 - (a) Draw $\eta \sim \mathcal{N}(\alpha_t, a^2 I)$
 - (b) For each word:
 - i. Draw $Z \sim \text{Mult}(\pi(\eta))$.
 - ii. Draw $W_{t,d,n} \sim \text{Mult}(\pi(\beta_{t,z}))$.

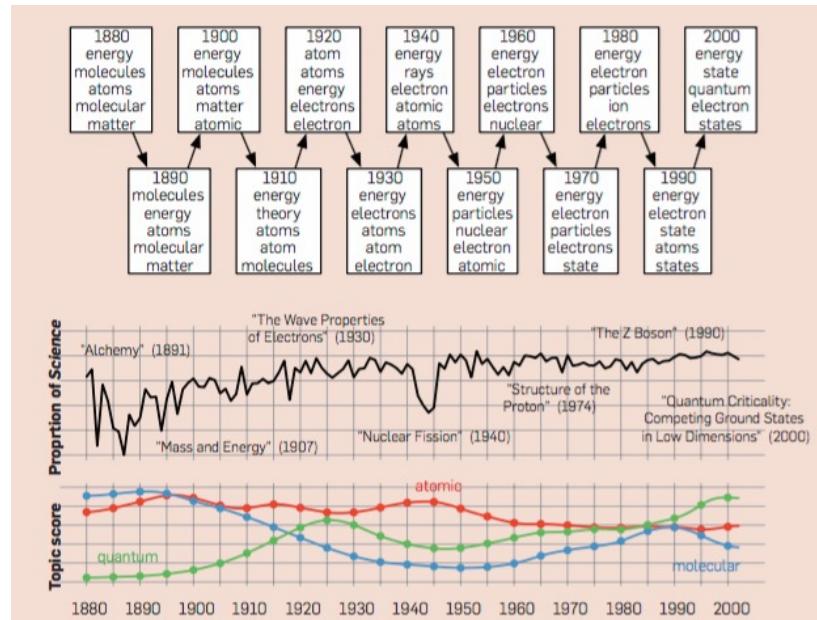
from natural parameters to normalized parameters

81



data from *Science* between 1880 and 2002 (Blei, 2012)

83



data from *Science* between 1880 and 2002 (Blei, 2012)

82

And for “our” scientific articles?

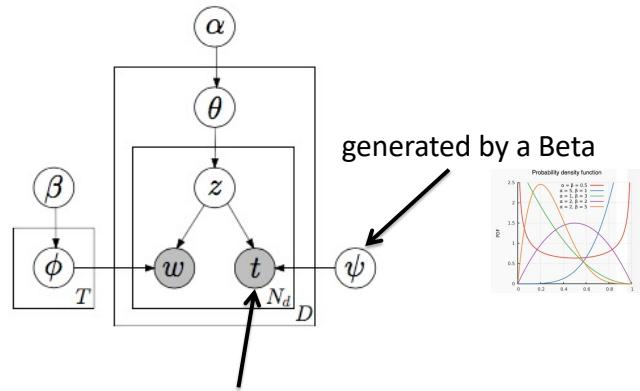
Excerpt for the “database” topic:

	1996	1997	1998	1999	2000	2001	2002	2003
database	data							
data	database	database	query	query	query	query	query	query
query	query	query	queries	queries	queries	queries	queries	queries
system	system	queries	system	system	web	web	web	xml
systems	queries	system	system	web	web	web	xml	web
object	systems	systems	web	system	xml	xml	xml	web
queries	object	performance	systems	systems	system	system	system	system
performance	performance	databases	performance	paper	paper	paper	paper	paper
databases	databases	paper	paper	performance	systems	performance	performance	performance
management	paper	information	information	information	information	information	information	relational

84

Topic Over Time

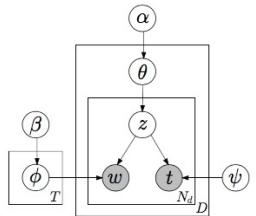
(Wang et McCallum, 2006)



Temporal dimension as an observed random variable

85

Generative process

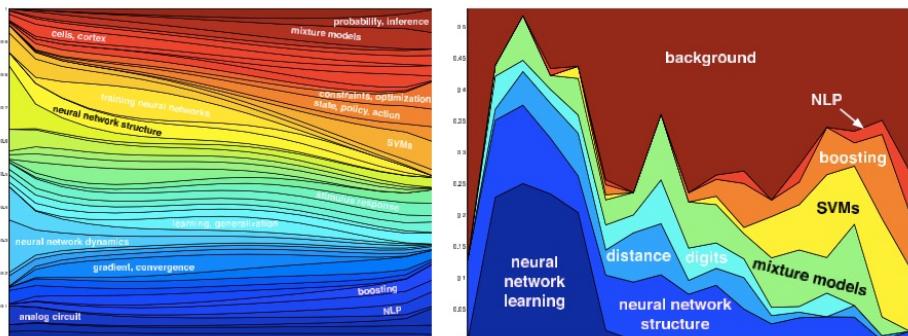


1. Draw T multinomials ϕ_z from a Dirichlet prior β , one for each topic z ;
2. For each document d , draw a multinomial θ_d from a Dirichlet prior α ; then for each word w_{di} in document d :
 - (a) Draw a topic z_{di} from multinomial θ_d ;
 - (b) Draw a word w_{di} from multinomial $\phi_{z_{di}}$;
 - (c) Draw a timestamp t_{di} from Beta $\psi_{z_{di}}$.

yes, two words of the *same* document can be associated to *different* timestamps!

86

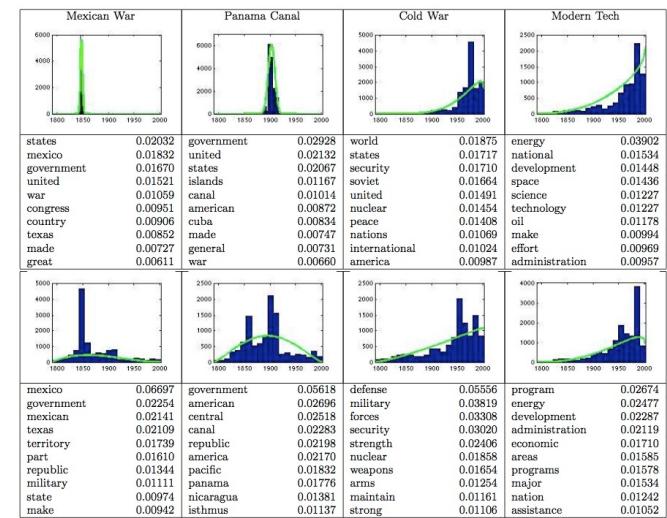
Results on scientific trends



Data extracted from the proceedings of NIPS between 1987 and 2003

87

Results on the state of the union



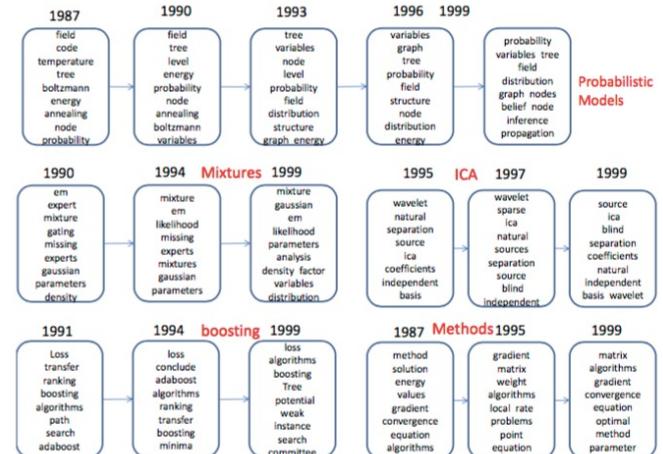
88

Some follow-up

- Time discretization
 - cDTM (Wang et al., 2008)
- Vocabulary evolution
 - Online LDA with infinite vocabulary
(Zhai and Boyd-Graber, 2013)
- Number of topics
 - DP process based evolutionary clustering
(Xu et al., 2008)
 - Infinite DTM (Ahmed and Xing, 2010)

89

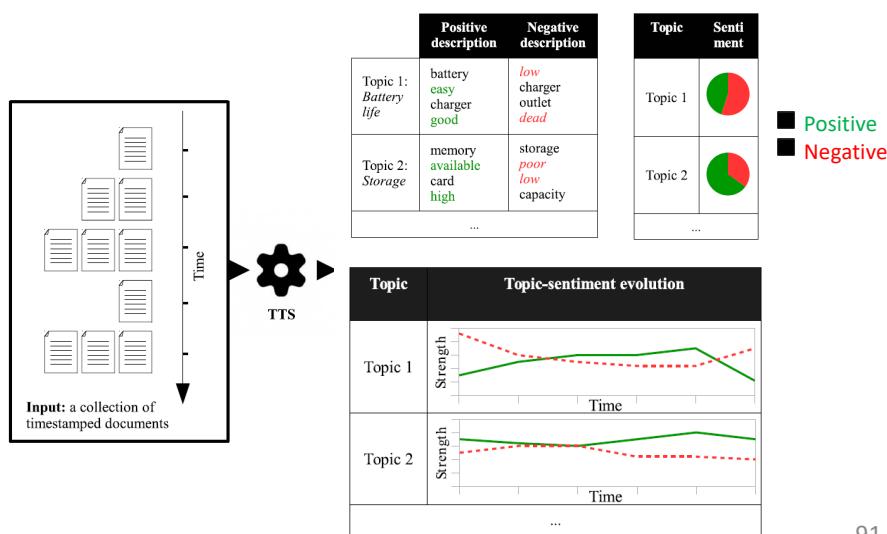
Infinite DTM



Data extracted from the proceedings of NIPS (Ahmed and Xing, 2010)

90

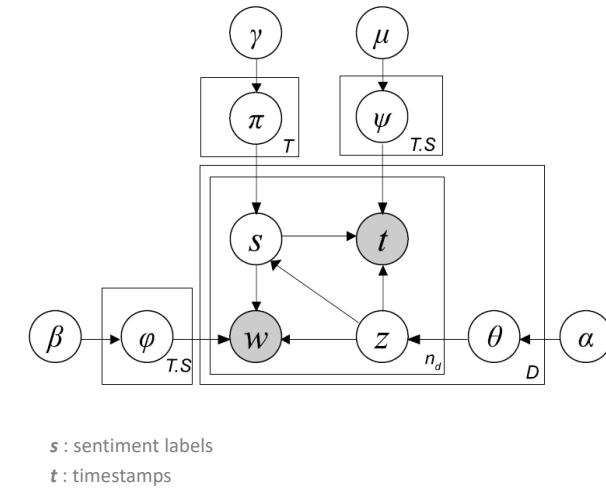
TTS : joint analysis of topic, sentiment and time



91

Time-aware Topic x Sentiment model

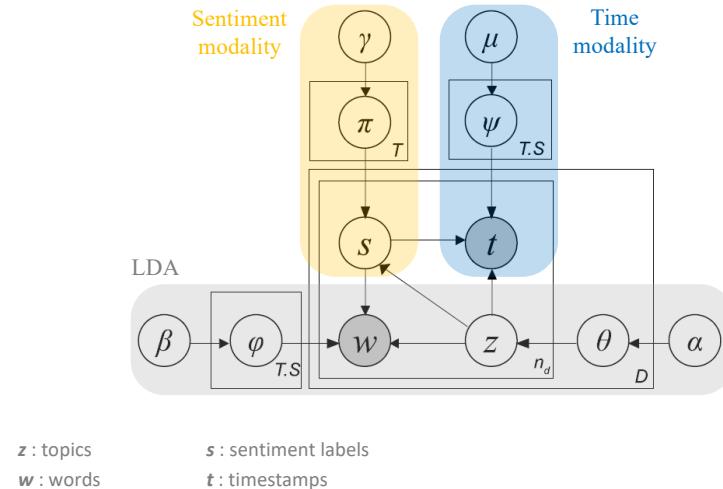
(Dermouche et al., 2014)



z : topics
 w : words
 s : sentiment labels
 t : timestamps
 n_d

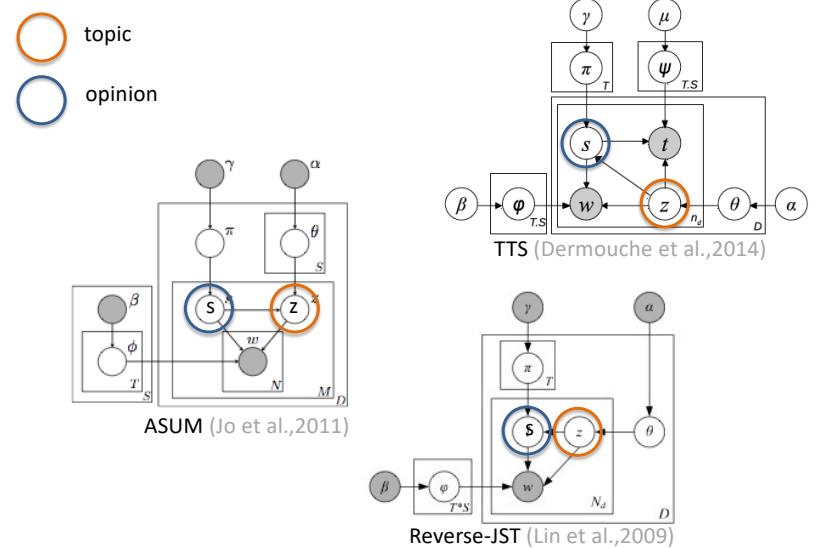
92

TTS – graphical model



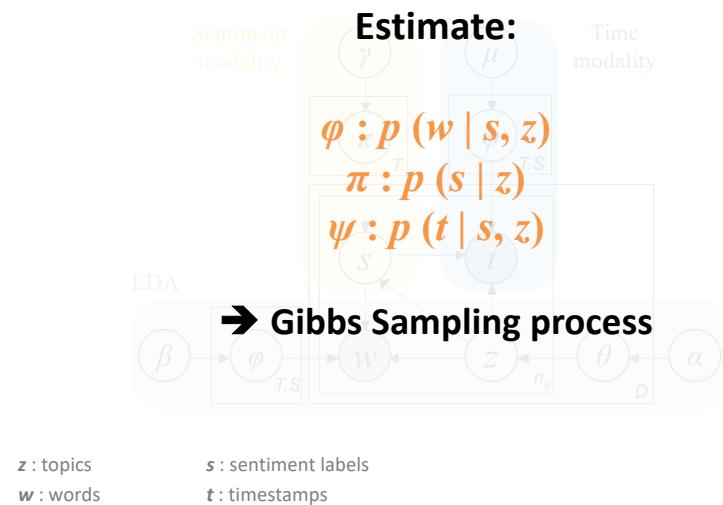
93

Different models, different assumptions



94

TTS – graphical model



95

TTS – parameters estimation (sketch)

- Joint probability: $p(w, t, s, z | \alpha, \beta, \gamma, \mu) = p(w|s, z) \cdot p(t|s, z) \cdot p(s|z) \cdot p(z|\alpha)$
- Use it for deriving the marginal probability $p(s, z | .)$ and the parameters updates:

$$\phi_{j,k,i} = \frac{n_{i,j,k} + \beta}{n_{j,k} + V \cdot \beta} \quad \theta_{d,j} = \frac{n_{d,j} + \alpha_j}{n_d + \sum_{j'} \alpha_{j'}} \quad \text{a.s.o.}$$
- Integrate opinionated lexicon knowledge
- Weight the temporal dimension by $1/n_d$

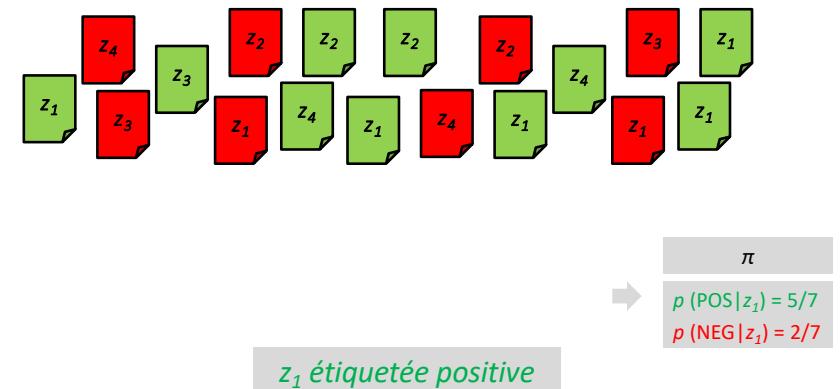
96

Evaluation scheme

- Two datasets
 - MDS (Amazon reviews) ➔ $\approx 29,000$ reviews
 - NYSK (Dominique Strauss-Kahn case) ➔ $\approx 10,000$ newswires
- Accuracy of sentiment prediction at *document* level
➔ not the main purpose of TTS model
- KL distance between “estimation” and “reality”
 - Q_s = distance between topic distributions over sentiments (π)
 - Q_t = distance between topic-sentiment distributions over time (ψ)

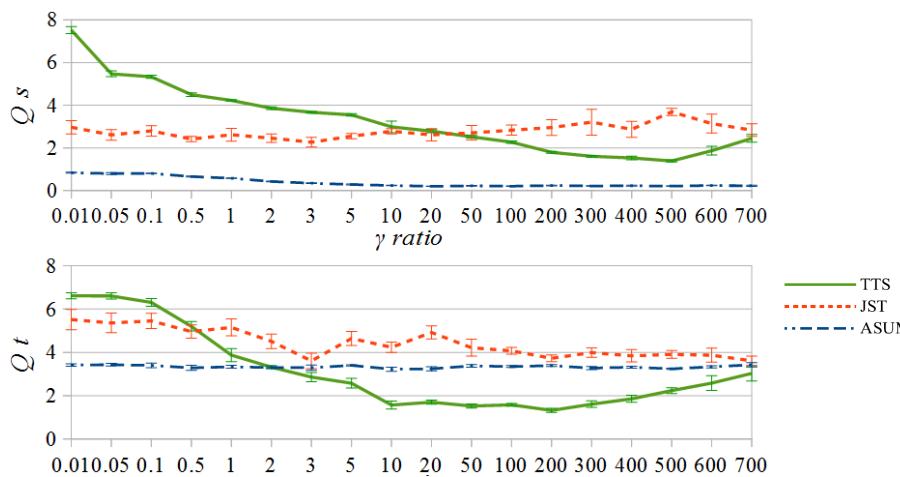
97

Building the gold standard Q_s



98

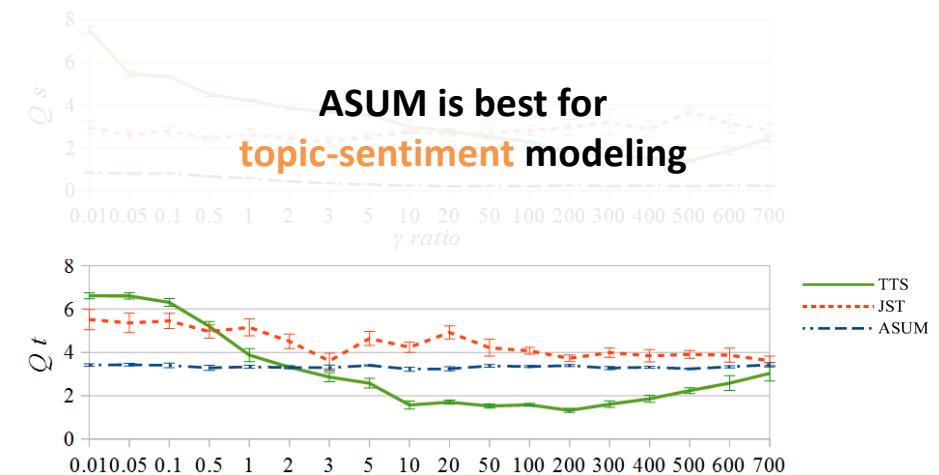
Results – accuracy on MDS



Note that γ can be estimated dynamically (Dermouche et al., 2015)

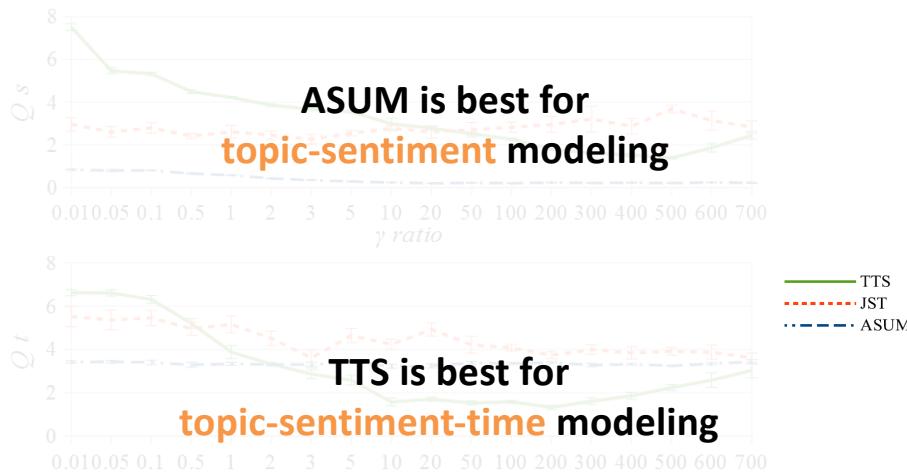
99

Results – accuracy on MDS



100

Results – accuracy on MDS



101

Results – examples on MDS

z_1 : computer & video games		z_2 : beauty		z_3 : software		z_4 : gourmet food	
positive	negative	positive	negative	positive	negative	positive	negative
game	way	hair	smell	use	comput	use	comput
play	player	scent	shaver	software	xp	software	xp
one	bore	great	shave	work	upgrad	work	upgrad
fun	puzzl	thick	dri	internet	support	internet	support
good	run	eye	feel	crash	file	crash	file
graphic	long	iron	recommend	system	manual	system	manual
level	bad	disappoint	featur	featur	slow	featur	slow
better	hard	clean	irrit	easi	connect	easi	connect

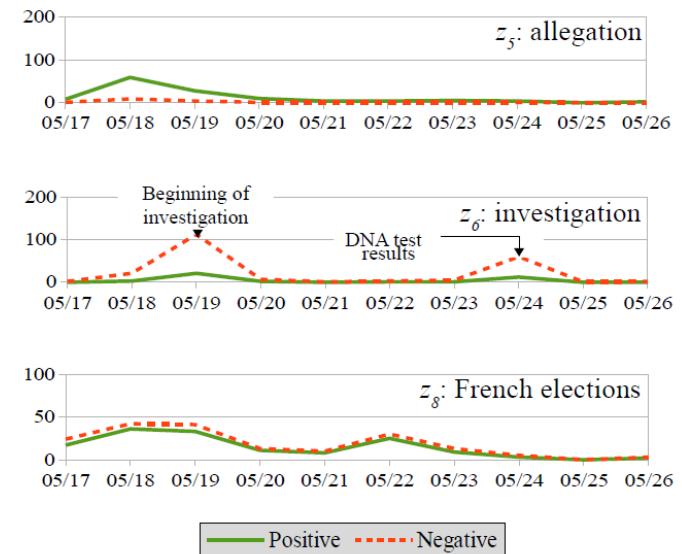
102

Results – examples on MDS

z_1 : computer & video games		z_2 : beauty		z_3 : software		z_4 : gourmet food	
positive	negative	positive	negative	positive	negative	positive	negative
game	way	hair	smell	use	comput	tast	tea
play	player	scent	shaver	software	xp	coffe	flavor
one	bore	great	shave	work	upgrad	drink	drink
fun	puzzl	thick	dri	internet	support	milk	milk
run	eye	eye	feel	crash	file	brand	brand
graphic	feel	feel	system	manual	crash	fat	fat
level	chocol	chocol	featur	connect	system	chocol	textur
	sweet	sweet	connect			sweet	treat

103

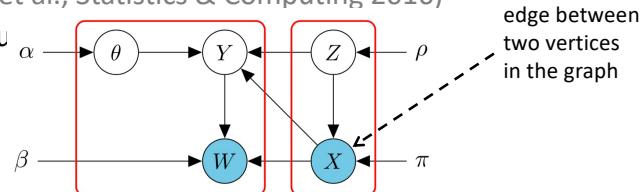
On the second case study: NYSK



104

Stochastic Topic Block Model

- Work of Rawya Zreik at SAMM, Paris
(Bouveyron et al., Statistics & Computing 2016)
- One contribu

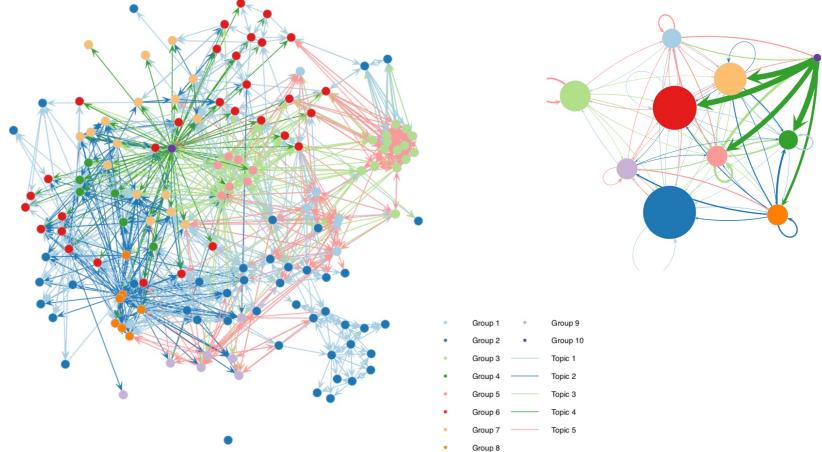


- It combines the stochastic block model (Wang & Wong, 1987) to LDA (Blei et al., 2003)
- Valorized as a professional tool (see <http://linkage.fr>)

105

Stochastic Topic Block Model (con't)

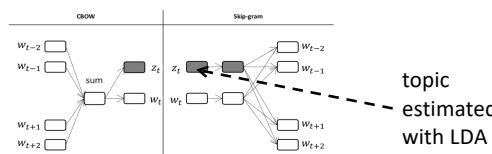
- Example of clustering result with STBM on the Enron data set (Sept.-Dec. 2001)



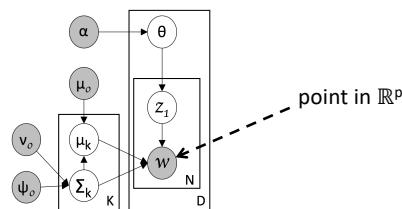
106

Possible links with distributional semantics

- Topic2Vec: LDA and word2vec (Niu et al., IALP 2015)



- Gaussian LDA: word2vec and LDA
(Das et al., ACL 2015)



- Many other attempts

(Li et al., ACL 2016) (Xun et al., IJCAI 2017)

107

Some personal conclusions

- Positive aspects of probabilistic approaches
 - “clean” dependency modeling (even with heterogeneous data) for probabilistic models vs. BERTTopics and prompt-based models
 - techniques for parameter estimation in reasonable time
- Some difficulties on the road
 - parameter estimation
 - model interpretation
 - granularity level
- Personal perspectives
 - flaw in the evaluation framework currently used (?)
 - well-grounded comparison between traditional TM and neuronal TM

108

References

- Blei, D.M., A.Y. Ng and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: pp. 993–1022.
- Blei, D.M. (2011). Probabilistic Topic Models. Tutorial at KDD. <https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf>
- Graham S. et al. (2012). Getting Started with Topic Modeling and MALLET. Online lesson. <http://programminghistorian.org/lessons/topic-modeling-and-mallet>
- McCallum, A. (cours, 2011) <https://people.cs.umass.edu/~mccallum/courses/gm2011>
- Wang, Y. (2008). Distributed gibbs sampling of latent topic models: The gritty details. Tech. Rep. <https://cxwangyi.files.wordpress.com/2012/01/lit.pdf>
- Xia, H. and P. Luo (2006). Graphical Representation, Generative Model, Gibbs Sampling. Tutorial (available online).
- Weingart, S. (2012). Topic Modeling for Humanists: A Guided Tour. <http://www.scottbot.net/HIAL/?p=19113>