UNIVERSITÉ DE LYON

ICOM INSTITUT DE LA COMMUNICATION

INFO STAT

Topic learning

Master Data Mining

Julien Velcin

Outline

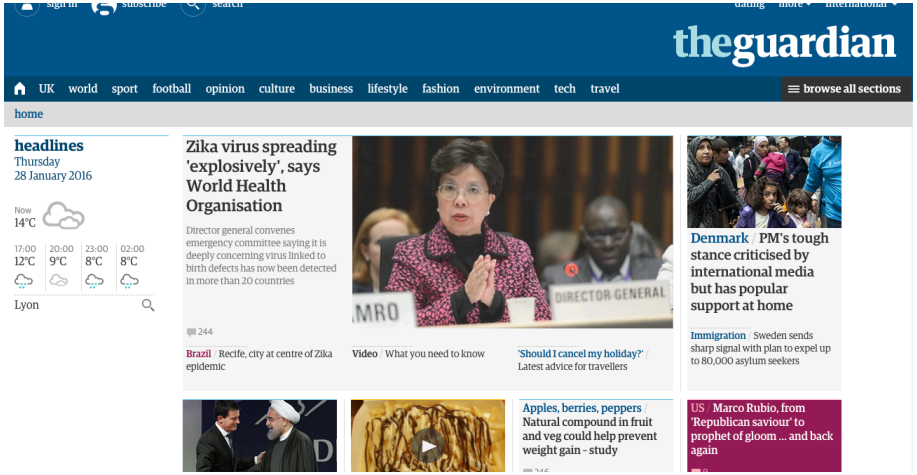
- Why topic learning
- Probabilistic graphical models
- Latent Dirichlet Allocation
- Illustration on several case studies
- More graphical models

Foreword

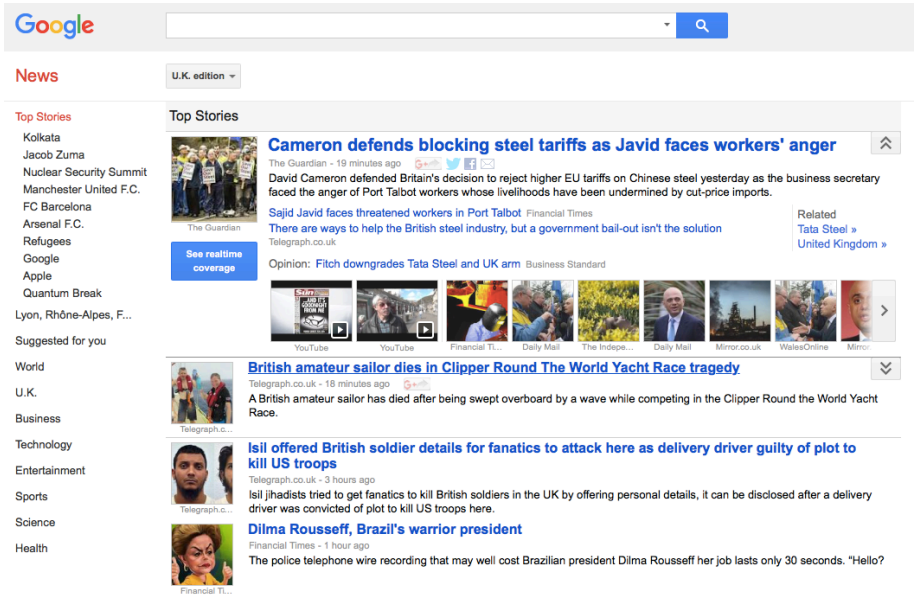
- Several slides are directly taken from the talk given by David M. Blei for KDD (2011)
  https://www.cs.princeton.edu/~blei/kdd-tutorial.pdf
- Some of them related to graphical model come from the HP team (2006)
  http://home.in.tum.de/~xiaoh/pub/3G_talk.pdf
- The derivation for the Gibbs sampling is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

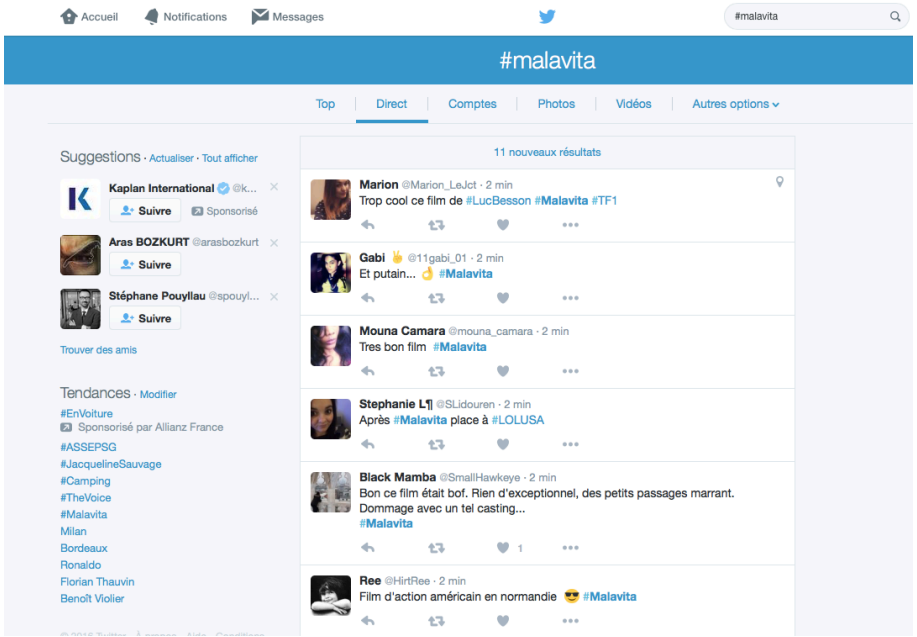
Outline

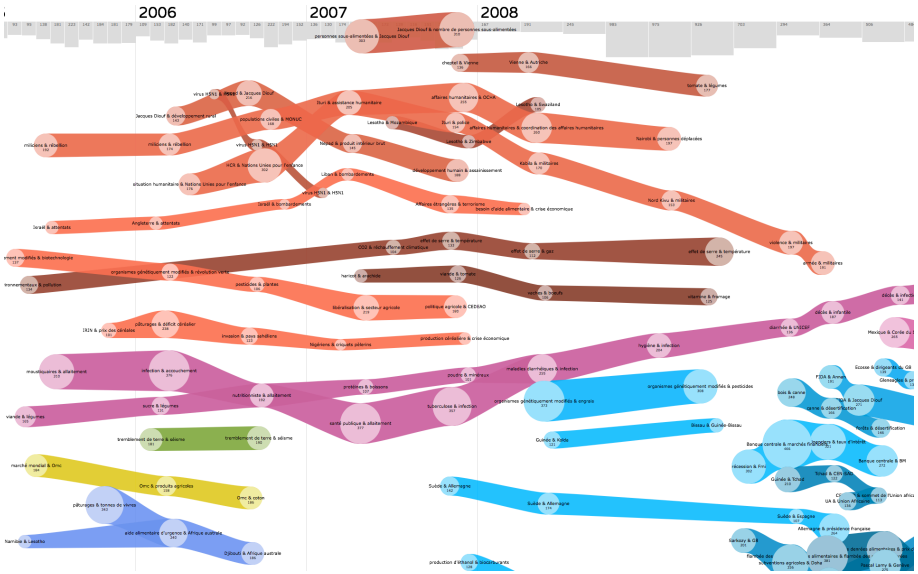
- **Why topic learning**
- Probabilistic graphical models
- Latent Dirichlet Allocation
- Illustration on several case studies
- More graphical models



Google News







Pulseweb project: http://pulseweb.cortext.net

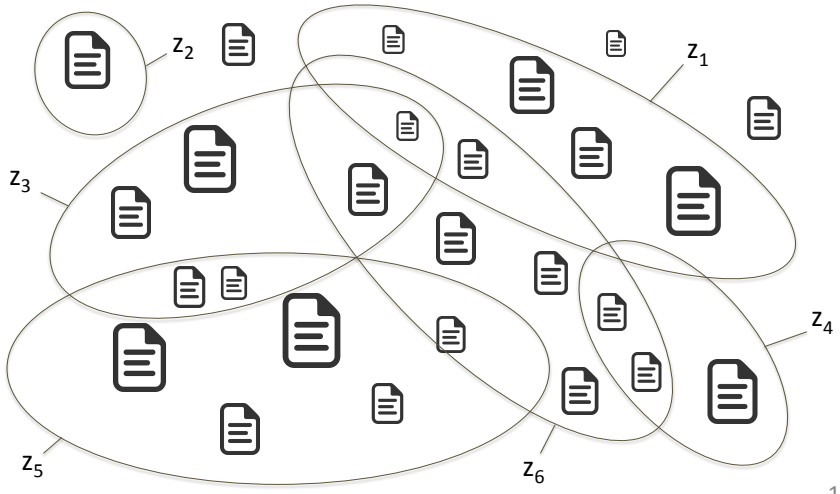
Why topic learning?

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

Difference with clustering

(Xie and Xing, 2013)



Various approaches for topic learning

- Algebraic approaches
  - LSA (Deerwester et al., 1990)
  - NMF (Paatero et Tapper, 1994)
  - Dictionary learning (Jenatton et al., 2010)
- Geometrical approaches
  - TDT (Allan et al., 1998) (Pons-Porrata et al., 2003)
  - AGAPE (Velcin and Ganascia, 2007)
- Probabilistic approaches
  - pLSA, LDA… (see the following)

Some background on text mining

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

- Classic input:

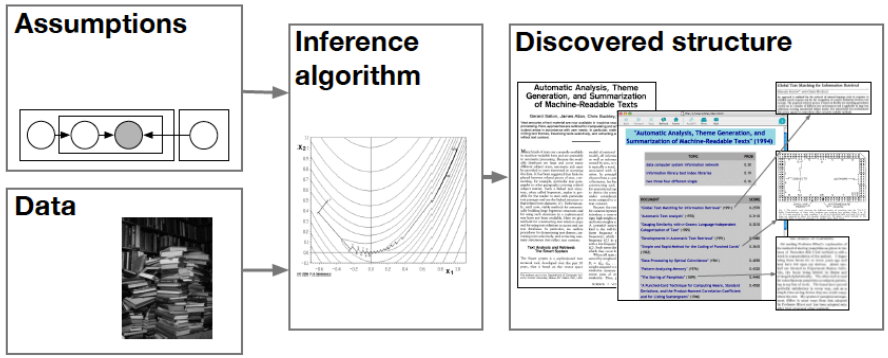
	Docs																			
Terms	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
data	1	1	0	0	2	0	0	0	0	0	1	2	1	1	1	0	1	0	0	0
examples	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
introduction	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
mining	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
network	0	0	0	0	0	0	0	0	0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

Outline

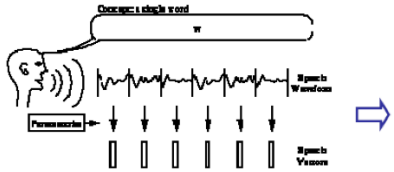
- Why topic learning
- **Probabilistic graphical models**
- Latent Dirichlet Allocation
- Illustration on several case studies
- More graphical models

Probabilistic graphical models

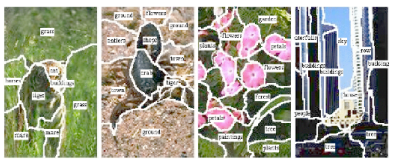
« If you remember one picture » :



Examples given by A. McCallum



Speech recognition



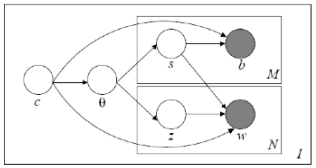
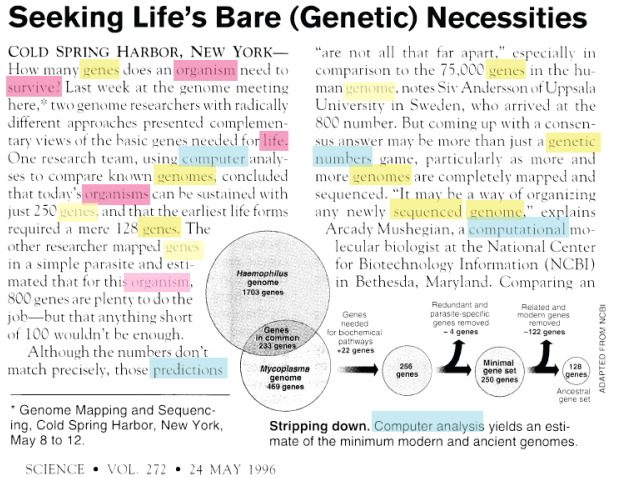


Image analysis

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.

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

SCIENCE • VOL. 272 • 24 MAY 1996

Simple intuition: Documents exhibit multiple topics.

Discovering latent structures

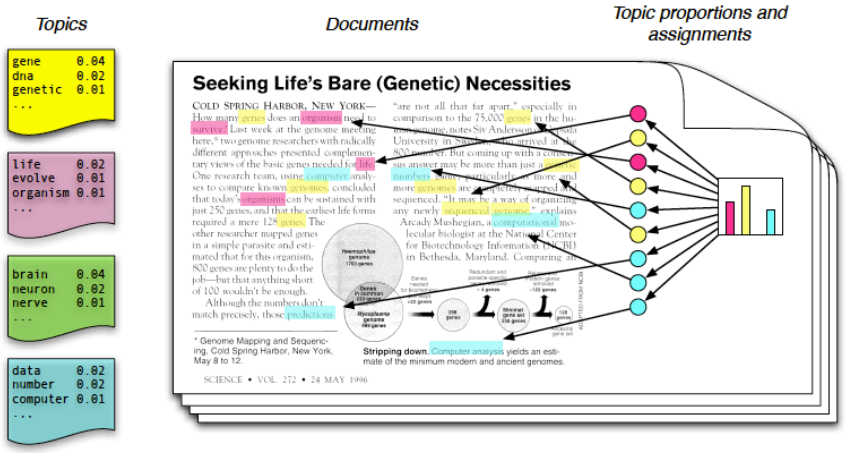


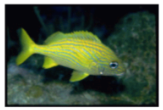
Image				
Ground Truth	field, foals, horses, mare	beach, horizon, people, water	waved, albatross, flight, sky	coast, sky, water, waved
PLSA-WORDS Annotation	grass, foals, horses, garden, trees	water, trees, beach, flowers, garden	city, flight, ceremony, pond, swallow-tailed	trees, sky, snow, clouds,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

(Blei et al., 2003)

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

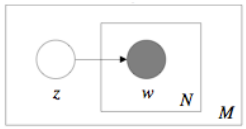
True caption
clouds jet plane
Corr–LDA
sky plane jet mountain clouds
GM–LDA
sky water people tree clouds
GM–Mixture
sky plane jet clouds pattern

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:

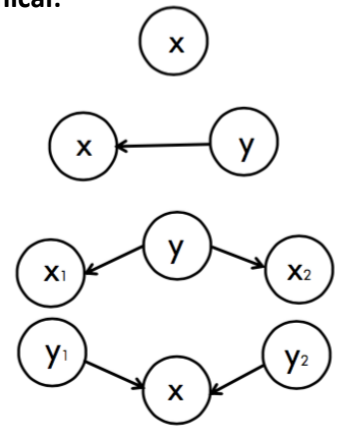


Algebraic vs. graphical

Algebraic :

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

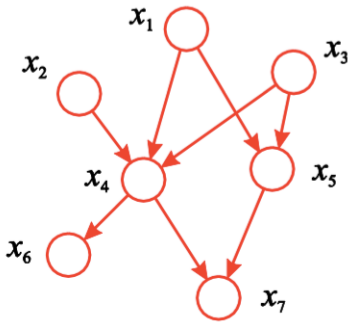
Graphical:



Joint distribution

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

where pa_i stands for the parents of node i



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:

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

Example of a generative process

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

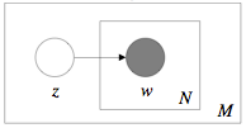
For each document M:

- draw a topic z

$$z \sim Mult(p)$$

- for each token N :
  - draw a word w given z

$$w \sim Mult_z(p)$$



Outline

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

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.

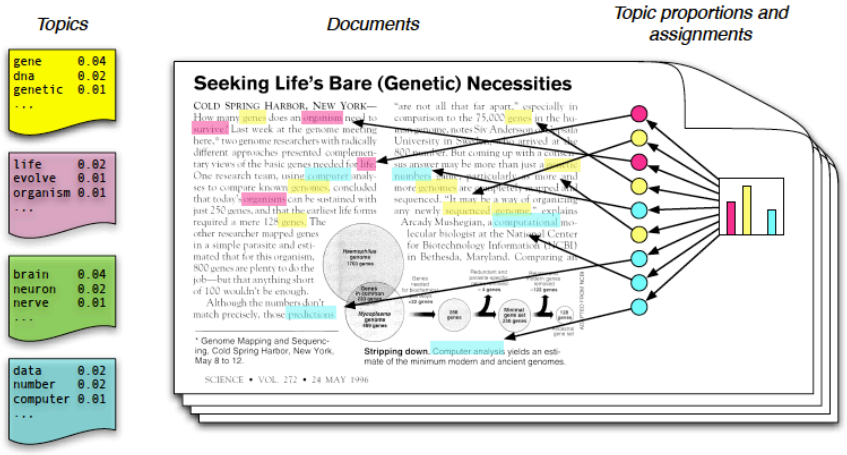
Infer the hidden structure using **posterior inference**

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?

Generative model for LDA



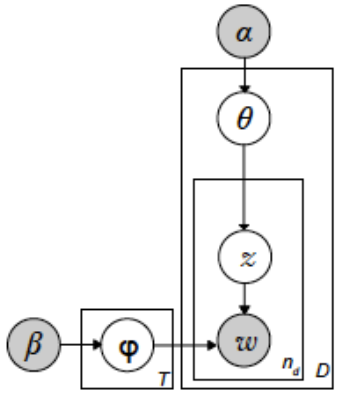
LDA as a graphical model

ϕ : $p(w_i/z_j)$

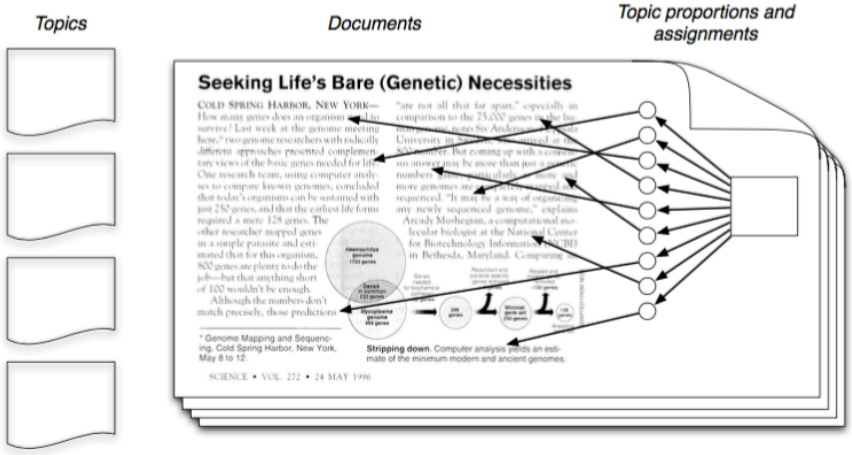
θ : $p(z_j/d_m)$

α : prior $p(\theta)$

β : prior $p(\phi)$



The posterior distribution



Infering the latent variables

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

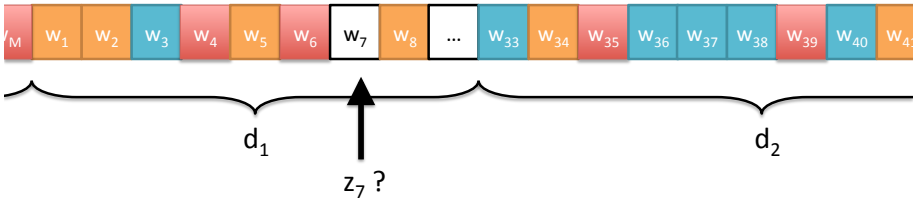
Intractable!

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

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)



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

Estimation with MCMC (2)

- Joint distribution for z and w:

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

- First term:

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

$$\underbrace{\prod_{k=1}^{K} \prod_{v=1}^{V} \phi_{k,v}^{\psi_{k,v}} \quad \prod_{k=1}^{K} \frac{1}{B(\beta)} \prod_{v=1}^{V} \phi_{k,v}^{\beta_v - 1}}_{\text{conjuguacy}}$$

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

Beta function

prior on word v

Estimation with MCMC (4)

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

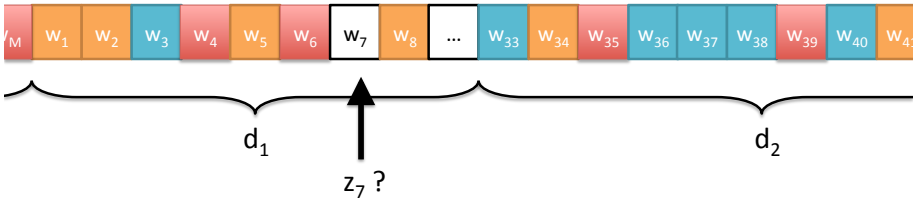
$$p(w/z,\beta) = \prod_{k=1}^{K} \frac{B(\psi_k+\beta)}{B(\beta)}$$

The same reasoning holds for $p(z/\alpha)$

So that we can calculate $p(z,w|\alpha,\beta)$

…and drives the Gibbs sampling procedure

Estimation with MCMC (5)



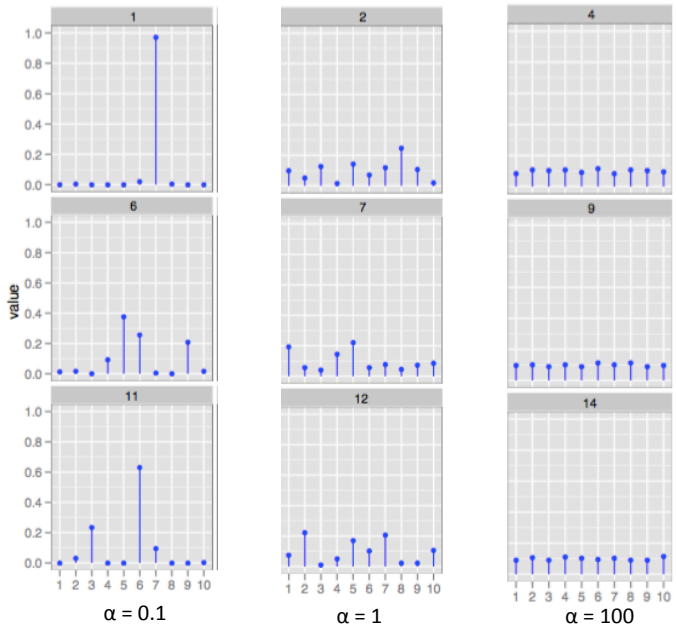
updating rules

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

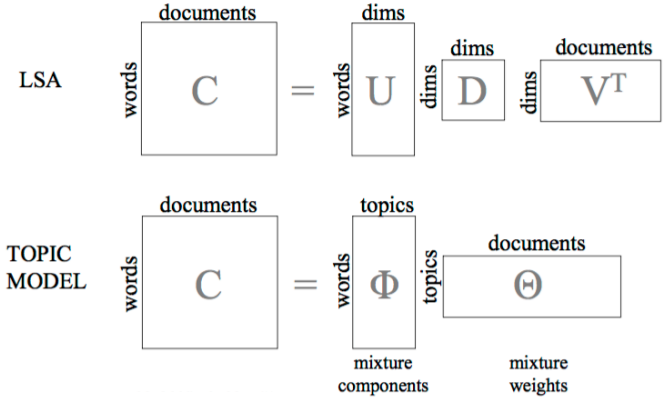
$$\theta_{d,k} = \frac{\Omega_{d,k}+\alpha_k}{\sum_{k'=1}^{K} \Omega_{d,k'}+\alpha_{k'}}$$

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.



Not so far from LSA



(Giffiths and Steyvers, 2002)

Outline

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

Different types of datasets

- Scientific articles
- 20 Newsgroups

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

- Discharge summaries

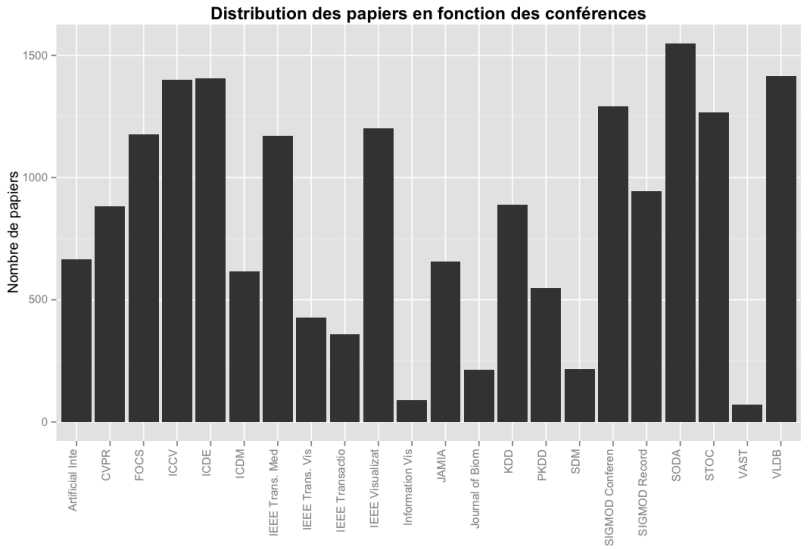
in collaboration with S. Chevret, R. Flicoteau and M. Dermouche (INSERM – APHP)

Scientific articles

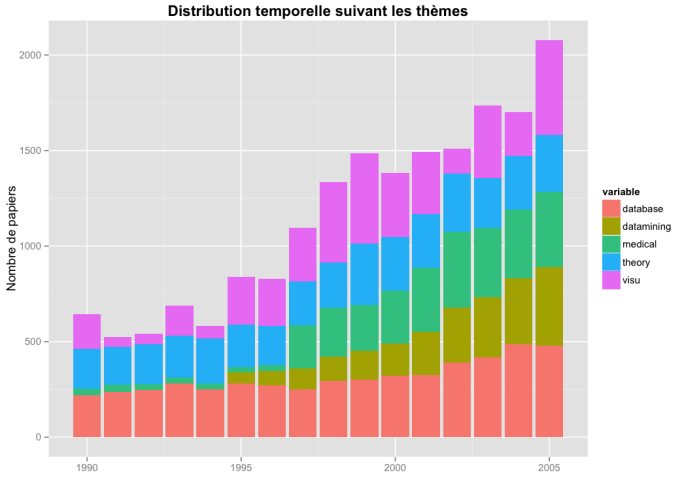
+ de 18,000 titles with or w/o abstracts published between 1990 and 2005 (Tang et al., 2012)

- database: ICDE, VLDB, SIGMOD…
- data mining (after 1994) : KDD, ICDM…
- visualization: CVPR, InfoViz, ICCV…
- theoretical computer science: FOCS, SODA…
- medical informatics: JAMIA, AIME…

Distribution over venues



Temporal evolution of topics



Topics extracted with LDA
(vocabulary of 5000 words)

lda 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…

lda 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…

lda 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…

lda 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…

lda 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)

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

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

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

lda 3 : database systems - query processing - database system - query optimization - xml data - data management - query language - database management - management systems - relational database…

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

20 NewsGroups

- 20,000 texts distributed in 20 categories
  http://qwone.com/~jason/20Newsgroups/

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

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?

category:
comp.windows.x

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.
* Reliablity data includes launches through Dec, 1990. Reliabity 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.

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

category: sci.space (con't)

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

Excerpt of the 20 extracted topics:

lda 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…

lda 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…

lda 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…

lda 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…

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

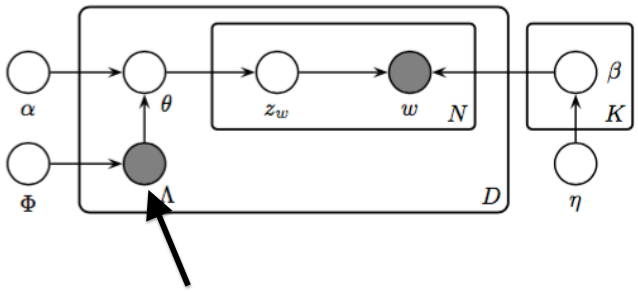
N40

Hyperplasie de la prostate

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 aigue 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

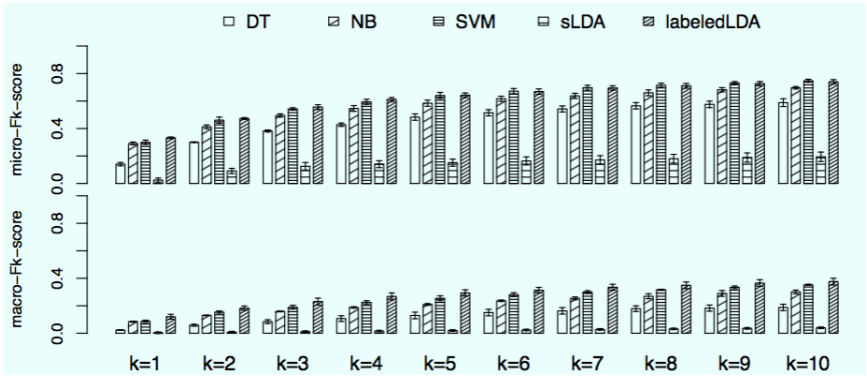
Supervised topic modeling

Labeled-LDA (Ramage et al., 2009)



supervision (here, medical code)

Some comparative results



(Dermouche et al., 2016)

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)	*Kardegic*	*motif* (cause)
mini	*toux* (cough)	*DFG* (GFR)	*diurne* (diurnal)	*chir* (surgery)
capsulaire (capsular)	*bud* (urodynam. test)	*laparoscopique*[5]	*surtout* (especially)	*ATCD* (med. history)
élevé (high)	*rééducation*[5]	*contre* (against)	*fonctionnel*(functional)	*clinique-uro*
extension	*urgenturie*[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**: *Macroglobulinmie de Waldenström* (Waldenström's macroglobulinemia)	**D46.2**: *Anémie réfractaire avec excès de blastes* (refractory anemia with excess of blasts)	**C83.0**: *Lymphome à petites cellules B* (small B-cell 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*[5]	*remise* (redelivery)	*accepter (accept)*	*formule* (formula)
classique (classical)	*macroglobulinémie*[5]	*blaste* (blast)	*consentement(consent)*	*BNP*
panoramique(panoramic)	*monoclonal*	*AREB* (RAEB)	*aborder (approach)*	*VCD*
escalade (escalation)	*béta* (beta)	*leuco*	*attendu (expected)*	*évolution* (evolution)
étoposide (etoposide)	*créatininémie*[5]	*Vidaza*	*logistique (logistics)*	*dosage* (dose)
BEAM	*sup* (increased)	*myélodysplasique*[5]	*version*	*pro*
SPI (IPS)	*stabilité* (stability)	*BHC*	*objectif (goal)*	*arriver* (reach)
nodulaire (nodular)	*cérébral* (cerebral)	*mgX* (m.g.)	*contrainte(constraint)*	*immunochimique*[5]
#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

…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

Quidditch
wood 0.04386071
quidditch 0.03491612
team 0.02060479
quaffle 0.01702696
snitch 0.01613250
game 0.01523804
catch 0.01344912
play 0.01255466
flint 0.01255466
seeker 0.01166021

Weasley family
weasley 0.04050274
percy 0.03038466
fred 0.02869831
george 0.02448244
twins 0.01942340
year 0.01858022

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
cupboard 0.012015328

But also…

said 0.08080992
get 0.02566909
know 0.02420258
back 0.02317602
see 0.01877648
something 0.01862983
think 0.01804323
now 0.01730997
going 0.01716332
well 0.01511021

looked 0.03222237
like 0.02515218
eyes 0.02122431
long 0.02122431
little 0.01886758
black 0.01886758
took 0.01611806
hat 0.01572528
face 0.01533249
pulled 0.01533249

hagrid 0.06322251
yeh 0.06136366
ter 0.04835173
yer 0.03719865
said 0.02170826
dragon 0.01861018
fer 0.01799056
gringotts 0.01737095
got 0.01675133
don 0.01613172

harry 0.07567148
one 0.03453437
first 0.02718846
time 0.01739391
much 0.01616959
next 0.01543500
never 0.01494527
day 0.01298636

door 0.03846168
open 0.02367537
cloak 0.02121098
looking 0.01874660
two 0.01874660
floor 0.01677509
forward 0.01677509

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 ,Äî	**quill**	george and lee	**order of the phoenix**	diagon
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

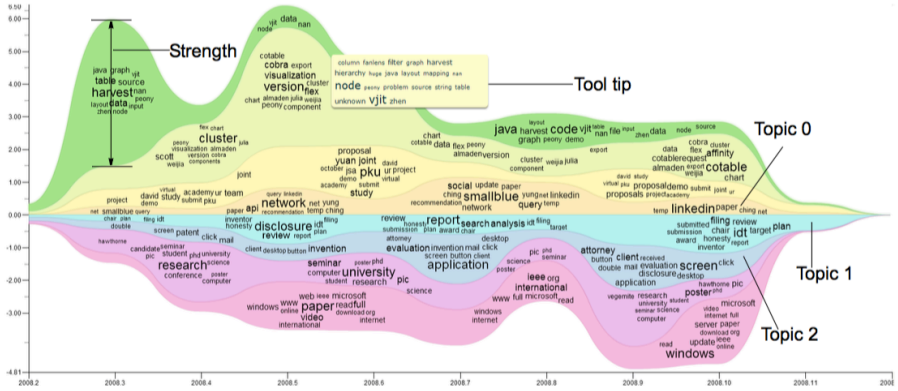
Outline

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

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)

Topic dynamics

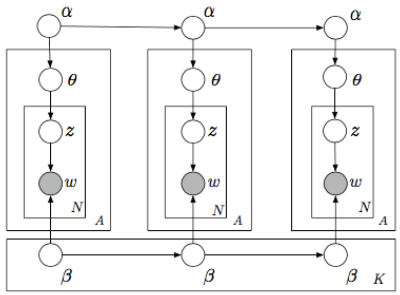


TIARA (Liu et al., 2009) (Wei et al., 2010)

Dynamic Topic Model

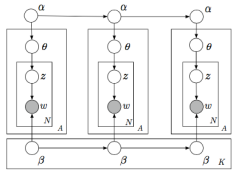
(Blei and Lafferty, 2006)

- Graphical model:



- Links inspired by Brownian motion

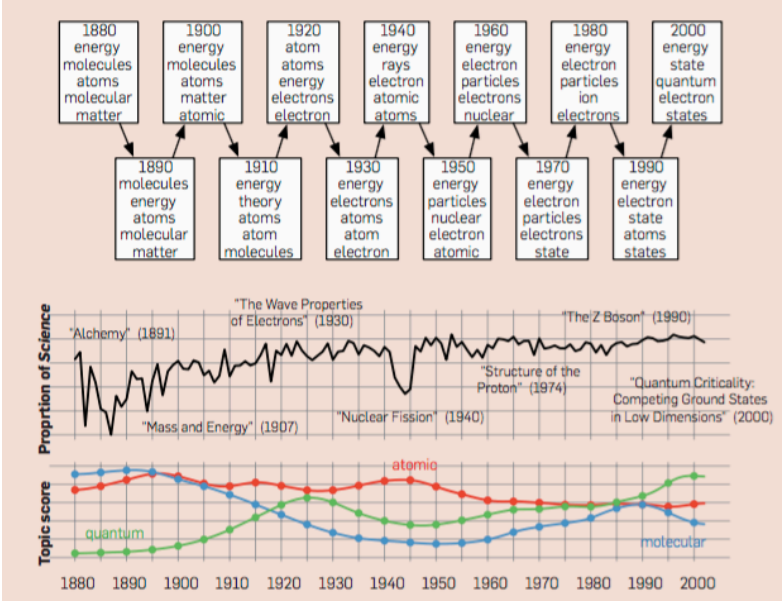
Generative process



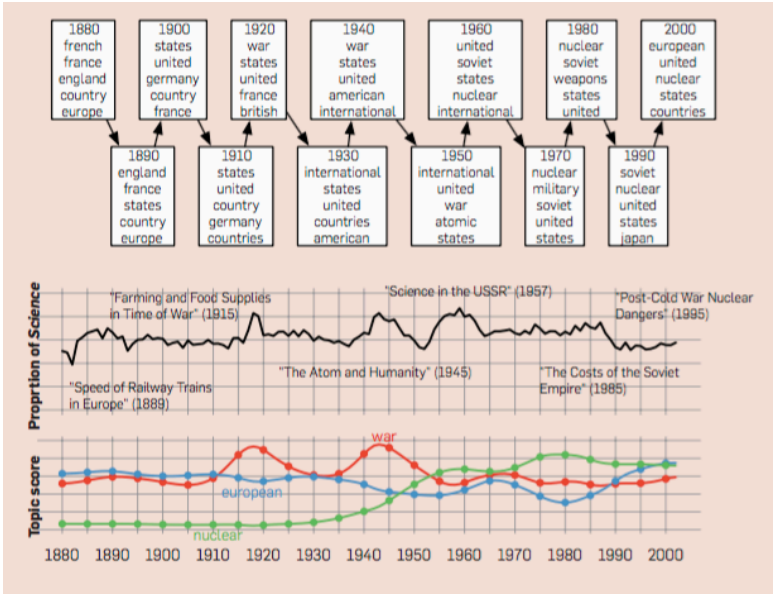
Brownian motion here

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

from natural parameters to normalized parameters



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



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

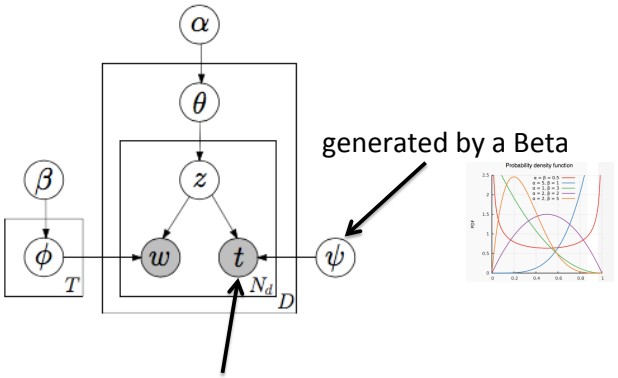
And for “our” scientific articles?

Excerpt for the “database” topic:

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

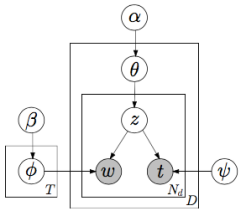
Topic Over Time

(Wang et McCallum, 2006)



Temporal dimension as an observed random variable

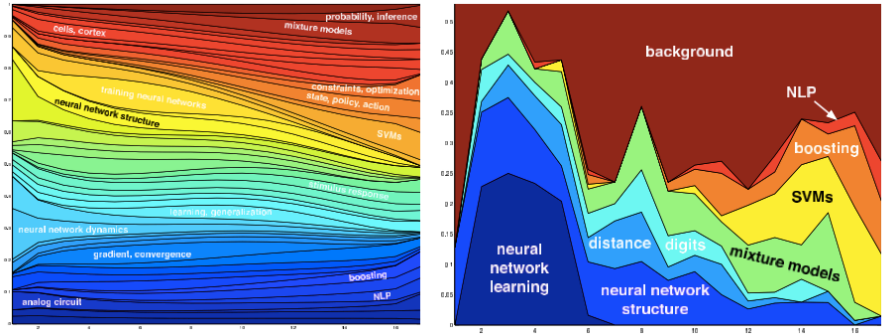
Generative process



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 $\psi_{z_{di}}$.

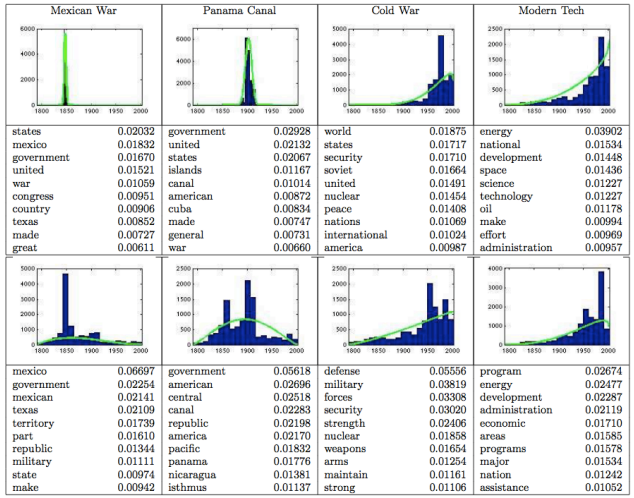
yes, two words of the *same* document can be associated to *different* timestamps!

Results on scientific trends



Data extracted from the proceedings of NIPS between 1987 and 2003

Results on the state of the union

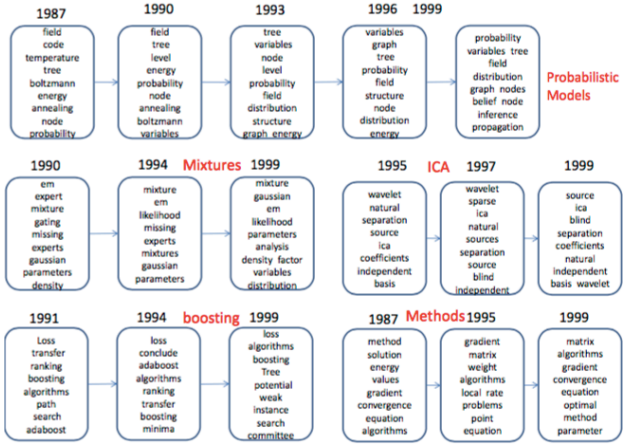


Mexican War		Panama Canal		Cold War		Modern Tech	
states	0.02032	government	0.02928	world	0.01875	energy	0.03902
mexico	0.01832	united	0.02132	states	0.01717	national	0.01534
government	0.01670	states	0.02067	security	0.01710	development	0.01448
united	0.01521	islands	0.01167	soviet	0.01664	space	0.01436
war	0.01059	canal	0.01014	united	0.01491	science	0.01227
congress	0.00951	american	0.00872	nuclear	0.01454	technology	0.01227
country	0.00906	cuba	0.00834	peace	0.01408	oil	0.01178
texas	0.00852	made	0.00747	nations	0.01069	make	0.00994
made	0.00727	general	0.00731	international	0.01024	effort	0.00969
great	0.00611	war	0.00660	america	0.00987	administration	0.00957
mexico	0.06697	government	0.05618	defense	0.05556	program	0.02674
government	0.02254	american	0.02696	military	0.03819	energy	0.02477
mexican	0.02141	central	0.02518	forces	0.03308	development	0.02287
texas	0.02109	canal	0.02283	security	0.03020	administration	0.02119
territory	0.01739	republic	0.02198	strength	0.02406	economic	0.01710
part	0.01610	america	0.02170	nuclear	0.01858	areas	0.01585
republic	0.01344	pacific	0.01832	weapons	0.01654	programs	0.01578
military	0.01111	panama	0.01776	arms	0.01254	major	0.01534
state	0.00974	nicaragua	0.01381	maintain	0.01161	nation	0.01242
make	0.00942	isthmus	0.01137	strong	0.01106	assistance	0.01052

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 evolutinary clustering (Xu et al., 2008)
  - Infinite DTM (Ahmed and Xing, 2010)

Infinite DTM



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

Some personal conclusions

- Positive aspects of probabilistic approaches
  - "clean" dependency modeling (even with heterogeneous data)
  - techniques for parameter estimation in reasonable time
- Some difficulties on the road
  - posterior estimation and inference
  - model interpretation and model checking
  - same issues than with clustering (e.g., granularity, number k)
- Personal perspectives
  - topic labeling and summary
  - links with word embedding
  - non parametric models
  - temporal evolution and change points

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/llt.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