Résumé de vastes corpus textuels : quelques cas d'application aux données ouvertes

Julien Velcin – Laboratoire ERIC
http://mediamining.univ-lyon2.fr/velcin

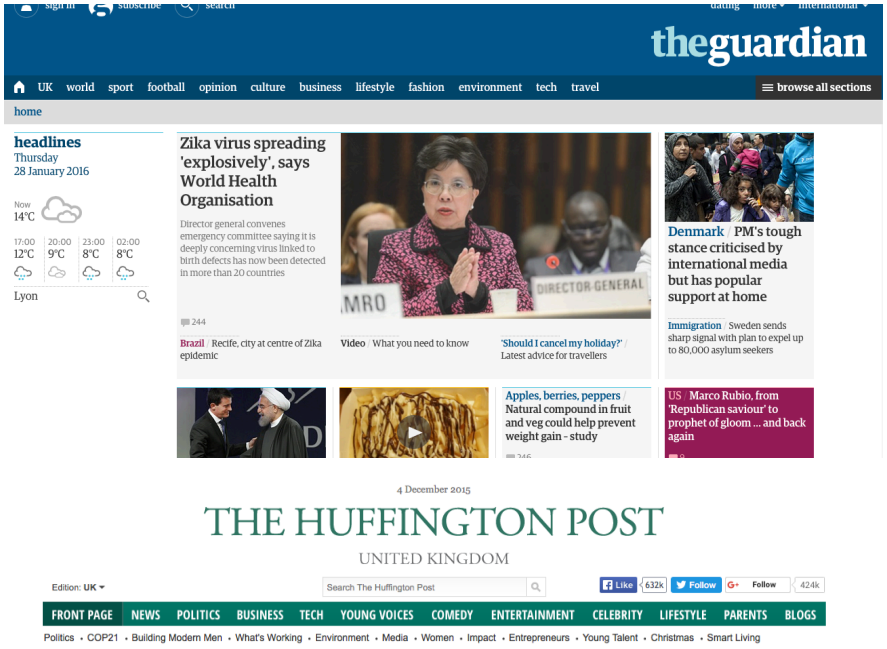
Journées « Humanités numériques et données ouvertes »
24 et 25 mai 2016 – Lyon

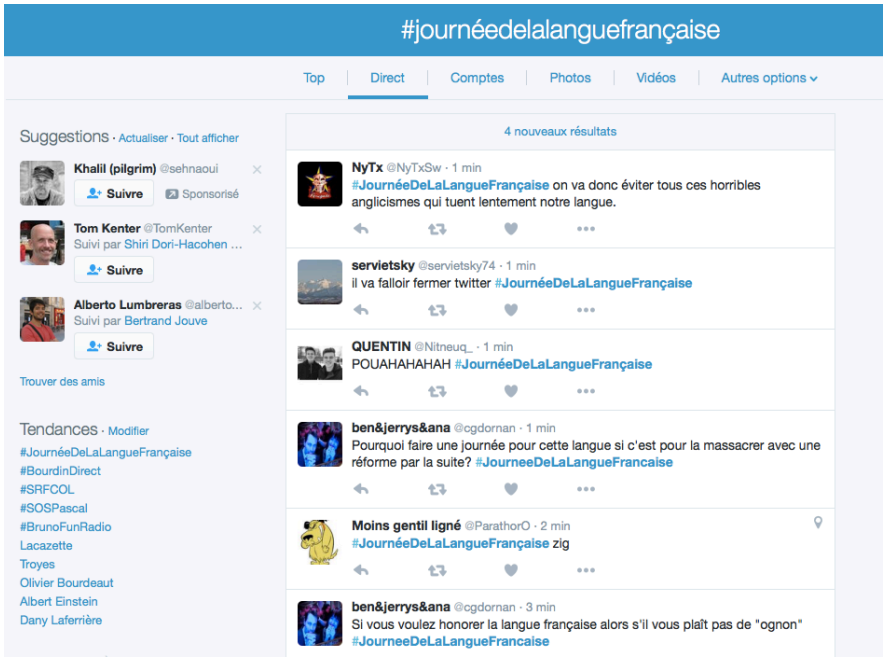
Plan de la présentation

- Résumer par des modèles thématiques
- Illustration sur plusieurs cas d'étude
- Aller plus loin : cas des données temporelles
- Conclusion et perspectives

Plan de la présentation

- **Résumer par des modèles thématiques**
- Illustration sur plusieurs cas d'étude
- Aller plus loin : cas des données temporelles
- Conclusion et perspectives



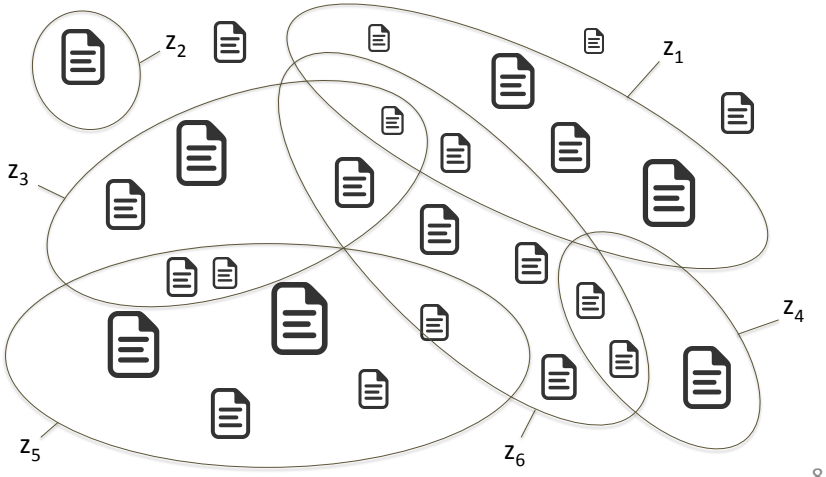


Apprendre les thématiques ?

- Les modèles de thématiques ont pour objectif d'organiser **automatiquement** de vastes collections de données textuelles
  - découvrir les thèmes qui « émergent »
  - indexer les documents à l'aide de ces thèmes
  - utiliser ces annotations pour résumer, naviguer, rechercher… donc mieux appréhender les données

Clustering vs. thématiques

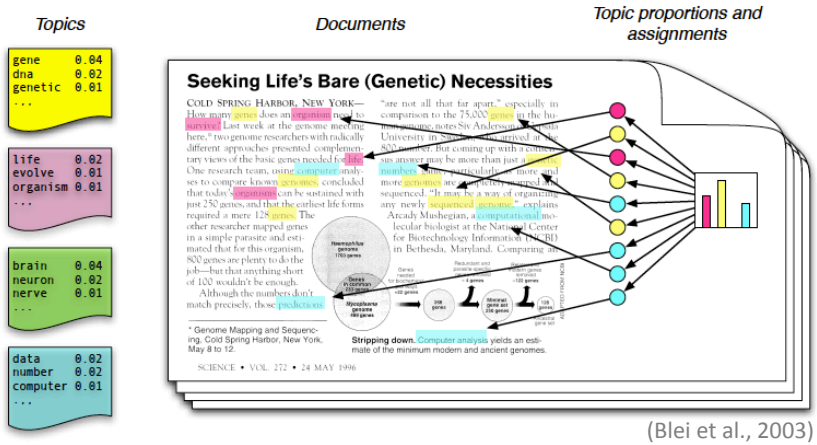
(Xie and Xing, 2013)



Eléments d'état de l'art

- Approches algébriques
  - LSA (Deerwester et al., 1990)
  - NMF (Paatero et Tapper, 1994)
  - apprentissage de dictionnaires (Jenatton et al., 2010)
- Approches géométriques
  - issus du TDT (Allan et al., 1998) (Pons-Porrata et al., 2003)
  - AGAPE (Velcin and Ganascia, 2007)
- Approches probabilistes
  - pLSA, LDA… (voir la suite de l'exposé)

Modèle de thématiques : cas de LDA



(Blei et al., 2003)

Manipuler des données textuelles

- Hypothèse du sac de mots
- Etape de prétraitements :
  - suppression des « mots-outils »
  - suppression des marques de ponctuations
  - suppression des nombres

- En entrée :

	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

Plan de la présentation

- Résumer par des modèles thématiques
- **Illustration sur plusieurs cas d'étude**
- Aller plus loin : cas des données temporelles
- Conclusion et perspectives

Illustration avec quelques jeux de données (oui, qui ne sont pas tous libres…)

- Articles scientifiques
- 20 Newsgroups

en collaboration avec P. Poncelet, M. Roche et J.A. Lossio (LIRMM)

- Série Harry Potter

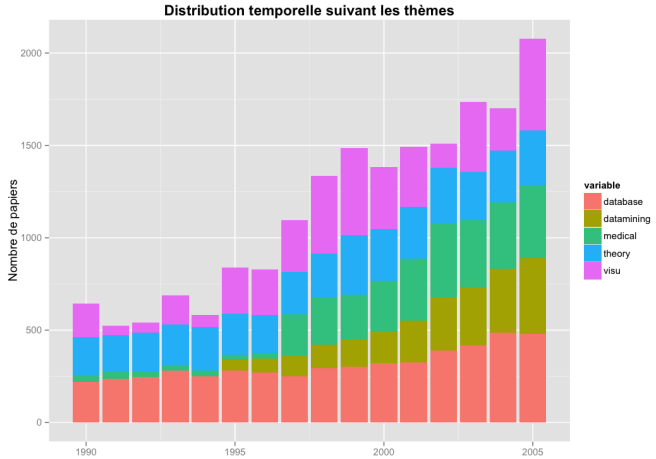
en collaboration avec C. Gravier (LHC)

Articles scientifiques

+ de 18000 titres et/ou résumés d'articles publiés entre 1990 et 2005 (Tang et al., 2012)

- base de données : ICDE, VLDB, SIGMOD…
- data mining (après 1994) : KDD, ICDM…
- visualisation : CVPR, InfoViz, ICCV…
- informatique théorique : FOCS, SODA…
- informatique médicale : JAMIA, AIME…

Evolution des thématiques



Thématiques extraites par LDA (vocabulaire de 5000 mots)

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…

Thématiques extraites par LDA
(vocabulaire de 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 textes répartis dans 20 catégories
  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?

catégorie :
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

catégorie :
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

catégorie :
sci.space
(suite)

Thématiques extraites par LDA (vocabulaire de 10000 mots)

Extrait des 20 thématiques demandées :

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…

Et dans Harry Potter

extrait sur
20 thématiques :

Maisons de sorcellerie
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

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

Professeurs de Poudlard
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

Mais aussi…

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

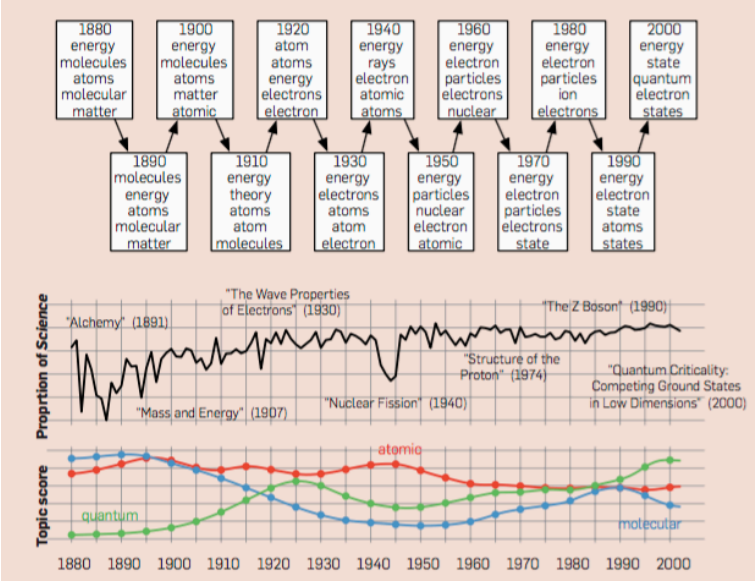
hagrid 0.06322251
yeh 0.06136366
ter 0.04835173
yer 0.03719865
said 0.02170826
dragon 0.01861018
fer 0.01799056
gringotts 0.01737095
got 0.01675133
don 0.01613172

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

Plan de la présentation

- Résumer par des modèles thématiques
- Illustration sur plusieurs cas d'étude
- **Aller plus loin : cas des données temporelles**
- Conclusion et perspectives

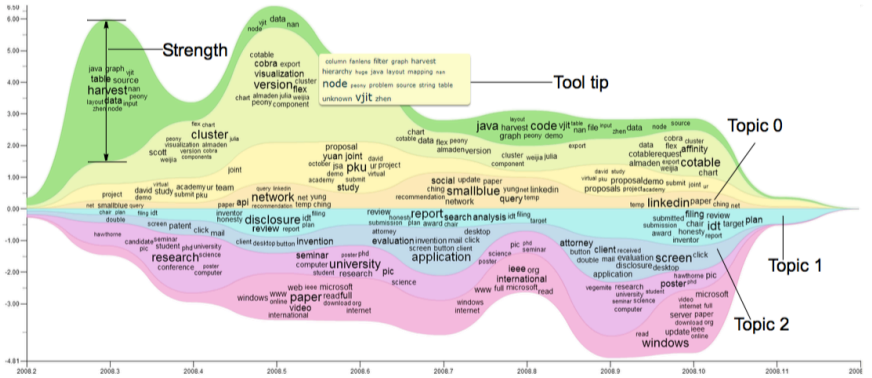


données tirées de *Science* entre 1880 et 2002 (Blei, 2012)

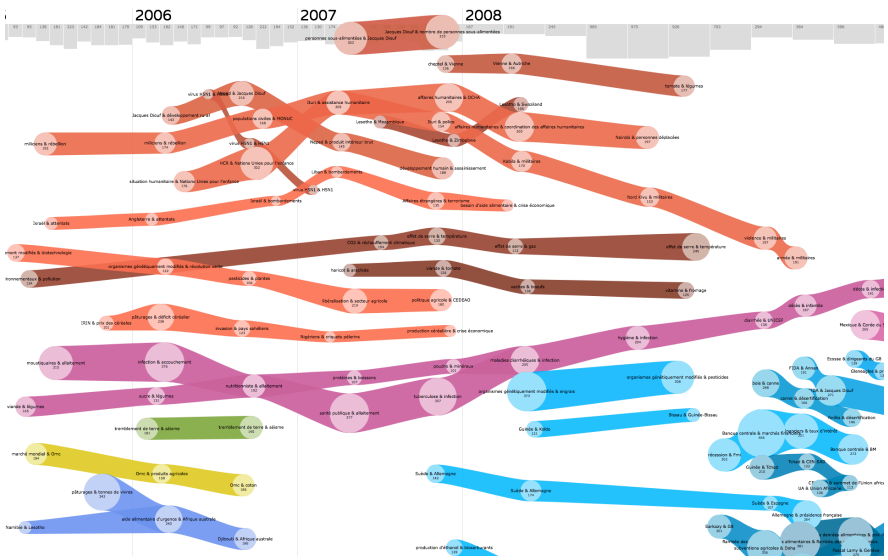
Résultats sur le discours de l'union

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

Dynamique des thématiques



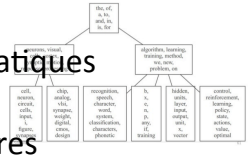
TIARA (Liu et al., 2009) (Wei et al., 2010)



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

Quelques questions intéressantes

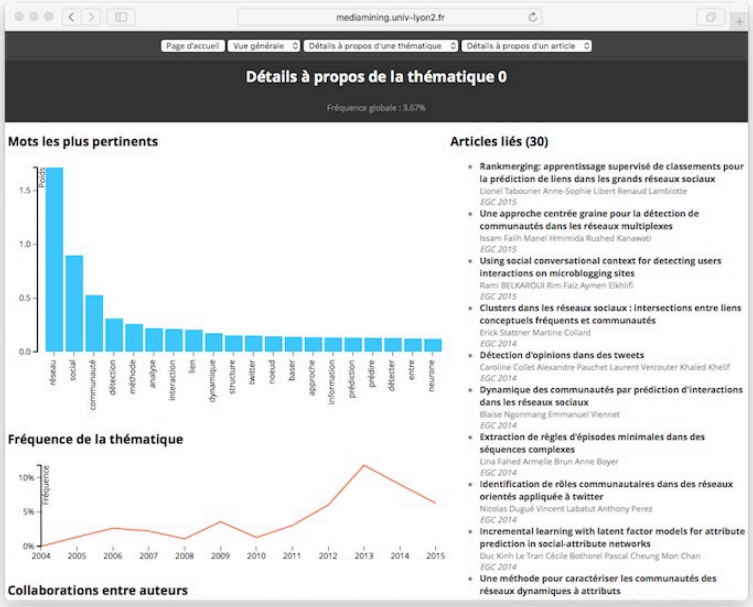
- Trouver le « meilleur » nombre de thématiques
  HDP (Teh et al., 2004)
- Résumer l'information portée par une thématique
  Topic labeling (Mei et al., 2007)
- Trouver une structure entre les thématiques
  hLDA (Griffiths et al., 2004) CTM (Blei et al., 2006)
- Combiner les thématiques avec d'autres informations : auteur, opinion, structure, etc.
  Author-Topic (Rosen-Zvi et al., 2006)
  Topic-Opinion (Dermouche et al., 2014) (et tant d'autres…)
- Liens avec le « plongement de mots » (Das et al., 2015)



Plan de la présentation

- Résumer par des modèles thématiques
- Illustration sur plusieurs cas d'étude
- Aller plus loin : cas des données temporelles
- **Conclusion et perspectives**

TOM @ERIC: http://mediamining.univ-lyon2.fr/people/guille/egc2016/



Références orientées DH

- Blei, D.M., A.Y. Ng and M. I. Jordan (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: pp. 993–1022.
- Burton M. (2013). The Joy of Topic Modeling: A bag of words. http://mcburton.net/blog/joy-of-tm
- Goldstone A. and T. Underwood (2012). What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship? http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone
- Graham S. et al. (2012). Getting Started with Topic Modeling and MALLET. Online lesson. http://programminghistorian.org/lessons/topic-modeling-and-mallet
- Weingart, S. (2012). Topic Modeling for Humanists: A Guided Tour. http://www.scottbot.net/HIAL/?p=19113