Analyser le mediascape : retour sur un projet pluridisciplinaire

Julien Velcin
Laboratoire ERIC
Université de Lyon
julien.velcin@univ-lyon2.fr
@jvelcin

Outline of my talk

1) Exploring the informational landscape
2) Topic modeling as a major tool in NLP
3) Studying the mediascape
4) A « user in the loop » perspective
5) Conclusion in a DH point of view

Outline of my talk

1) **Exploring the informational landscape**
2) Topic modeling as a major tool in NLP
3) Studying the mediascape
4) A « user in the loop » perspective
5) Conclusion in a DH point of view

theguardian

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

home

headlines
Thursday
28 January 2016

Now 14°C

17:00	20:00	23:00	02:00
12°C	9°C	8°C	8°C

Lyon

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

244

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

Brazil / Recife, city at centre of Zika epidemic

Video / What you need to know

'Should I cancel my holiday?' / Latest advice for travellers

4 December 2015

THE HUFFINGTON POST

UNITED KINGDOM

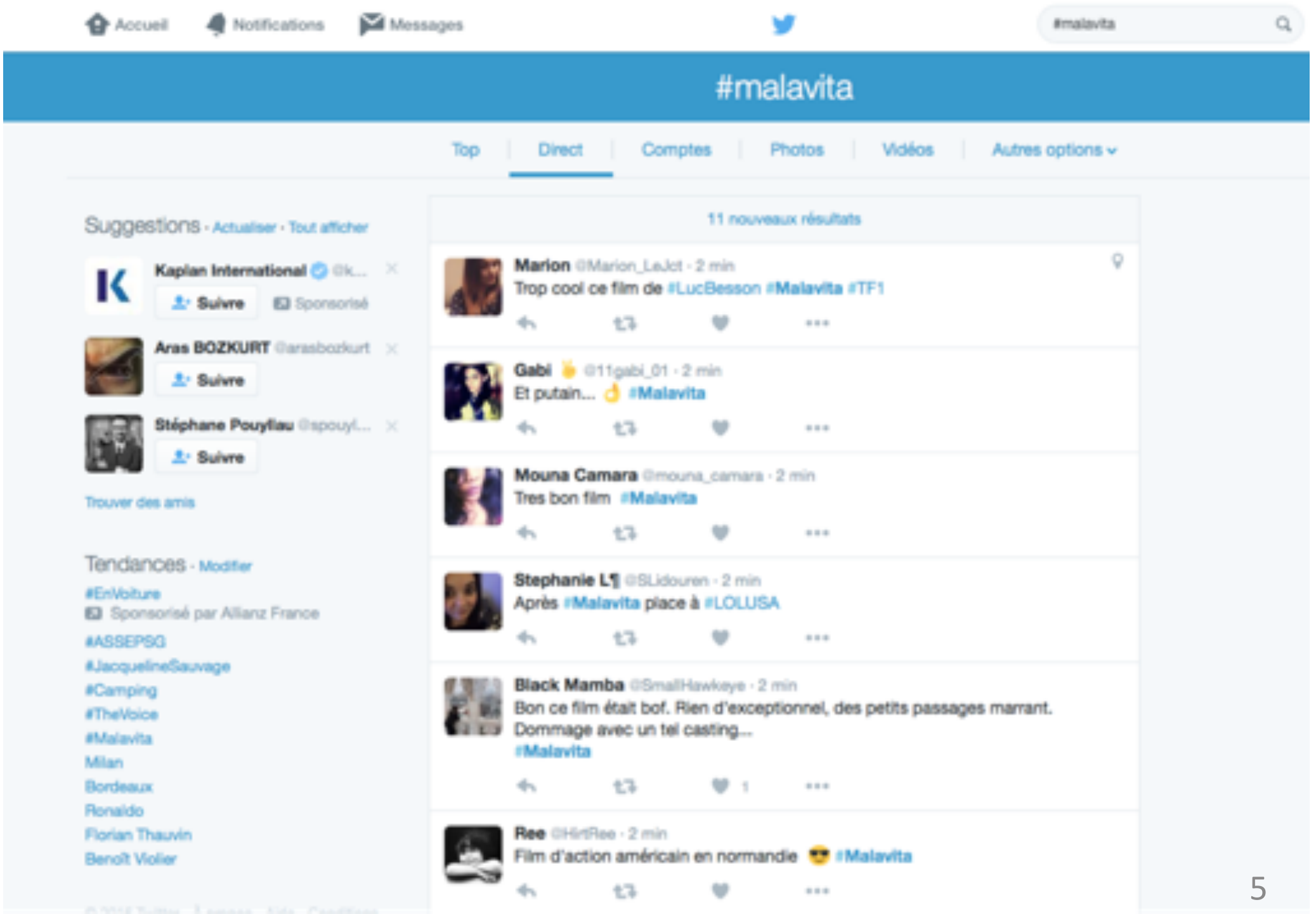
Edition: UK

Search The Huffington Post

Like 632k Follow Follow 424k

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


Accueil
Notifications
Messages
#malavita
#malavita
Top
Direct
Comptes
Photos
Vidéos
Autres options
Suggestions · Actualiser · Tout afficher
Kaplan International @k...
Suivre
Sponsorisé
Aras BOZKURT @arasbozkurt
Suivre
Stéphane Pouyllau @spouyl...
Suivre
Trouver des amis
Tendances · Modifier
#EnVoiture
Sponsorisé par Allianz France
#ASSEPSG
#JacquelineSauvage
#Camping
#TheVoice
#Malavita
Milan
Bordeaux
Ronaldo
Florian Thauvin
Benoît 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

Exploring textual corpora

- Top-down approaches
  - regexp (e.g., “terror[a-z]*”)
  - keyword-based queries
  - (supervised) classification schemas
- Bottom-up approaches
  - basic statistics
  - projection to low-level spaces (e.g., PCA)
  - structuration by (co-)clustering or topic modeling
  - More advanced summarization techniques (e.g., metromaps)
- Mixed approaches
  - weakly-supervised clustering
  - active learning schemas

Outline of my talk

1) Exploring the informational landscape
2) **Topic modeling as a major tool in NLP**
3) Studying the mediascape
4) A « user in the loop » perspective
5) Conclusion in a DH point of view

Some background

- Bag-of-words assumption
- Usual preprocessing
  - remove numbers and punctuation
  - remove 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

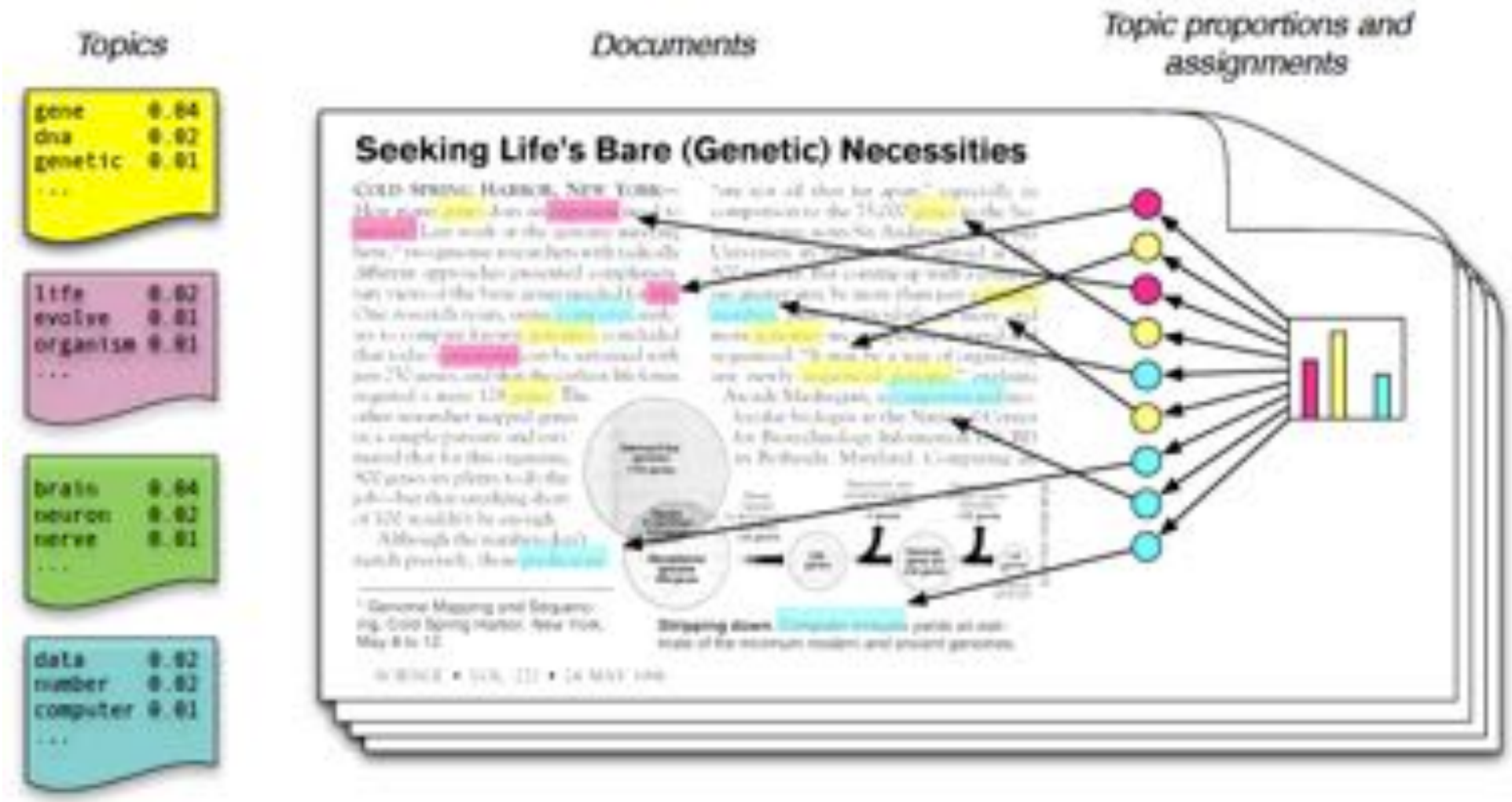
Why topic learning?

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

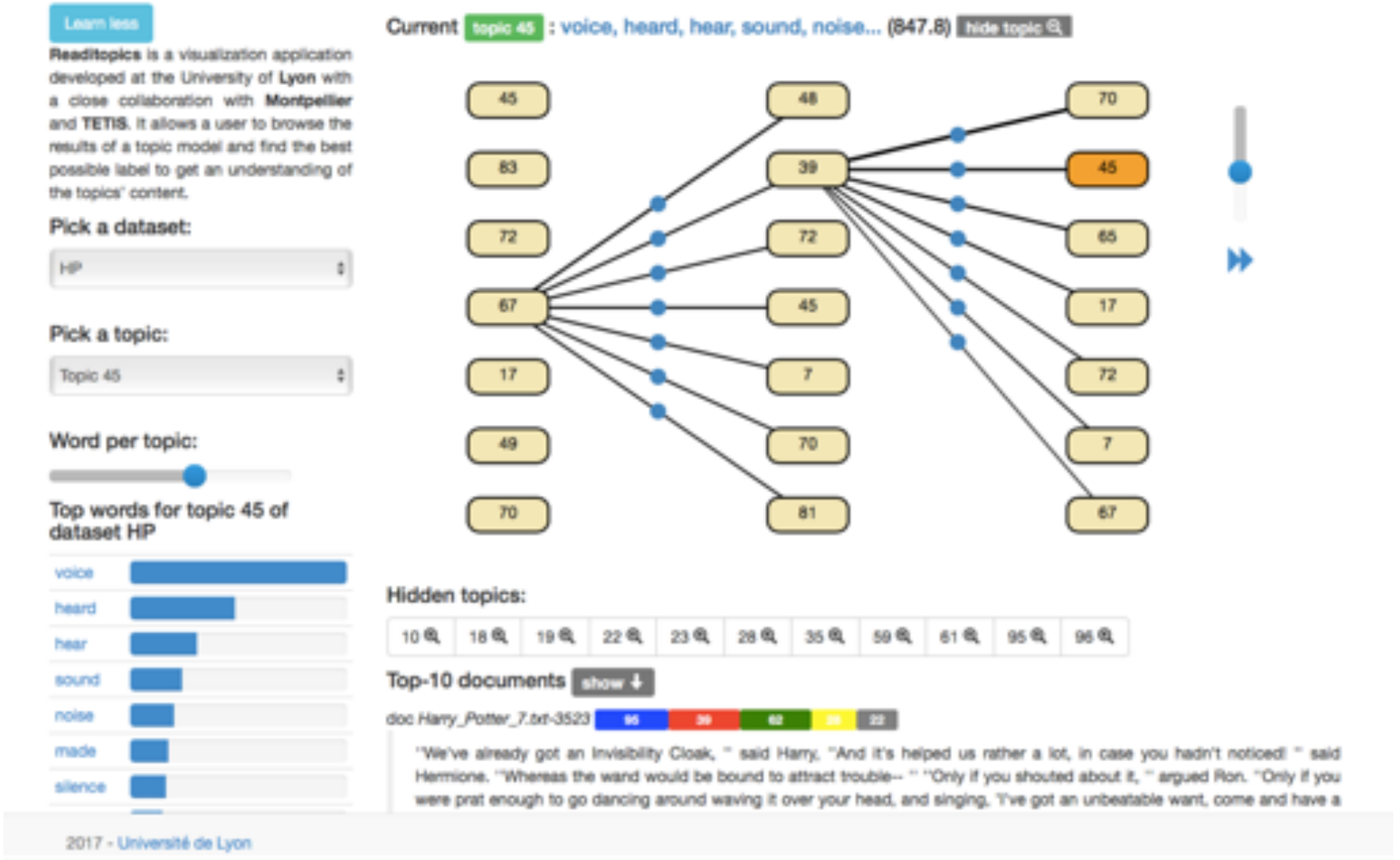
(some slides are taken from the talk of D. Blei for KDD 2011)

Discovering latent structures



(Blei et al., 2003)

ReadiTopics (Velcin et al., 2018)

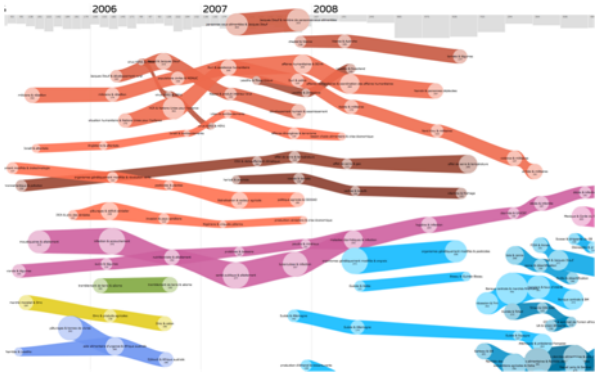


Outline of my talk

1) Exploring the informational landscape
2) Topic modeling as a major tool in NLP
3) **Studying the mediascape**
4) A « user in the loop » perspective
5) Conclusion in a DH point of view

What’s a « mediascape »?

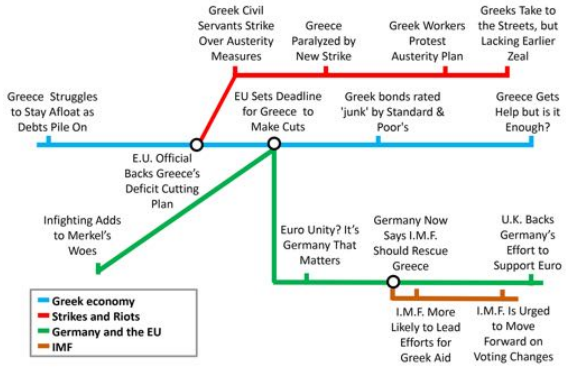
Appadurai A.: Disjuncture and Difference in the Global Cultural Economy, Theory Culture Society 1990; 7; 295

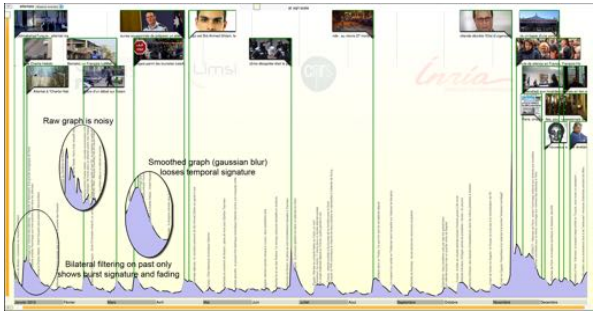


Projet Pulseweb
(Cointet, Chavalarias…)

http://pulseweb.cortext.net

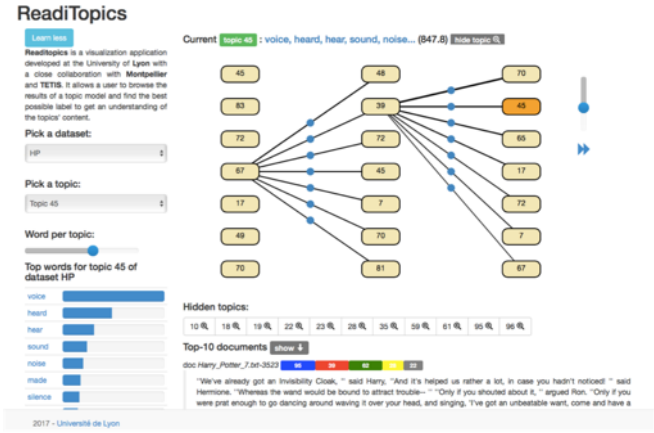
Metromaps
(Shahaf et al., 2015)





Chronolines
(Nguyen et al., 2014)

Readitopics
(Velcin et al., 2018)



https://github.com/Erwangf/readitopics

What we did

- Joint work with a sociologist (J.C. Soulages, Max Weber lab), in a project related to data journalism (Velcin et al., workshop @EGC 2017)
- As input: a collection of documents (here, newspapers from the **Huffington Post**)
- As output: distribution over topic categories
- Two levels of categories:
  - basic level, found by the topic model (here, LDA)
  - high level, labeled by experts (here, J.C. Soulages and partners from Brazil)

Comparing news media

- Usual preprocessing (tokenisation, stopwords…)
- Three versions of the same media (HuffPost):

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?

  -> associate each topic with one **given** category (e.g., sport or media)

  -> up to now, this is manual!

- Estimate the **importance** of every category (here, volume of words tagged by the covered topics)

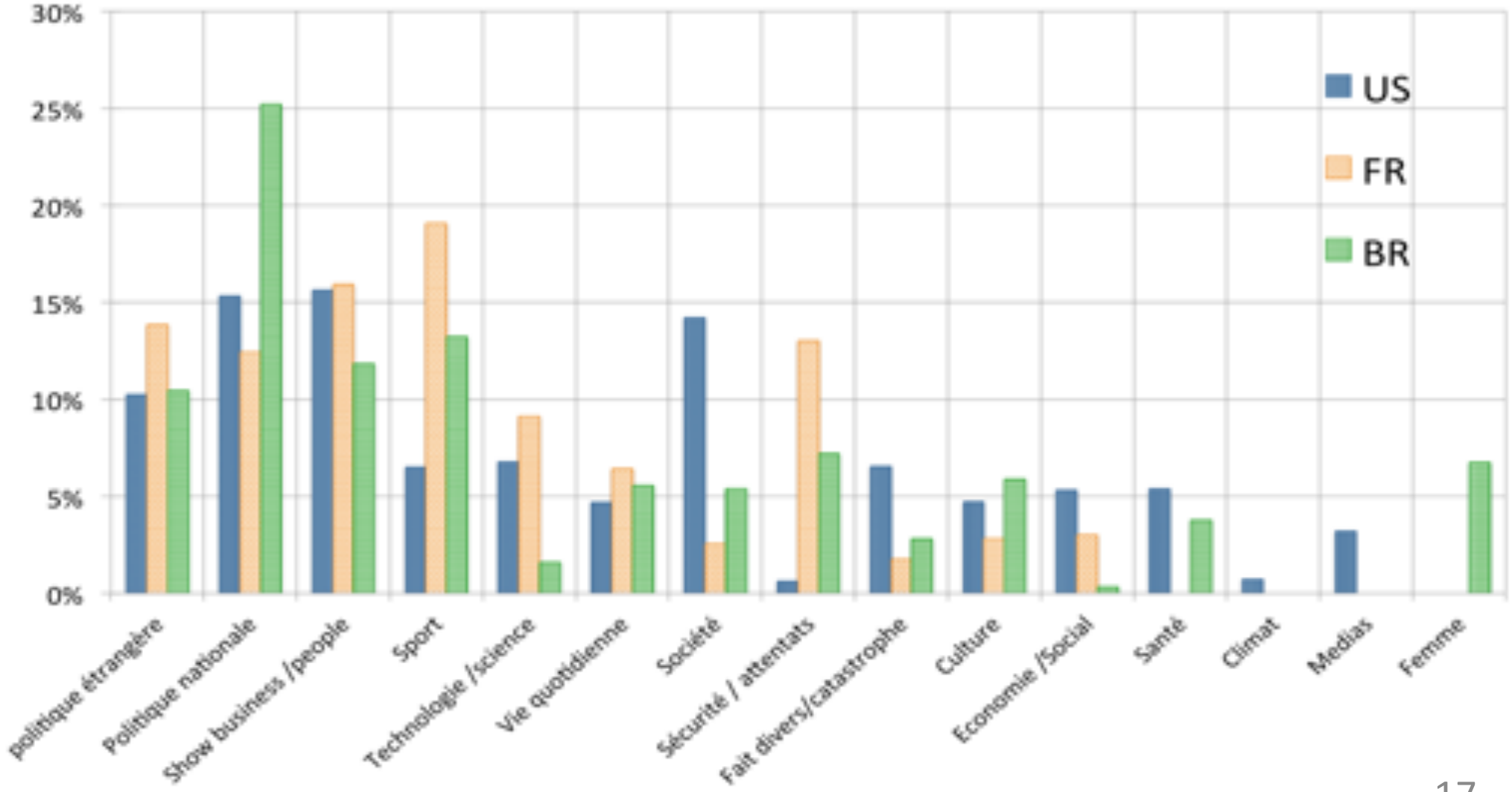
Some topics extracted by LDA

topic	#doc	cat.	mots les plus probables
			en français (sur 4133 articles) :
z_{18}	28	1	manifestation, paris, police, travail, loi, contre, syndicats, place, bastille, 2016
z_{19}	36	1	loi, travail, gouvernement, l'état, texte, l'assemblée, d'urgence, mois, projet, conseil
z_{25}	39	2	jeux, rio, olympiques, olympique, août, jo, athlètes, 2016, brésil, cérémonie
z_{47}	18	3	morandini, jean-marc, inrocks, catherine, l'animateur, lui, qu'il, europe, comédiens, plainte
z_{73}	47	4	nice, 14, l'attentat, anglais, promenade, camion, attentat, police, soir, christian
			en anglais (sur 12067 articles) :
z_{14}	92	5	refugees, children, refugee, people, countries, world, syrian, rights, million, year
z_{21}	74	2	gymnastics, biles, olympic, team, simone, olympics, gymnast, gold, rio, hernandez
z_{3}	46	6	pokemon, game, pokémon, playing, players, catch, «pokemon, go», pizza, play
z_{50}	56	7	muslim, religious, muslims, faith, church, god, christian, religion, hate, american
z_{27}	140	8	clinton, voters, trump, poll, polls, americans, election, support, vote, relationships
			en portugais (sur 2355 articles) :
z_{44}	52	8	dilma, presidente, impeachment, senado, senadores, processo, senador, rousseff, julgamento, defesa
z_{58}	7	9	sexo, menstruação, durante, rao, mccane, comédia, realmente, corpo, riso, menstruada
z_{71}	11	7	negros, brancos, negras, pessoas, racial, negra, racismo, país, movimento, black
z_{37}	57	2	brasil, vôlei, jogo, medalha, vitória, ouro, seleção, set, brasileiras, torcida
z_{99}	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.

Compared results

Normalized distribution over the 15 categories
(*remember that* each category can be associated to **multiple** topics)



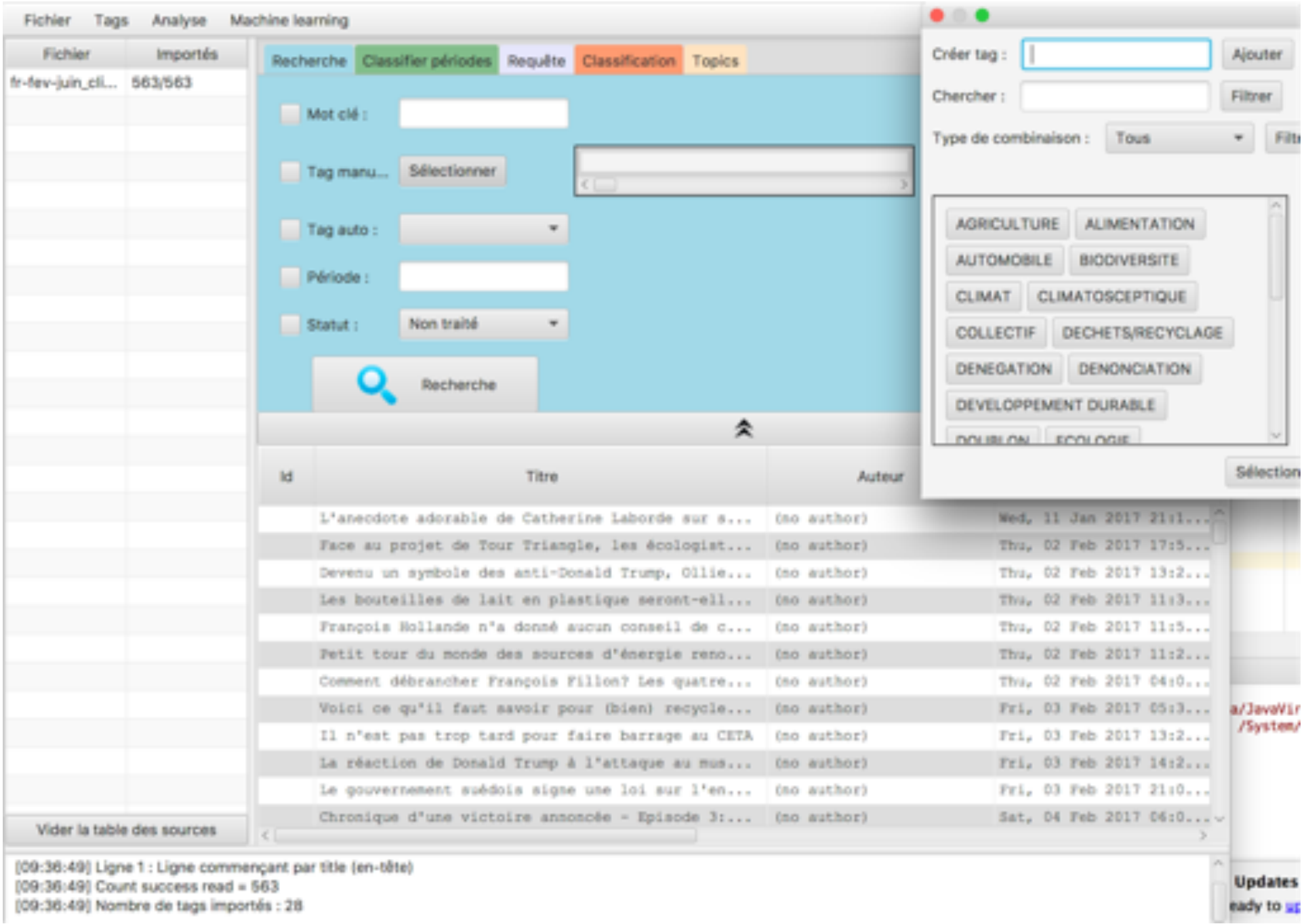
Outline of my talk

1) Exploring the informational landscape
2) Topic modeling as a major tool in NLP
3) Studying the mediascape
4) **A « user in the loop » perspective**
5) Conclusion in a DH point of view

Human in the loop for topic models

- Give more insight about the produced structure
  - work on topic coherence (Röder et al., 2015)
  - topic labeling (Mei et al., 2007) (Gourru et al., 2018)
- Let the user interfere with the model
  - interactive topic modeling (Hu et al., 2014)
  - make the model selection easer
- **Towards learning the (meta) categories**
  - **semi-supervised, active learning techniques**
  - **refine the categorical structure with humans!**

Newsbrowsers



Outline of my talk

1) Exploring the informational landscape
2) Topic modeling as a major tool in NLP
3) Studying the mediascape
4) A « user in the loop » perspective
5) **Conclusion in a DH point of view**

Take-home message

- For studying the informational landscape:
  - semi-supervised / active learning is the key
  - clear idea of the targeted application
  - lots of engineering work (needs funding)
- Collaboration with LLSSH needs:
  - mutual understanding
  - trust
  - time
- And by the way…
  - life-cycle of information (Davidson et al., 2020)

References

(Appadurai A., 1990) Appadurai A.: Disjuncture and Difference in the Global Cultural Economy, **Theory Culture Society**, 1990; 7; 295.

(Blei et al., 2003) Blei, D. M., Ng, A. Y., & Jordan, M. I.: Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022, 2003.

(Davidson et al., 2020) Davidson I., A. Gourru, J. Velcin and Y. Wu: Behavioral differences: insights, explanations and comparisons of French and US Twitter usage during elections. Social Network Analysis and Mining (**SNAM**), 10: 6.

(Gourru et al., 2018) Gourru, A., Velcin, J., Roche, M., Gravier, C., & Poncelet, P.: United we stand: Using multiple strategies for topic labeling. In International Conference on Applications of Natural Language to Information Systems (**NLDB**), pp. 352-363, 2018, Springer.

(Velcin et al. 2017) Velcin J., J.C. Soulages, S. Kurpiel, D.L. Otavio, M. Del Vecchio et F. Aubrun: Fouille de textes pour une analyse comparée de l'information diffusée par les médias en ligne : une étude sur trois éditions du Huffington Post. **Atelier Journalisme computationnel** @EGC, 2017.

(Velcin et al., 2018) Velcin, J., Gourru, A., Giry-Fouquet, E., Gravier, C., Roche, M., & Poncelet, P.: Readitopics: make your topic models readable via labeling and browsing. Proceedings of the 27th International Joint Conference on Artificial Intelligence (**IJCAI** 2018, demo track), pp.5874-5876, Stockholm, Sweden.