

Categorization and machine learning

Julien Velcin
<http://eric.univ-lyon2.fr/jvelcin>

ERIC Lab, Université Lyon 2

Séminaire doctoral, Université de Lyon

Context and motivation

Julien Velcin – ERIC Lab, Université Lyon 2 – Séminaire Udl 2022

3

Outline of the talk

- Context of the ERIC Lab / Lyon 2
- Dilemma of categorization
- Illustration with various partners
 - once upon a time, ImagiWeb (political science)
 - topic models as « explanation » of given categories (health)
 - studying the informational landscape (information science)
- Some lessons I've learnt
- Ongoing work

2

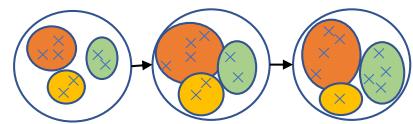
Some context

- **ERIC Lab:** Univ. Lyon 1 + Lyon 2 <http://eric.msh-lse.fr>
(some keywords: data science, machine learning, business intelligence, social media analysis, digital humanities...)
- Two teams: SID and DMD
- The lab is a member of **MSH-LSE** <https://www.msh-lse.fr>
- Many applications to Social Sciences and Humanities
(projects in literature, political sciences, archeology...)

4

Dilemma of categorization

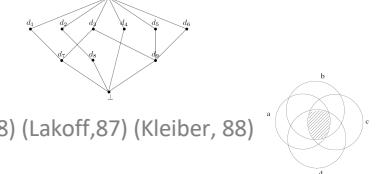
- Origin of (natural) categorization not very well known (kind of chicken and egg problem)
- Previous works in psychology, linguistics, cognitive science
- It boils down to knowing what comes from **previous knowledge** (can be seen as an *a priori*) and from the **data**
- Relates to « human in the loop » / interactive approaches
- Illustration with temporal data (Christophe et al., 2021)



5

Some categorization theories

- **Logical theory** (Aristotle)
- **Prototypes** (Rosch,73,75,78) (Lakoff,87) (Kleiber, 88)
- **Stereotypes** (Lippman,22)
- **Family resemblance** (Wittgenstein,53)



6

Once upon a time: the ImagiWeb project

Julien Velcin – ERIC Lab, Université Lyon 2 – Séminaire Udl 2022

7

ImagiWeb project

- Studying the image (representation) of entities emitted from the social media and its evolution over time (Velcin et al., 2014) (Boyadjian et Velcin, 2017)
 - Aspects:
 - Political line
 - Future project
 - Balance sheet
 - Ethic
 - Injunction
 - Communication
 - etc.
- ANR grant (2012-2015)



8

Collaborative project



9

Structuration of the project



Data acquisition:

- Case study n°1: image of two French politicians (10M tweets from March 2012 til Jan. 2013 + public polls + socio-demographic meta-data)
- Case study n°2 : image of the EDF company (9k blog posts from Jan. 2011 til Sept. 2012)

Guide & platform open source for the annotation

Manual annotation of a subset:

- Case study n°1: 11,527 annotations of 7,283 tweets (20 annotators, between 1 and 3 per tweet)
- Case study n°2: 600 blog posts (passages) (2 annotators)

10

The screenshot shows the IMAGIWEB annotation interface with the following details:

- Guide rapide:**
 - 998 textes à annoter. Aller au texte : 5726 : 36 textes déjà annotés. Revoir le texte : Aucun :
 - Sélectionnez le commentaire avec la souris.
 - Atribuez une polarité d'opinion en appuyant sur une des 5 catégories proposées pour l'annotation à droite.
 - Représentez la cible du commentaire sélectionné et écrivez-la dans le champ de texte réservé à gauche.
- Cibles:**
 - Personne/vie privée
 - Aucune
 - Attribut:Sondage
 - Attribut:Sujetien
 - Attribut:Autre
 - Bilan:Ecologie
 - Bilan:Economie
 - Bilan:Sociétal
 - Bilan:Autre
 - Compétence:Expertise
 - Compétence:Gouverner
 - Compétence:Autre
 - Ethique:Affaire
 - Ethique:Honnêteté
 - Ethique:Autre
 - Injonction
 - Performance:Prestations
 - Performance:Global
 - Performance:Autre
 - Personne:Apparence
- Opinion:**
 - Concerne l'image.
 - Ne concerne pas l'image.
- A propos du système:**
 - By http://twitter.com/PierreCourade/status/263711917636460544 (Image de: Hollande) N°5726 de PierreCourade le 31/10/2012: @Francetv2012 "François Hollande préfère rester en... @Francetv2012 "François Hollande préfère rester en retrait" ?! Il ne s'est pas précipité à Grenoble ? Il ne s'invite pas dans les médias ?!"
 - Confiance assurée des annotations
 - Confiance faible des annotations
- JSON :** {pertinence:concerne,confiance:assuree,ambigu:{}},trespositif:{},positif:{},"Hollande préfère rester en retrait"},neutre:{},negatif:{1:"Il ne s'invite pas dans les médias"},tresnegatif:{}}

11

Guide and categories for the annotation (excerpt)

Guide d'annotation
Version du 12/10/12
Auteur : J. Vélin

Public concerné

Ce document est destiné aux personnes qui participent à l'annulation ImaginWeb. Il sera notamment utilisé lors des journées d'organisées à Montpellier.

Corpus ciblés

Les deux corpus sur lesquels porte l'annotation ont été extraits de la plate-forme fournie par AMI Software. L'extraction a été paramétrée en place par les partenaires utilisateurs, à savoir le CEPEL et l'ERIC, comme des nécessaires d'opinion sur les hommes politiques et les campagnes présidentielles en France (et en particulier Sarkozy). Le corpus d'EDF concerne le thème des énergies et les corpus comportent à la fois des textes issus de blogs (textes pour tweets (messages limités à 140 caractères). Normalement, ils sont français, et ils peuvent comporter des fautes d'orthographes, de tirs et des hashtags (pour Twitter), etc.

Les annotations

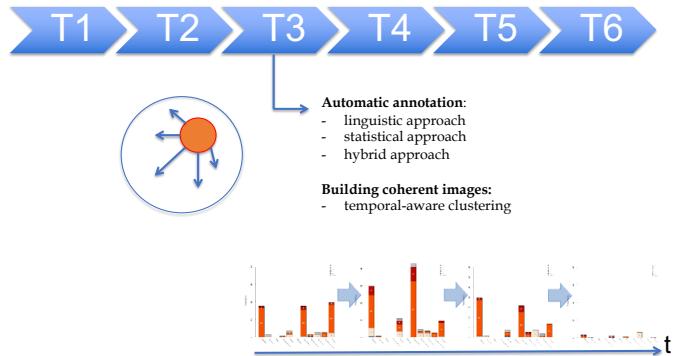
Les annotations prévues sont de deux types :

- Annotation des marqueurs d'opinion : il s'agit de définir si l'exemple une phrase, ou une partie de phrase) exprime un tonalité suivante (commentée plus loin dans ce tableau).
 - a. Xxx +2 (très positif)
 - b. Xxx +1 (positif)
 - c. 0 (neutre, pas d'opinion exprimée)
 - d. -1 (négatif)
 - e. Xxx -2 (très négatif)

- Annotation de la cible de l'opinion : il s'agit de définir un mot qui constitue la cible sur laquelle porte l'opinion. Il faut pas que ce mot soit analysé (ici EDF, ou un homme politique). Il s'agit d'attributs, propriétés qui caractérisent l'entité, comme par exemple un « producteur d'électricité », que l'entreprise de photovoltaïque, ou qu'elle même des « chantiers ». Pour

		envers le candidat Hollande qu'il soutient selon toute vraisemblance.
		Opinions portant sur les décisions que l'homme politique a pu prendre par le passé au sujet de l'environnement et de l'écologie.
		Exemple : « il faut bien le reconnaître, sarko a plus fait pr l'environnement qu'Hollande » (nb : ici l'entité visée est Sarkozy).
C2. Bilan	Economie	Opinions portant sur les décisions que l'homme politique a pu prendre par le passé au sujet de l'emploi, de l'industrie, de la finance, de la compétitivité, du déficit, de la dette, de la fiscalité, de la sécurité sociale et de toute question relative à l'économie.
		Exemple : « Sarko va recevoir le "Gérard de l'armateur" pour réussir à donner des conférences en économie quand on a rien à lui donner de sérieux ? »
	Sociétal	Opinions portant sur les décisions que l'homme politique a pu prendre par le passé au sujet de questions sociétales (immigration, droit des homosexuels, lutte des sans-papiers, questions de laïcité/religieuses, etc.)
		Exemple : « Les débats sous #Hollande ont une autre gueule. #MariagePourTous #FinDeVie Sous Sarko, c'était #IdentitéNationale #Immigration #Roms » (nb : ici l'entité visée est Sarkozy).
C3. Compétence	Expertise	L'homme politique est-il considéré comme très compétent ou au contraire très incompetent dans un domaine donné (comme celui de l'économie par exemple)
		Exemple : « @AndreIrwin Oui, #Cop21 est à l'honnêteté et l'humilité, ce qu'Hollande est à la compétence économique. Un désastre. » (nb : ici l'entité visée est Hollande).
	Gouverner	Contrainte à la sous-cible « expertise », la sous-cible « gouverner » est plus large et ne vise pas un domaine de compétence précis. Elle concerne la

Let's use machine learning



13

Classification results – case 1

ID lot	Méthode employée	Macro F-Score	Micro F-Score
7	Combinaison linéaire pondérée	0.62	0.80
8	Combinaison linéaire pondérée	0.71	0.81
9	Combinaison linéaire pondérée	0.59	0.71
8	Cosinus-LIA	0.64	0.71
9	Cosinus-LIA	0.56	0.65
7	Xerox	0.37	0.53
8	Xerox	0.48	0.50
9	Xerox	0.43	0.46

ID lot	Méthode employée	Macro F-Score	Micro F-Score
7	Combinaison linéaire pondérée	0.36	0.50
8	Combinaison linéaire pondérée	0.61	0.69
9	Combinaison linéaire pondérée	0.32	0.48
7	Xerox	0.24	0.22
8	Xerox	0.26	0.37
9	Xerox	0.22	0.34
8	Cosinus-LIA	0.40	0.54
9	Cosinus-LIA	0.25	0.40

14

Classification results – case 2

Batch 2 divided into batch 3 (posterior validation) and batch 4 (blind validation)

polarity	ID lot	Méthode employée	Macro F-Score	Micro F-Score
2	2	Méthode Xerox	0.60	0.75
2	2	Cosinus-LIA	0.68	0.80
2	2	Combinaison linéaire pondérée	0.73	0.83
3	3	Combinaison linéaire pondérée	0.79	0.85
4	4	Combinaison linéaire pondérée	0.62	0.78

target	ID lot	Méthode employée	Macro F-Score	Micro F-Score
2	2	Méthode Xerox	0.59	0.60
2	2	Cosinus-LIA	0.65	0.68
2	2	Combinaison linéaire pondérée	0.70	0.71
3	3	Combinaison linéaire pondérée	0.64	0.74
4	4	Combinaison linéaire pondérée	0.59	0.65

15

La France est une république indivisible, **démocratique**, laïque et sociale, voilà mon **engagement**. #H2012 → (Ethique, ++)

→ (Positionnement, +)

Geste fort du président **Hollande** qui participera ce jeudi à la journée des mémoires, de la traite, de l'esclavage et de leurs abolitions. → (Injonction, +)

Pourquoi **j'aime bien** Mélenchon et **je voterai** Hollande
http://t.co/TVM8RwoH via @***** → (Injonction, +)

→ (Personne:Charisme, -)

#Delanoë "ce qui me frappe ds la campagne de **Hollande** c son honnêteté intellectuelle alors que **Sarkozy** dit tout et n importe quoi" → (Ethique:Honnêteté, +)

→ (Personne:Charisme, -)

@aut-1154 Neuilly sur Seine 61100 habitants , France 65000 000.Votez Hollande. → (Injonction, +)

→ (Personne:Charisme, -)

@***** Hollande n'a aucun charisme ! Il fait honte à la France et aux Français ! → (Personne:Charisme, -)

→ (Personne:Charisme, -)

Sympatiques, ce Hollande. Et cultivé avec ça. On a parlé saucisses toute la soirée. → (Personne:Charisme, -)

→ (Personne:Charisme, -)

Je savais qu'Hollande était un gros mou de socialiste. Mais là si ce n'est pas du reniement ou du renoncement ?Libertédeconscience

→ (Ethique:Honnêteté, -)

François Hollande : le mensonge c'est maintenant: C'est cela un président . Il y a pas comme un léger bug → (Ethique:Honnêteté, -)

→ (Ethique:Honnêteté, -)

Copé appelle Hollande à "reprendre en main" son gouvernement "incompétent" http://t.co/IPanwi5r via @LePoint

→ (Compétence, -)

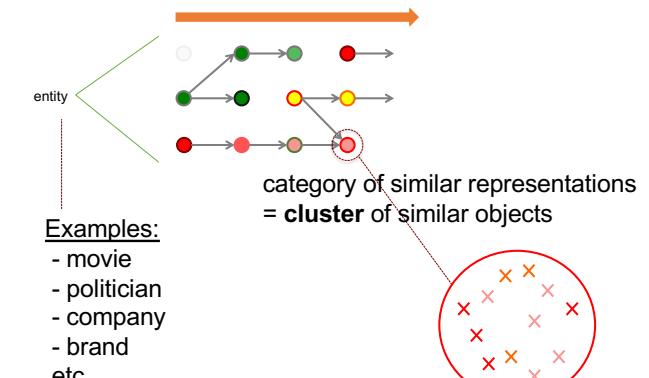
16

Sparse matrix as input

		description features								
Author	Time	f ₁	f ₂	f ₃	f ₄	f ₅	f ₆	...	f _{n-1}	f _n
pseudo1	t1		1				2		1	
pseudo1	t2		1				1			
pseudo1	t3				2					2
pseudo2	t1		3	1						1
pseudo3	t1				3					
pseudo3	t2				2					
pseudo3	t3				2					
pseudo4	t3	3					1			
pseudo5	t3					3				2

17

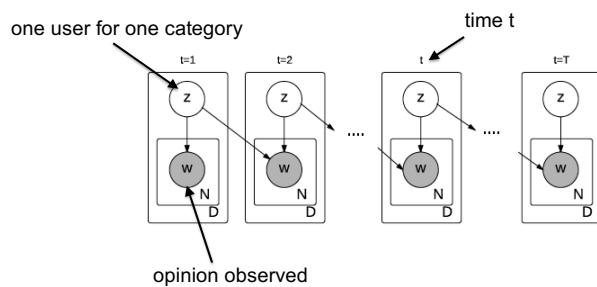
Temporal evolution of entities



18

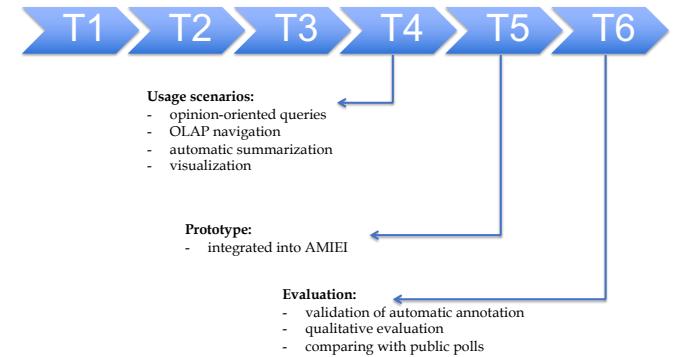
Temporal Mixture Model

- TMM = probabilistic generative model (Kim et al., 2015)



19

Structuration of the project



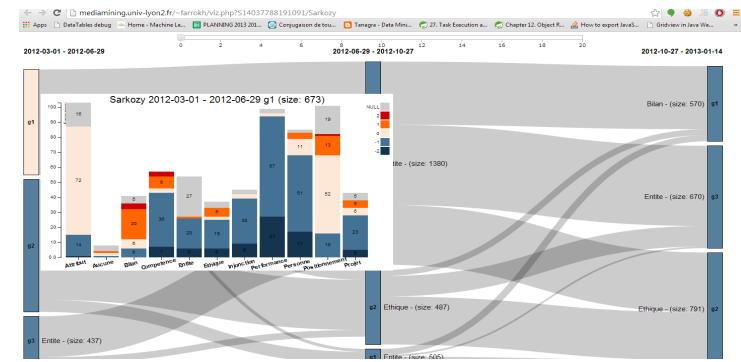
20

Example of visualization



21

Example of visualization (con't)



22

Topic models as « explanations » of given categories

Julien Velcin – ERIC Lab, Université Lyon 2 – Séminaire UDL 2022

23

Discharge summaries (joint work with M. Dermouche, S. Loudcher, R. Flicoteau, S. Chevret)

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

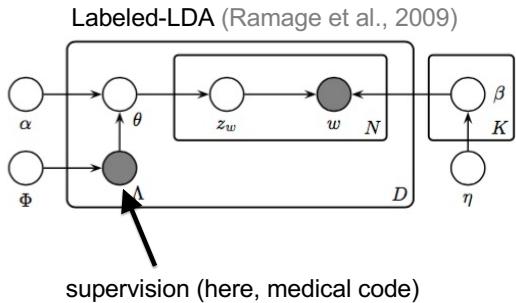
N40

masculin Antécédents médicaux Bloc de branche Arthroscie 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 Des suites opératoires ont été simples Arrêt des lavages à J2 et ablation de la sonde vésicale à J3 Episode de rétention aigüe d urines nécessitant un sondage en urgence à J3 Ablation de la sonde vésicale à J4 et reprise spontanée des mictions Conclusion Le patient sera revu en consultation dans un mois Antalgiques IXPRIM 1cp 4 fois jour si douleurs importantes par Dr Louis FROGER

Hyperplasie de la prostate

24

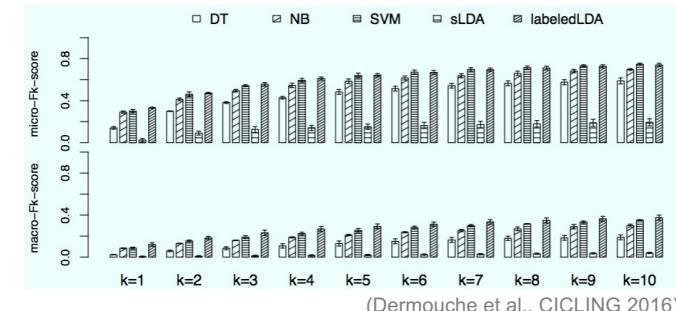
Supervised topic modeling



25

Some comparative results

joint work with APHP, St-Louis Hospital



26

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 aigüe (Acute cystitis)	S30.2: Contusion des organes génitaux externes (Congestion of the external genitalia)
prostatectomie ⁴	incontinent ⁵	prélèvement (sample)	pontage (bypass)	observer (watch)
radical	bandelette (band)	faueur (favour)	arterielle (arterial)	hospitalisé (inpatient)
laparotomie (laparotomy)	effort (stress)	manuel (hand-operated)	Ditropan	med
score	trans-obturatrice ⁵	artère (artery)	post-mictionnelle ⁵	ext
lobe (lobus)	urodynamique ⁵	assisté (assisted)	Kardégic	motif (cause)
mini	touz (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	urgenture ⁵	apparenté (related)	impériosité (urge)	fan (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
CS1.9: Lymphome de Hodgkin (Hodgkin's lymphoma)	CS8.0: Macroglobulinémie de Waldenström (Waldenström's macroglobulinemia)	D46.2: Anémie réfractaire avec excès à petites cellules B de blastes (refractory anemia with excess of blasts)	C83.0: Lymphome B généralisée secondaire (small B-cell lymphoma with excess of blasts)	E85.3: Amylose
ABVD	IgM	senior	critère (criterion)	amylose
IVOX	lymphoplasmocytaire	multirésistant(resistant)	participer(participate)	troponine (troponin)
classique (classical)	macroglobulinémie ⁵	remise (redelivery)	accepter (accept)	formule (formula)
panoramique(panoramic)	monoclonal	blast (blast)	consentement(consent)	BNP
escalade (escalation)	béta (beta)	AREB (RAEB)	aborder (approach)	VCD
étoposide (etoposide)	créatininémie ⁵	leuco	attendu (expected)	évolution (evolution)
BEAM	sup (increased)	Vidaza	logistique (logistics)	dosage (dose)
SPI (IPS)	stabilité (stability)	myelodysplasique ⁵	version	pro
nodulaire (nodular)	cérébral (cerebral)	BHC	objectif (goal)	arriver (reach)
#documents=168	F ₁ -score=0.75	mgX (m.g.)	contrainte(constraint)	immunochimique ⁵
#documents=72	F ₁ -score=0.74			
#documents=37	F ₁ -score=0.78			
#documents=38	F ₁ -score=0.38			

27

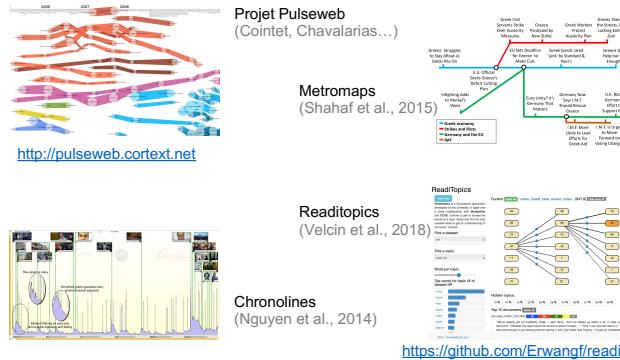
Studying the informational landscape

Julien Velcin – ERIC Lab, Université Lyon 2 – Séminaire UDL 2022

28

Studying the « mediascape »

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



29

What we did

- Joint work with a researcher in information science (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)

30

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)

31

Some topics extracted by LDA

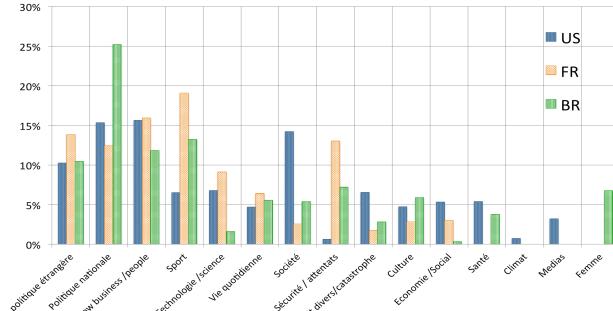
topic	#doc	cat.	mots les plus probables
<i>en français (sur 4133 articles) :</i>			
z18	28	1	manifestation, paris, police, travail, loi, contre, syndicats, place, bastille, 2016
z19	36	1	loi, travail, gouvernement, l'état, texte, l'assemblée, d'urgence, mois, projet, conseil
z25	39	2	jeux, rio, olympiques, olympique, août, jo, athlètes, 2016, brésil, cérémonie
z47	18	3	morandini, jean-marc, inrocks, catherine, l'amateur, lui, qu'il, europe, comédiens, plainte
z73	47	4	nice, 14, l'attentat, anglais, promenade, camion, attentat, police, soir, christian
<i>en anglais (sur 12067 articles) :</i>			
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
<i>en portugais (sur 2356 articles) :</i>			
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.

32

Compared results

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



33

Newsbrowsers

The screenshot shows a software interface for managing news articles. It includes a search bar, filter options for tags and periods, and a list of articles with columns for ID, Title, and Author. A sidebar lists various topic categories.

ID	Titre	Auteur
...	L'assassinat adorabile de Catherine Laborde sur ...	(no author)
...	Face au projet de Tour Triangle, les écologistes...	(no author)
...	Devenu un symbole des anti-Donald Trump, Ollier...	(no author)
...	Les dernières rumeurs sur la mort de François Holl...	(no author)
...	François Hollande s'est donné comme conseil de ...	(no author)
...	Petit tour du monde des sources d'énergie renou...	(no author)
...	Comment débarrasser François Fillon les quatres...	(no author)
...	Voici ce qu'il faut savoir pour bien recycler...	(no author)
...	Il n'est pas trop tard pour faire barrage au CETA	(no author)
...	Le législateur de l'Assemblée à l'origine de la crise	(no author)
...	Le gouvernement audacieux signe ici une loi qui ...	(no author)
...	Chronique d'une victoire annoncée - Episode 31...	(no author)

34

Conclusion and future work

Julien Velcin – ERIC Lab, Université Lyon 2 – Séminaire Udl 2022

35

Some lessons I've learnt

- Draw the line between *a priori* knowledge and what can be learnt from the data
- Don't think partners in LLSSH have clear (and unique) categorization in mind
- Never underestimate the time (and cost) needed for building a categorization framework (think hard about the annotations)
- Collaboration with LLSSH needs:
 - mutual understanding
 - trust
 - time

36

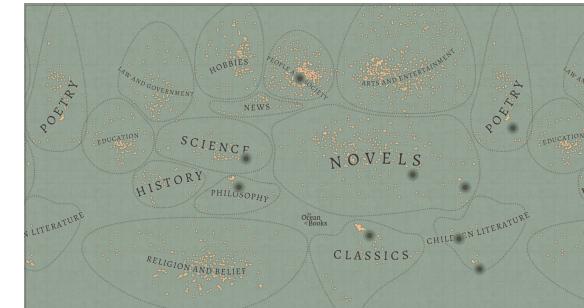
Ongoing work

- Collaboration with Ian Davidson (UCD Davis): clustering + XAI
- Project **POIVRE** (ERIC, EDF): Viewpoint detection on energy issues through Twitter
- Project **TIGA** (Lyon metropolis, IMU labex, ERIC...): L'industrie [Re]connectée et intégrée à son territoire et à ses habitants
- Project **LIFRANUM** (MARGE, ERIC, BnF): Identify and structure the corpus of digital French literatures
- Project **DIKé** (LHC, ERIC, NAVER Labs): Bias, fairness and ethics of compressed NLP models

37

LIFRANUM project

- Partners: MARGE, ERIC, BnF



38

Some references

- (Boadjian et Velcin, 2017) De l' « opinion mining » à la sociologie des opinions en ligne. Pour une approche interdisciplinaire de l'étude du web politique. Question de communication, 2017.
- (Christophe et al., 2021) Change detection in textual classification with unexpected dynamics. ESWA 2021
- (Dermouche M. et al., 2016) Supervised Topic Models for Diagnosis Code Assignment to Discharge Summaries, CICLING 2016.
- (Kim et al., 2015) Temporal multinomial mixture for instance-oriented evolutionary clustering, ECIR 2015.
- (Velcin et al., 2014) Investigating the Image of Entities in Social Media: Dataset Design and First Results, LREC 2014.
- (Velcin J. et al., 2017) 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 J. et al., 2018) Readitopics: Make Your Topic Models Readable via Labeling and Browsing , IJCAI 2018.

Thank you!

39