At the frontier of unsupervised and supervised learning

…some lessons I've learnt along the way with researchers in SSH

Julien Velcin
http://eric.univ-lyon2.fr/~jvelcin
ERIC Lab, Université Lyon 2

MASHS-2022
Models and Learning in the Humanities and Social Sciences (MASHS)

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)
- Lessons I've learnt
- Ongoing work

Context and motivation

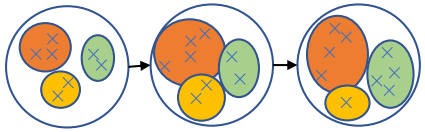
Julien Velcin – ERIC Lab, Université Lyon 2 – Séminaire MASHS 2022

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…)

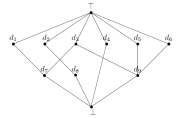
Dilemma of categorization

- Origin of (natural) categorization not very well known (kind of chicken and egg problem)
- Previous works in pyschology, 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:

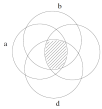


Some categorization theories

- Logical theory (Aristotle)

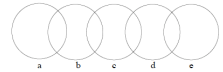


- Prototypes (Rosch,73,75,78) (Lakoff,87) (Kleiber, 88)



- Stereotypes (Lippman,22)

- Family resemblance (Wittgenstein,53)



Once upon a time: the ImagiWeb project

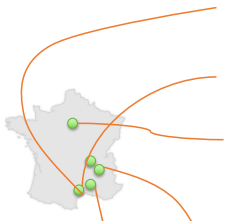
Julien Velcin – ERIC Lab, Université Lyon 2 – Séminaire MASHS 2022

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)

- ANR grant (2012-2015)

Collaborative project



ami — éditeur logiciel
spécialistes de la veille sur Internet

CEPEL — Utilisateurs (cas d'étude n°1)
spécialistes en sciences politiques

EDF — Utilisateurs (cas d'étude n°2)
informaticiens, veilleurs, sémiologues

eric — Fouille de données
Apprentissage automatique

xerox Research Centre Europe — Traitement automatique des langues
Extraction d'information

LABORATOIRE INFORMATIQUE D'AVIGNON — Recherche d'information
Analyse des opinions

Structuration of the project

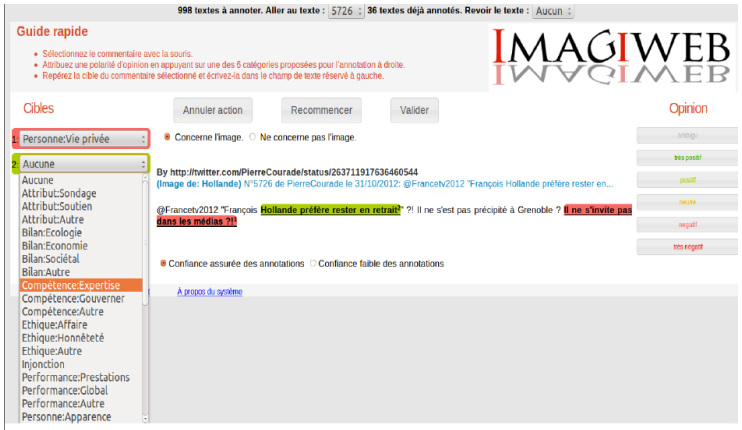
T1 T2 T3 T4 T5 T6

Management

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

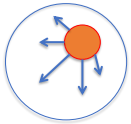
Manual annotation of a subset:
- Case study n°1: 11,527 annotations of 7,283 tweets
  (20 annotators, between 1 and 3 per tweet)
- Case study n°2: 600 blog posts (passages)
  (2 annotators)



JSON : {pertinence:concerne,confiance:assuree,ambigu:{},trespositif:{},positif:{2:"Hollande préfère rester en retrait"},neutre:{},negatif:{1:"Il ne s'invite pas dans les médias"},tresnegatif:{}}

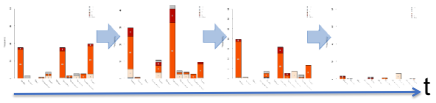
Let's use machine learning

T1 T2 T3 T4 T5 T6



Automatic annotation:
- linguistic approach
- statistical approach
- hybrid approach

Building coherent images:
- temporal-aware clustering



Classification results – case 1

7=FH (validation)
8=NS (validation)
9=NS (blind)

polarity

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

target

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

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	Méthode Xerox	0.60	0.75
2	Cosinus-LIA	0.68	0.80
2	Combinaison linéaire pondérée	0.73	0.83
3	Combinaison linéaire pondérée	0.79	0.85
4	Combinaison linéaire pondérée	0.62	0.78

target

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

La France est une république indivisible, démocratique, laïque et sociale, voilà mon engagement. #FH2012 → (Ethique, + +)

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

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

#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, +)

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

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

Sympatisch, ce Hollande. Et cultivé avec ça. On a parlé saucisses toute la soirée. → (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, - -)

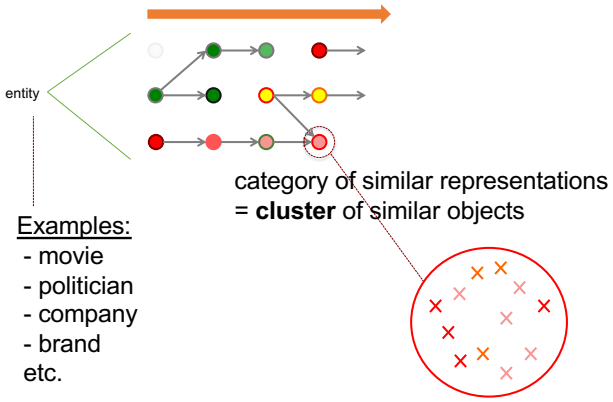
Copé appelle Hollande à "reprendre en main" son gouvernement "incompétent" http://t.co/lPanwi5r via @LePoint → (Compétence, -)

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

Temporal evolution of entities

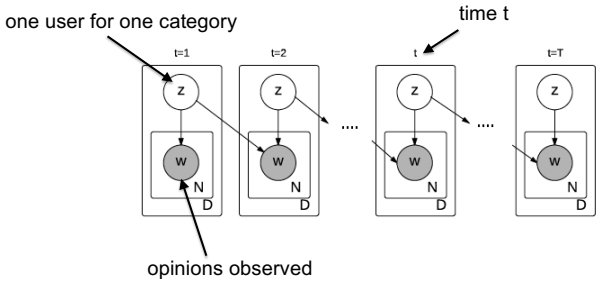


Examples:
- movie
- politician
- company
- brand

etc.

Temporal Mixture Model

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



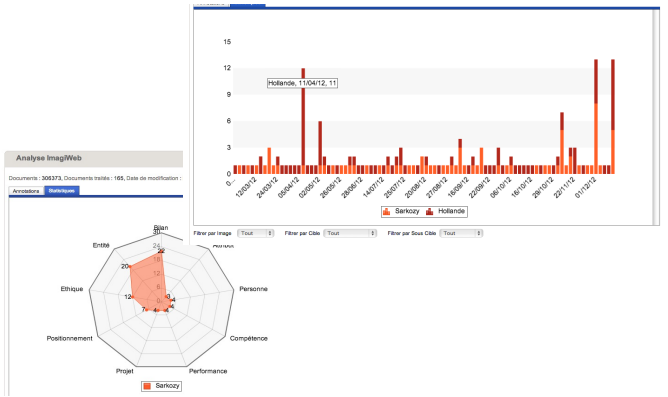
Structuration of the project

Usage scenarios:
- opinion-oriented queries
- OLAP navigation
- automatic summarization
- visualization

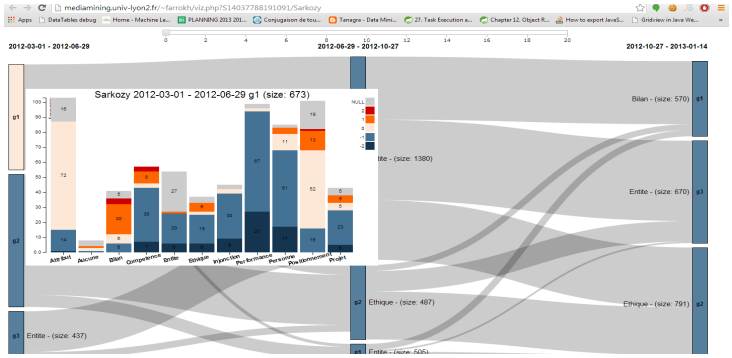
Prototype:
- integrated into AMIEI

Evaluation:
- validation of automatic annotation
- qualitative evaluation
- comparing with public polls

Example of visualization



Example of visualization (con't)



Topic models as « explanations » of given categories

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

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

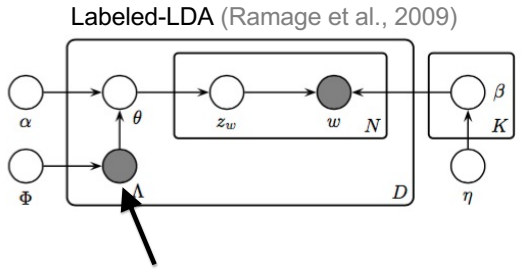
N40

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 lavages à J2 et ablation de la sonde vésicale à J3 Episode de rétention aigue d urines nécessitant un sondage en urgence à J3 Ablation de la sonde vésicale à J4 et reprise spontanée des mictions Conslusion Le patient sera revu en consultation dans un mois Antalgiques IXPRIM 1cp 4 fois jour si douleurs importantes par Dr Louis FROGER

Supervised topic modeling

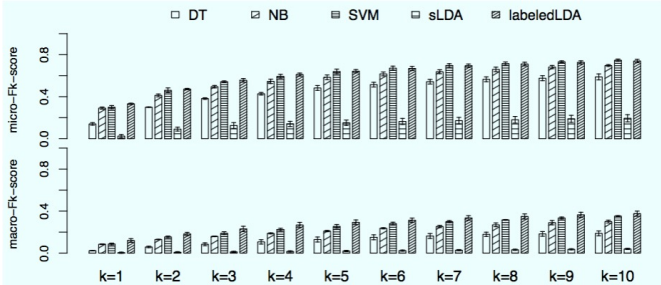
Labeled-LDA (Ramage et al., 2009)



supervision (here, medical code)

Some comparative results

- joint work with APHP, St-Louis Hospital



(Dermouche et al., CICLING 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)	*artérielle* (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

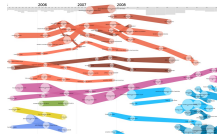
C81.9: *Lymphome de Hodgkin* (Hodgkin's lymphoma)	**C88.0**: *Macroglobulinmie de Waldenström* (Waldenström's macroglobulinemia)	**D46.2**: *Anémie réfractaire avec excès de blastes* (refractory anemia with excess of blasts)	**C83.0**: *Lymphome à petites cellules B* (small B-cell lymphoma)	**E85.3**: *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

Studying the informational landscape

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

Studying the « mediascape »

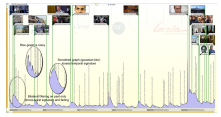
Appadurai A.: Disjuncture and Difference in the Global Cultural Economy, Theory Culture Society 1990; 7; 295



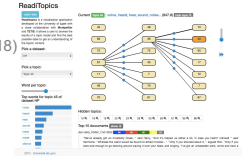
Projet Pulseweb
(Cointet, Chavalarias…)

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 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)

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)

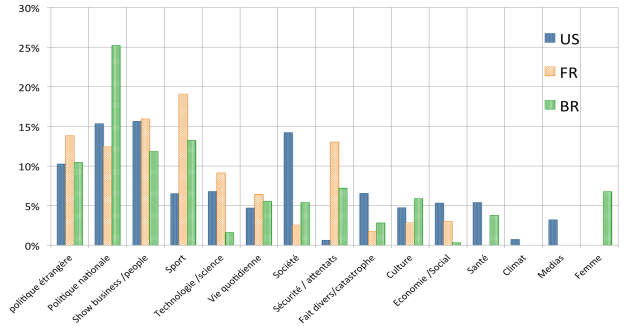
Some topics extracted by LDA

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

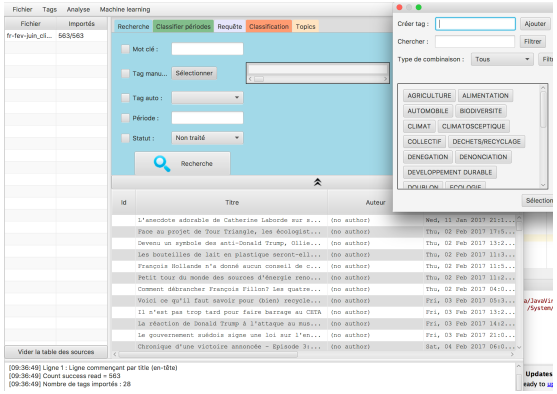
Les catégories attribuées ici (cat.) correspondent à : 1- Economie / Social, 2- Sport / JO, 3- Show business / people, 4- Sécurité / attentats, 5- Politique étrangère, 6- Technologie / science, 7- Société, 8- Politique nationale, 9- Santé.

Compared results

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



Newsbrowsers



Conclusion and future work

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

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
- Collaboration with LLSSH needs:
  - mutual understanding
  - trust
  - time

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

Some references

(Boyadjian 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.

(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!