



Exemple 1

Classification (partitionnement) de textes

Exemple : classification de textes

```
Nom de fichier;Titre;Auteur(s);Affiliation(s);Revue ou monographie;ISSN;e-ISSN;ISBN;e-ISBN;Éditeur;Type de publication;Date de publication;Catégories WoS;Catégories Science-Metrix;Catégories Scopus;Catégories INIST;Score qualité;Version PDF;XML structuré;Identifiant ISTEX;ARK;DOI
s_000002;Structures and diseases;"K Ulrich Wendt <sup>1</sup> ; Manfred S Weiss <sup>2</sup> ; Patrick Cramer <sup>3</sup> ; Dirk W Heinz <sup>4</sup> ; Sanofi-Aventis, Frankfurt, D-65926, Germany ; European Molecular Biology Laboratory, c/o DESY, Hamburg, D-22603, Germany ; Gene Centre, Ludwig Institute for Structural Biology, Helmholtz Centre for Infection Research, Braunschweig, D-38124, Germany";Nature Structural & Molecular Biology;1545-9999;structural biology is making significant contributions toward an understanding of molecular constituents and mechanisms underlying human diseases at the 10th International Conference on Structural Biology of Disease Mechanisms held in September 2007 in Murnau, Germany.";"1 - science ; 2 - cell biology ; 2 - life sciences ; 2 - biomedical research ; 3 - developmental biology";"1 - Life Sciences ; 2 - Biochemistry, Genetics and Molecular Biology ; 3 - Genetics and Molecular Biology ; 3 - Structural Biology";1 - sciences humaines et sociales;5.26;1.4;Absent;C20103CC68E46DEBF1A30871D342FC7F50B
7
```

Corpus SARS-MERS-Export																							
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S					
1	Nom de fichier	Titre	Auteur(s)	Affiliation(s)	Revue ou monographie	ISSN	e-ISSN	ISBN	e-ISBN	Éditeur	Type de publication	Date de publication	Langue(s) du Résumé	Mots-clés / Catégories WoS	Catégories Science-Metrix	Catégories Scopus	Catégories INIST	Score qualité	Version PDF	XML structuré	Identifiant ISTEX	ARK	DOI
2	sars-mens_00001	Structures and diseases	L. Ulrich Wendt ¹; Manfred S Weiss ²	Department of Chemical and Analytical Sciences at Sandford-Avantis, Frankfurt, D-65926, Germany	1545-9999	1545-9985				Wiley	journal	review-article	2008	Anglais	Structural biology is making significant contributions toward 1 - science ; 2 - cell biology ; 2 - life sciences ; 2 - biomedical research ; 3 - developmental biology	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
3	sars-mens_00002	Valuating Euro-Mediterranean Stephen C. Calleja		Research and Development, American Red Cross, London Labour British Journal	0007-1048	1365-2141	9,7807612 9,7807612+12	9,7807612+12	9,7807612+12	Elsevier	journal	conference	2008	Anglais	What are the prospects for the future of the Euro-Mediterranean region? A valuation exercise	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
4	sars-mens_00003	Emerging pathogens and their Roger Y. Fouquet		Y. Jaf CNRS, UMI 3189, 13015, Marseille, France ; Faculté de médecine Bulletin de la Société de pathol.	0007-0985	1961-9049				Wiley	journal	review-article	2012	Anglais	The threat of infection by conventional transplant blood transfusion	1 - science ; 1 - health sci.1	1 - health sci.1	1 - health sci.1	1 - health sci.1				
5	sars-mens_00004	Planetary science: Mission to David J. Stevenson		California Institute of Technology, Pasadena, California 91125, United States	0028-0836					Nature	journal	research-article	2011	Anglais	Not science fiction, but a technically feasible plan to probe 1 - science ; 1 - general ; 1 - Gen	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
6	sars-mens_00005	Advances in the study of the H5N1 virus in J. Remy		Fédération des maladies infectieuses et tropicales de l'Institut National Africain d'Hôpital-Gaston Bergeron, Paris, France	0033-1554	1954-2112				Lavoisier	journal	research-article	2009	Anglais	Advances in the study of the H5N1 virus in J. Remy	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
7	sars-mens_00006	RNA aptamer-based sensing: Daniel J. Abelson ¹; Jili Jin ²; Li-Jia Ju ³; ...		Jung D Department of Biotechnology, Young University, Seoul 120-440, South Korea	0033-1554	1364-5528				RSC	journal	research-article	2009	Anglais	Severe acute respiratory syndrome coronavirus (SARS-CoV) 1 - science ; 1 - natural sci.1	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
8	sars-mens_00007	Short burst oxygen therapy for C. Roberts		Correspondence to C. Roberts Department of Respiratory and Critical Care Medicine, King's College London, London, United Kingdom	0066-6376	1468-2396				BMJ	journal	editorial	2004	Anglais	Severe acute respiratory syndrome coronavirus (SARS-CoV) 1 - science ; 1 - natural sci.1	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
9	sars-mens_00008	Regulation of the art of control D. S. Robinson		Correspondence to D. S. Robinson Leukemia Biology Section, Thomas Jefferson University, Philadelphia, PA, United States	0066-6376	1468-2396				BMJ	journal	editorial	2004	Anglais	Regulation of the art of control D. S. Robinson	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
10	sars-mens_00009	Anterior segment and cornea W. Yeung ¹; ...		D 5 C Holt Department of Ophthalmology, University of Hong Kong, Hong Kong, China	0066-6376	1468-2396				BMJ	journal	editorial	2004	Anglais	Anterior segment and cornea W. Yeung ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
11	sars-mens_00010	Advances in the study of the H5N1 virus in J. Remy		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	editorial	2009	Anglais	Advances in the study of the H5N1 virus in J. Remy	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
12	sars-mens_00011	Initial investigation of the Jiangxi Tan ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	Initial investigation of the Jiangxi Tan ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
13	sars-mens_00012	An initial investigation of the Jiangxi Tan ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	An initial investigation of the Jiangxi Tan ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
14	sars-mens_00013	Advantage to being different: Bill et al.		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	Advantage to being different: Bill et al.	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
15	sars-mens_00014	New approaches to glycan arrays C. R. Wilson ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	New approaches to glycan arrays C. R. Wilson ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
16	sars-mens_00015	Strategies, expertise et territoire Jacques Lourdeau ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	Strategies, expertise et territoire Jacques Lourdeau ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
17	sars-mens_00016	Strategies, expertises et territoires J. Lourdeau ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	Strategies, expertises et territoires J. Lourdeau ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
18	sars-mens_00017	Strategies, expertises et territoires J. Lourdeau ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	Strategies, expertises et territoires J. Lourdeau ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
19	sars-mens_00018	New approaches to glycan arrays C. R. Wilson ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	New approaches to glycan arrays C. R. Wilson ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
20	sars-mens_00019	Matière préliminaire Alain Pellerin		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	Matière préliminaire Alain Pellerin	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
21	sars-mens_00020	Recueil des cours		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				BMJ	journal	other	2005	Anglais	Recueil des cours	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
22	sars-mens_00021	Association of CAMG Genetic Kohen Y. Chen ¹; ...		Center for Biostatistics, University of Michigan Health System, Ann Arbor, MI, United States	0022-1899	1537-6013				OUP	journal	research-article	2007	Anglais	Association of CAMG Genetic Kohen Y. Chen ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
23	sars-mens_00022	Open Reading Frame 8a of the Chikungunya virus		Institute of Pulmonary Disease, Taiwan, Republic of China	0022-1899	1537-6013				OUP	journal	research-article	2007	Anglais	Open Reading Frame 8a of the Chikungunya virus	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
24	sars-mens_00023	Advances in the study of the H5N1 virus in J. Remy		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2007	Anglais	Advances in the study of the H5N1 virus in J. Remy	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
25	sars-mens_00024	Viral lower respiratory tract disease severity in children H. M. van der Veen ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2007	Anglais	Viral lower respiratory tract disease severity in children H. M. van der Veen ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
26	sars-mens_00025	Chapter I - Preliminary Issues A. M. Strickland		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2007	Anglais	Chapter I - Preliminary Issues A. M. Strickland	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
27	sars-mens_00026	Survey of the year 2004 comm. Rebecca L. Rich ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Survey of the year 2004 comm. Rebecca L. Rich ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
28	sars-mens_00027	L'importance de la couleur et de l'épaisseur R. R. Röder		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	L'importance de la couleur et de l'épaisseur R. R. Röder	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
29	sars-mens_00028	Advantages and challenges of Confocal T. Kägi ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Advantages and challenges of Confocal T. Kägi ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
30	sars-mens_00029	CHAPTER 9 - Viral-coded ion C. Stephen Griffin		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	CHAPTER 9 - Viral-coded ion C. Stephen Griffin	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
31	sars-mens_00030	Investigation d'une épidémie I.C. Aumeran ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Investigation d'une épidémie I.C. Aumeran ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
32	sars-mens_00031	Alternative sequence analysis J. E. Phillips ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Alternative sequence analysis J. E. Phillips ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
33	sars-mens_00032	Advantages and portable & portable G. Gould ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Advantages and portable & portable G. Gould ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
34	sars-mens_00033	Content		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Content	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
35	sars-mens_00034	Chapter 3 - Pharmacophore by Christian Lagger ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Chapter 3 - Pharmacophore by Christian Lagger ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
36	sars-mens_00035	Investigation avancée de la fréquence pour les séries de séquences (FMT) ; V. Boulard ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Investigation avancée de la fréquence pour les séries de séquences (FMT) ; V. Boulard ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
37	sars-mens_00036	Contenu		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Contenu	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
38	sars-mens_00037	Virology: SARS virus infection Ryon E. E. Martina ¹; ...		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Virology: SARS virus infection Ryon E. E. Martina ¹; ...	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				
39	sars-mens_00038	Subject index		Carrier-resolved phylogeny, Fudan University, Shanghai, China	0033-1554	1954-2112				OUP	journal	research-article	2005	Anglais	Subject index	1 - health sci.1	1 - life sciences	1 - health sci.1	1 - health sci.1				

ISTEX

fichier .CSV sur 27 colonnes
(méta-données ISTEX)

```

import pandas as pd

fichierCSV = "//Users/Patrice/PycharmProjects/ANF2021/ANF/test2.csv"
# load train data
data = pd.read_csv(fichierCSV, sep=";", header=0, error_bad_lines=False, encoding="utf_8")
data.head(10)

```

	Nom de fichier	Titre	Auteur(s)	Affiliation(s)	Revue ou monographie	ISSN	e-ISSN	ISBN	e-ISBN	Éditeur	...	Catégories Science-Metrix	Catégories Scopus	Catégories INIST	S
0	sars-mers_00002	Structures and diseases	K Ulrich Wendt ¹; Manfred S Weiss ...	Department of Chemical and Analytical Sciences...	Nature Structural & Molecular Biology	1545-9993	1545-9985	NaN	NaN	Nature	...	1 - health sciences ; 2 - biomedical research ...	1 - Life Sciences ; 2 - Biochemistry, Genetics...	1 - sciences humaines et sociales	5.
1	sars-mers_00003	Evaluating Euro-Mediterranean Relations	Stephen C. Calleya	NaN	Evaluating Euro-Mediterranean Relations	NaN	NaN	9.780715e+12	9.780203e+12	taylor-francis	...	NaN	NaN	NaN	8.
2	sars-mers_00005	Emerging pathogens and their implications for ...	Roger Y. Dodd	Research and Development, American Red Cross, ...	British Journal of Haematology	0007-1048	1365-2141	NaN	NaN	Wiley	...	1 - health sciences ; 2 - clinical medicine ; ...	1 - Health Sciences ; 2 - Medicine ; 3 - Hemat...	1 - sciences appliquées, technologies et medec...	7.2
3	sars-mers_00006	Pandémie grippale A/H5N1 et niveau de préparat...	E. d'Alessandro ¹; G. Soula <sup>2...>	CNRS, UMI 3189, 13015, Marseille, France ; fac...	Bulletin de la Société de pathologie exotique	0037-9085	1961-9049	NaN	NaN	Lavoisier	...	NaN	NaN	NaN	8.
4	sars-mers_00007	Planetary science: Mission to Earth's core — a...	David J. Stevenson	California Institute of Technology, Pasadena, ...	Nature	0028-0836	NaN	NaN	NaN	Nature	...	1 - general ; 2 - general science & technology...	1 - General ; 2 - Multidisciplinary ; 3 - Mult...	1 - sciences humaines et sociales	4.
5	sars-mers_00008	Première étude sur le dépistage et la prise en...	A. -J. Rémy	Fédération des unités médicales des centres de...	Journal Africain d'Hépato-Gastroentérologie	1954-3204	1954-3212	NaN	NaN	Lavoisier	...	NaN	NaN	NaN	2.
6	sars-mers_00395	RNA aptamer-based sensitive detection of ...	Dae-Gyun Ahn ¹; ...	Department of Biotechnology, Yonsei	The Analyst	0003-2654	1364-5528	NaN	NaN	RSC [journals]	...	1 - natural sciences ; 2 - chemistry ;	1 - Physical Sciences ; 2 - Chemistry ; 3 -	1 - sciences appliquées, technologies	7.9

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9 entries, 0 to 8
Data columns (total 27 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Nom de fichier    9 non-null      object  
 1   Titre              9 non-null      object  
 2   Auteur(s)          9 non-null      object  
 3   Affiliation(s)    8 non-null      object  
 4   Revue ou monographie  9 non-null      object  
 5   ISSN               8 non-null      object  
 6   e-ISSN             7 non-null      object  
 7   ISBN               1 non-null      float64 
 8   e-ISBN              1 non-null      float64 
 9   Éditeur            9 non-null      object  
 10  Type de publication 9 non-null      object  
 11  Type de document   9 non-null      object  
 12  Date de publication 9 non-null      int64   
 13  Langue(s) du document 9 non-null      object  
 14  Résumé             6 non-null      object  
 15  Mots-clés d'auteur 4 non-null      object  
 16  Catégories WoS     6 non-null      object  
 17  Catégories Science-Metrix 6 non-null      object  
 18  Catégories Scopus   6 non-null      object  
 19  Catégories INIST    4 non-null      object  
 20  Score qualité      9 non-null      float64 
 21  Version PDF         9 non-null      float64 
 22  XML structuré       9 non-null      object  
 23  Identifiant ISTEX  9 non-null      object  
 24  ARK                9 non-null      object  
 25  DOI                9 non-null      object  
 26  PMID               6 non-null      float64 
dtypes: float64(5), int64(1), object(21)
```

Identifier les colonnes (in)utiles

Enregistrement automatique		Corpus SARS-MERS-Export.csv - Lecture seule																															
Accueil		Insertion		Dessin		Mise en page		Formules		Données		Révision		Affichage		Acrobat		Dites-le-nous		Partager						Commentaires							
Calibri (Corps)		12		A A		Renvoyer à la ligne automatiquement		Standard		Fusionner et centrer		Mise en forme conditionnelle		Mette sous forme de tableau		Styles de cellule		Insérer Supprimer Mise en forme		Somme automatique		Remplissage		Trier et filtre		Rechercher et sélectionner		Idées		Créer et partager un PDF Adobe			
AA1	X	V	PMID	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA			
1	Nom de fichier	Titre	Auteur(s)	Affiliation(s)	Revue ou mi	ISSN	e-ISSN	ISBN	e-ISBN	Éditeur	Type de publ	Type de docu	Date de publ	Langue(s)	Résumé	Mots-clés d'	Catégories V	Catégories S	Catégories S	Catégories S	Catégories II	Score qualit	Version PDF	XML structur	Identifiant IS	ARK	DOI	PMID					
2	sars-mers_00002	Structures and diseases	K Ullrich We	Department	Nature Struct	1545-9993	1545-9985			Nature	journal	conference	2008	Anglais	Structural biology is making science ; 1 - health sci 1 - Life Sci 1 - sciences	15.26	1.4	Absent	C2010C68	ark:/67375/c2010c68	10.1038/nsr	18250627											
3	sars-mers_00003	Evaluating Early-Mediterranean Relatio	Stephen C. Cilleya	Relat	Evaluating E	1545-9993	1545-9985	9,7807E+12	9,7802E+12	Journal-franc	book	book	2005	Anglais	What are the prospects for the future of the Early-Mediterranean area and who is	8.92	1.6	Absent	B38466E2B2	ark:/67375/f10432a978202017647	10.4324/978202017647												
4	sars-mers_00005	Emerging pathogens and their implia	Roger Y. Doc	Researche	British Jour	0007-1044	1365-2141			Wiley	book	review-articl	2012	Anglais	The threat of blood transfu 1 - science ; 1 - health sci 1 - Health Sci 1 - sciences	7.792	1.3	Absent	A0410ED45	ark:/67375/10.1111/bjh.	22924410												
5	sars-mers_00006	Pandémie grippale A/H1N1 et niveau	E. d'Alessan	CNRS, UMI	Bulletin de l	0037-9085	1961-9049			Lavoisier	journal	research-art	2011	Français	Résumé: Dar Pandémie grippale ; Professionnels de santé ; Risque infectieux ;	8.702	1.3	Absent	888F984332	ark:/67375/i10.1007/s13149-011-0179	10.1007/s13149-011-0179												
6	sars-mers_00007	Planetary science: Mission to Earth's c	David J. Stev	California In	Nature	0028-0836				Nature	journal	research-art	2003	Anglais	Not science fiction, but a 1 - science ; 1 - general ; 1 - General ; 1 - sciences	4.012	1.4	Absent	4AA9A1CAT	ark:/67375/10.1038/423	12748631												
7	sars-mers_00008	Première étude sur le dépistage et	A.-J. Rémy	Fédération d	Journal Afric	1954-3204	1954-3212			Lavoisier	journal	research-art	2008	Français		2.77	1.3	Absent	D2F268A1FB	ark:/67375/10.1007/s1257-008-0044	10.1007/s1257-008-0044												
8	sars-mers_00395	RNA aptamer-based sensitive detectio	Dae-Gyun Al	Department	The Analyst	0003-2654	1364-5528			RSC [journ	journal	other	2009	Anglais	Severe acute respiratory s 1 - science ; 1 - natural sc 1 - Physical S 1 - sciences	7.925	1.6	Absent	A9D02044E	ark:/67375/10.1039/b90	19684916												
9	sars-mers_00009	Short burst oxygen therapy for relief	CM Roberts	Corresponde	Thorax	0040-6376	1468-3296			BMJ	journal	editorial	2004	Anglais	oxygen ; brez 1 - science ; 1 - health sci 1 - Health Sciences ; 2 - M	7.012	1.2	Absent	0E3A17F898	ark:/67375/i10.1136/thx.	15282379												
10	sars-mers_00010	Regulation: the art of control? Regulat	D S Robinson	Corresponde	Thorax	0040-6376	1468-3296			BMJ	journal	editorial	2004	Anglais	asthma ; allé 1 - science ; 1 - health sci 1 - Health Sciences ; 2 - M	7.012	1.2	Absent	CD157E2382	ark:/67375/i10.1136/thx.	15282380												
11	sars-mers_00011	Antiviral agents and corticosteroids in	[W CY ^u	Department	Thorax	0040-6376	1468-3296			RSC [journ	journal	other	2004	Anglais	severe acute 1 - science ; 1 - health sci 1 - Health Sciences ; 2 - M	7.012	1.2	Absent	6B109A2002	ark:/67375/10.1036/thx.	15282381												
12	sars-mers_00056	Contents and Highlights in Chemical Technology			The Analyst	0003-2654	1364-5528			RSC [journ	journal	other	2009	Anglais	1 - science ; 1 - natural sc 1 - Physical Sciences ; 2 - T	7.012	1.6	Absent	8432AC434	ark:/67375/10.1039/b9119616	10.1039/b9119616												
13	sars-mers_00012	An initial investigation of the associati	Jianglu Tan	Shanghai Un	Journal of Ep	0143-005X	1470-2738			BMJ	journal	other	2005	Anglais	Objective: Tc severe acute 1 - social sci 1 - health sci 1 - Health Sc 1 - sciences	9.187	1.3	Absent	8F5A9A4509	ark:/67375/i10.1136/jec	15709076												
14	sars-mers_00250	Carrier-resolved technology for homog	Huan Li ^u	School of Ph	The Analyst	0003-2654	1364-5528			RSC [journ	journal	other	2008	Anglais	For clinical diagnosis, a 1 - science ; 1 - natural sc 1 - Physical S 1 - sciences	10.12	1.2	Absent	60B36F0C01	ark:/67375/10.1039/b90	18709199												
15	sars-mers_00013	Advantages to being different	Lucy Bird	Nature Revie	1474-1733	1474-1741			Nature	journal	article	2004	Anglais	1 - science ; 1 - health sci 1 - Life Sciences ; 2 - Imm	2.776	1.5	Absent	7F29064632	ark:/67375/10.1038/nrt	1427													
16	sars-mers_00396	Syntheses, properties and uses of bacte	Nicholas Thc	Cavendish La	Soft Matter	1744-683X	1744-6848			RSC [journ	journal	other	2010	Anglais	Bacterial storage lipids in 1 - science ; 1 - natural sc 1 - Physical S 1 - sciences	8.332	1.6	Absent	49F04C2D80	ark:/67375/10.1039/b927559													
17	sars-mers_00014	Stratégies, espaces et territoires	Jacques Lau	ESC Rouen	Revue Fran	0338-4551	1777-5663			Lavoisier	journal	other	2008	Anglais		7.012	1.5	Absent	A9F8871D51	ark:/67375/10.3166/rfg.	184.91-103												
18	sars-mers_00065	Nucleotides and Nucleic Acids: Oligo-	David Loake	A medical Re	Organophos	0305-9800	1465-1882	978-1-84755	978-1-84793	RSC [e-book	book-series	other	2010	Anglais		7.012	1.3	Absent	32795E04BF	ark:/67375/10.1039/b9847930899													
19	sars-mers_00066	NMR of proteins and nucleic acids	P. J. Simpson	a Cross-Faci	Nuclear Mag	0305-9800	1465-1882	978-1-84755	978-1-84793	RSC [e-book	book-series	other	2010	Anglais		7.012	1.3	Absent	898D3F0103	ark:/67375/10.1039/b9847930846													
20	sars-mers_00434	New development of glycan arrays	Chung-Yi Wu	The Genomi	Organic & Bi	1477-0520	1477-0539			RSC [journ	journal	other	2009	Anglais	The development of glyca 1 - science ; 1 - natural sc 1 - Physical S 1 - sciences	8.032	1.2	Absent	1C574B6D04	ark:/67375/10.1039/b90	19462030												
21	sars-mers_00017	Matière préliminaire	Alain Pellec	Recueil des cours						Brill HACCO	reference-w	2007	Anglais		7.012	1.6	Absent	B5A1DA81	ark:/67375/10.1163/ej.978900416615														
22	sars-mers_00210	Association of ICAM3 Genetic Variant	Kevin Y. K. C	Department	The Journal	0022-1899	1537-6613			OUP	journal	research-art	2007	Indétermine	Genetic polymorphisms hi 1 - science ; 1 - health sci 1 - Health Sc 1 - sciences	6.92	1.2	Absent	74C96F666B	ark:/67375/10.1086/518	17570115												
23	sars-mers_00211	Open Reading Frame 8a of the Human C	Yi-Chen Xh	Institut de Pu	The Journal	0022-1899	1537-6613			OUP	journal	research-art	2007	Indétermine	Background. A unique gen 1 - science ; 1 - health sci 1 - Health Sc 1 - sciences	8.079	1.4	Absent	9E11CE49675	ark:/67375/10.1086/518	17597455												
24	sars-mers_00018	Host factors and disease severity in tw	Volker Riecke	Medizinische	Laboratorium	0342-3026	0225-8466			Degruyter [j	journal	research-art	2006	Anglais	Infection wit comorbidity 1 - science ; 1 - health sci 1 - Health Sc 1 - sciences	4.034	1.3	Absent	64636E6F8B	ark:/67375/10.1515/jln	2006.003												
25	sars-mers_00019	Viral lower respiratory tract infection	J B M van W	Emma Childi	BMJ	0959-8138	1468-5833			BMJ	journal	other	2003	Anglais	1 - health sci 1 - Health Sciences ; 2 - M	5.676	1.4	Absent	2AC0E3B946	ark:/67375/10.1136/bm	12842956												
26	sars-mers_00020	Chapter I - Preliminary issues	A.V.M. Struyven	Recueil des cours						Brill HACCO	reference-w	2004	Anglais		7.012	1.6	Absent	6914BE6729	ark:/67375/10.1163/ej.97890041553														
27	sars-mers_00221	Survey of the year 2004 commercial o	Rebecca L. R	Center for Bi	Journal of M	0952-3499	1099-1352			Wiley	journal	review-articl	2005	Anglais	The year 2004 affinity ; Bla 1 - science ; 1 - health sci 1 - Life Sci 1 - sciences	9.184	1.3	Absent	432AB0D01	ark:/67375/10.1002/jmr	16252250												
28	sars-mers_00022	L'enzyme de conversion des angiotens	Guillaume L	UMR M100	Journal de la	1295-0661	1760-6128			EDP Science	journal	research-art	2010	Anglais	L'Enzyme de Angiotensin-converting en 1 - health sci 1 - Life Sciences ; 2 - Blo	9.892	1.3	Absent	C554647AB6	ark:/67375/10.1051/bc/2009032													
29	sars-mers_00023	Déterminants de l'orientation Yvel Imen Zelli		Institut Supé	Revue Franc	0338-4551	1777-5663			Lavoisier	journal	research-art	2010	Anglais	Cet article fait le point sur l'état d'avancement des travaux traitant le Yvel Ma	7.756	1.4	Absent	EC156G5D49	ark:/67375/10.3166/rfg.	07-63-82												
30	sars-mers_00001	CHAPTER 9 - Virus-coded Ion Channels	Stephen Griff	a Leeds Insti	Successful S	2041-3203	2041-3211	978-1-84793	978-1-84793	RSC [e-book	book-series	research-art	2013	Anglais	Ion channels constitute effective drug targets for myriad human 1 - sciences	9.352	1.3	Absent	B8E946452F	ark:/67375/10.1039/b9847937814													
31	sars-mers_00025	Investigation d'une épidémie hospitali	C. Aumen	Service d'hyg	Réanimatior	1624-0693	1951-6959			Lavoisier	journal	research-art	2011	Anglais	Résumé: Les Épidémie ; Enquête ; Rougeole ; Méningococcémie ; Hôpital ; Outt	7.816	1.3	Absent	0FE15D0748	ark:/67375/10.1007/s13146-011-0345													
32	sars-mers_01744	Comparative sequence analysis of full	J. E. Phillips	Department	The Analyst	0003-2654	1364-5528			Springer [j	journal	research-art	2013	Anglais	Feline Infect: Feline infectious peritonitis virus ; Feline enteric coronavirus ; Pat	8.07	1.4	Absent	B6C960457B	ark:/67375/10.1007/s1162-011-0397													
33	sars-mers_00363	An inexpensive and portable microchip	Govind V. Ka	Applied Mini	RSC [journ	0003-2654	1364-5528			RSC [journ	journal	other	2008	Anglais	We present an inexpensive 1 - science ; 1 - natural sc 1 - Physical S 1 - sciences	9.28	1.6	Absent	742E4F89B9	ark:/67375/10.1039/b71	18299747												
34	sars-mers_00364	Contents		New Journal	1144-0546	1369-9261			RSC [journ	journal	other	2008	Anglais	1 - science ; 1 - natural sc 1 - Physical Sciences ; 2 - T	7.012	1.6	Absent	07324A7D8C	ark:/67375/10.1039/b8c2827N														
35	sars-mers_00004	Chapter 3 - Pharmacophore-based Virt	Christian Lag	Departmer	Chemoinform	0920-4593	978-0-85404	978-0-85404	978-0-85404	RSC [e-book	book-series	other	2008	Anglais		7.012	1.3	Absent	4A0A2FA34L	ark:/67375/10.1039/b9847558879													
36	sars-mers_00027	Introduction au Be Congrès int			Proceedings	0027-0098				RSC [journ	journal	other	2010	Anglais		7.012	1.3	Absent	64881C9D90	ark:/67375/10.1039/b90													
37	sars-mers_00028	Huitième congrès internationa																															
38	sars-mers_00029	Virology: SARS virus infection																															
39	sars-mers_00069	Subject Index																															
40	sars-mers_00031	Dix-huitième réunion du Comi																															
41	sars-mers_00032	Faster drugs for unknown bug:																															
42	sars-mers_00033	Structural genomics of infecti																															
43	sars-mers_00044	NMR of carbohydrates, lipids and me	W. Swieczka	a Institute of	Nuclear Mag	0305-9804	1465-1882	978-1-84793	978-1-84793	RSC [e-book	book-series	other	2011	Anglais		7.012	1.3	Absent	BDS5B1D5Z	ark:/67375/10.1039/b98479372796													
44	sars-mers_00035	CONTENTS																															
45	sars-mers_00036	Angiotensin-converting enzyme 2 is a	Wenhui Li ^u	Partners AID	Nature	0028-0836	1476-4679			Nature	journal	other	2003	Anglais	1 - science ; 1 - natural sc 1 - General ; 2 - Multidisc	6.478	1.4	Absent	88F9C21567	ark:/67375/10.1038/nat	14647384												
46	sars-mers_00016	Chapter 3 - Antisense Morpholino Olig	Hong M. Mo	1 AVI BioPha	Therapeutic	1757-7152	1757-7160	978-0-85404	978-0-85404	RSC [e-book	book-series	chapter	2008	Anglais		7.012	1.3	Absent	42020003	ark:/67375/10.1039/b984755875													
47	sars-mers_00468	Contents and Chemical Technology																															
48	sars-mers_00361	[18F]- and [11C]-Labeled N-benzyl-isaf	Dong Zhou <	Division of R	Organic & Bi	1477-0520	1477-0539			RSC [journ	journal	other	2007	Anglais	1 - science ; 1 - natural sc 1 - Physical Sciences ; 2 - T	7.012	1.6	Absent	65D4467060	ark:/67375/10.1039/b9174517													
49	sars-mers_00038	Le point sur la ventilation mécanique i	J. -D. Ricard	Service de R	Réanimatior	1624-0693</td																											

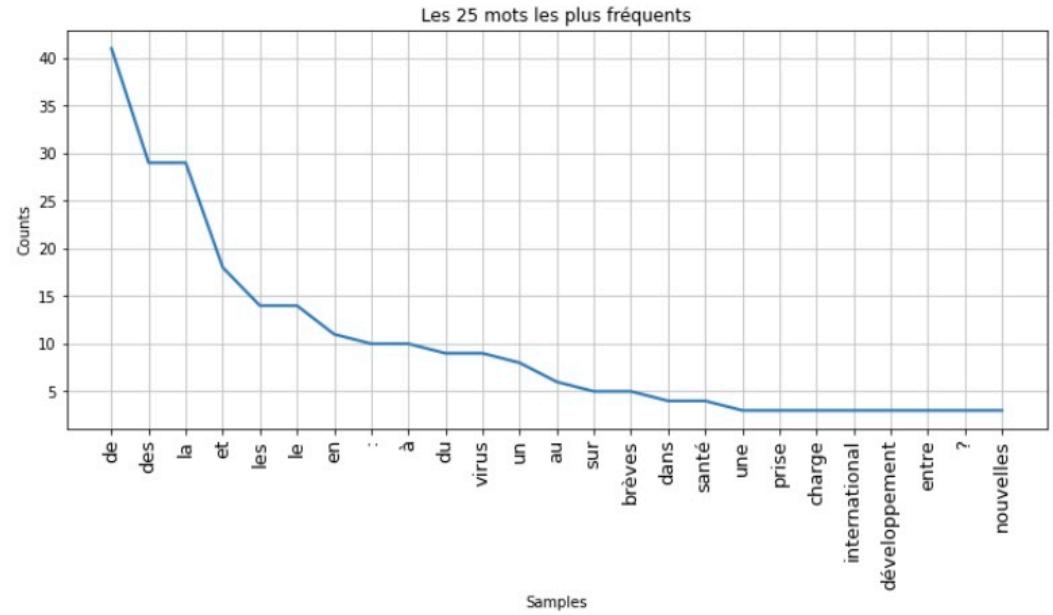
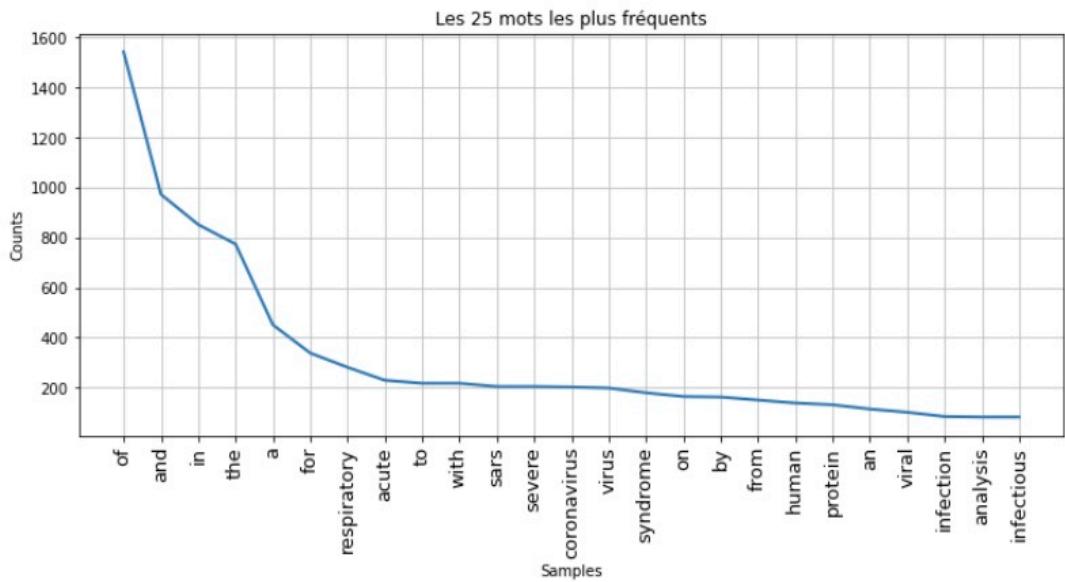
```
data = pd.read_csv(EichierCSV, sep=";", header=0, error_bad_lines=False, encoding="utf_8", usecols=[0,1,13,14,15,16])
```

```
print('Nombre total de documents : ', len(data), data.shape) data: {DataFrame: #Nombre de document par langue
print("Nombre de documents en français : ", len(data[data.Langue=='Français']))
for langue in data['Langue'].unique(): data: {DataFrame: (2532, 6)}
    print("Documents en ", langue, " : ", len(data[data.Langue==langue])) langue
data.groupby('Langue', dropna=False).describe() data: {DataFrame: (2532, 6)}
```

```
for langue in data['Langue'].unique(): data: {DataFrame: (2532, 6)}
    mots_des_titres = []
    for titre in list(data['Titre'][data.Langue==langue]): data: {DataFrame: (2532
        mots = titre.split() titre: Trends in der Impfstoffentwicklung. DNA- und z
        for mot in mots: mots: ['Für', 'eine', 'Reihe', 'von', 'Infektionskrankhei
            mots_des_titres.append(mot.lower()) mots_des_titres: ['wirksamkeit', 'p
    print("pour langue ", langue, " : ", Counter(mots_des_titres).most_common(25))
```

```
Nombre total de documents : 2532 (2532, 6)
Nombre de documents en français : 67
Documents en Anglais : 2197
Documents en Français : 67
Documents en Indéterminé : 230
Documents en Allemand : 38
Les mots les plus fréquents par langue dans les titres :
pour langue Anglais : [ ('of', 1543), ('and', 973), ('in', 852), ('the', 774), ('a', 451), ('for', 338), ('respiratory', 281), ('acute', 229), ('to', 217), ('with', 217), ('sars', 204), ('severe', 204), ('coronavirus', 202), ('virus', 198), ('syndrome', 179), ('on', 164), ('by', 162), ('from', 150), ('human', 138), ('protein', 131), ('an', 114), ('viral', 101), ('infection', 84), ('analysis', 82), ('infectious', 82)]
pour langue Français : [ ('de', 41), ('des', 29), ('la', 29), ('et', 18), ('les', 14), ('le', 14), ('en', 11), ('à', 10), ('du', 9), ('virus', 9), ('un', 8), ('au', 6), ('sur', 5), ('brèves', 5), ('dans', 4), ('santé', 4), ('une', 3), ('prise', 3), ('charge', 3), ('international', 3), ('développement', 3), ('entre', 3), ('?', 3), ('nouvelles', 3)]
pour langue Indéterminé : [ ('of', 170), ('and', 112), ('in', 112), ('the', 90), ('respiratory', 85), ('acute', 74), ('severe', 65), ('a', 57), ('with', 52), ('syndrome', 49), ('coronavirus', 49), ('for', 39), ('human', 37), ('by', 30), ('infection', 25), ('to', 23), ('viral', 20), ('virus', 17), ('patients', 17), ('influenza', 16), ('from', 14), ('clinical', 13), ('on', 12), ('disease', 12), ('is', 12)]
```

```
# plot word frequency distribution of first few words
plt.figure(figsize=(12,5))
plt.title('Les 25 mots les plus fréquents')
plt.xticks(fontsize=13, rotation=90)
fd = nltk.FreqDist(mots_des_titres)
fd.plot(25,cumulative=False)
```



Suppression des mots outils ? (*stopwords*)

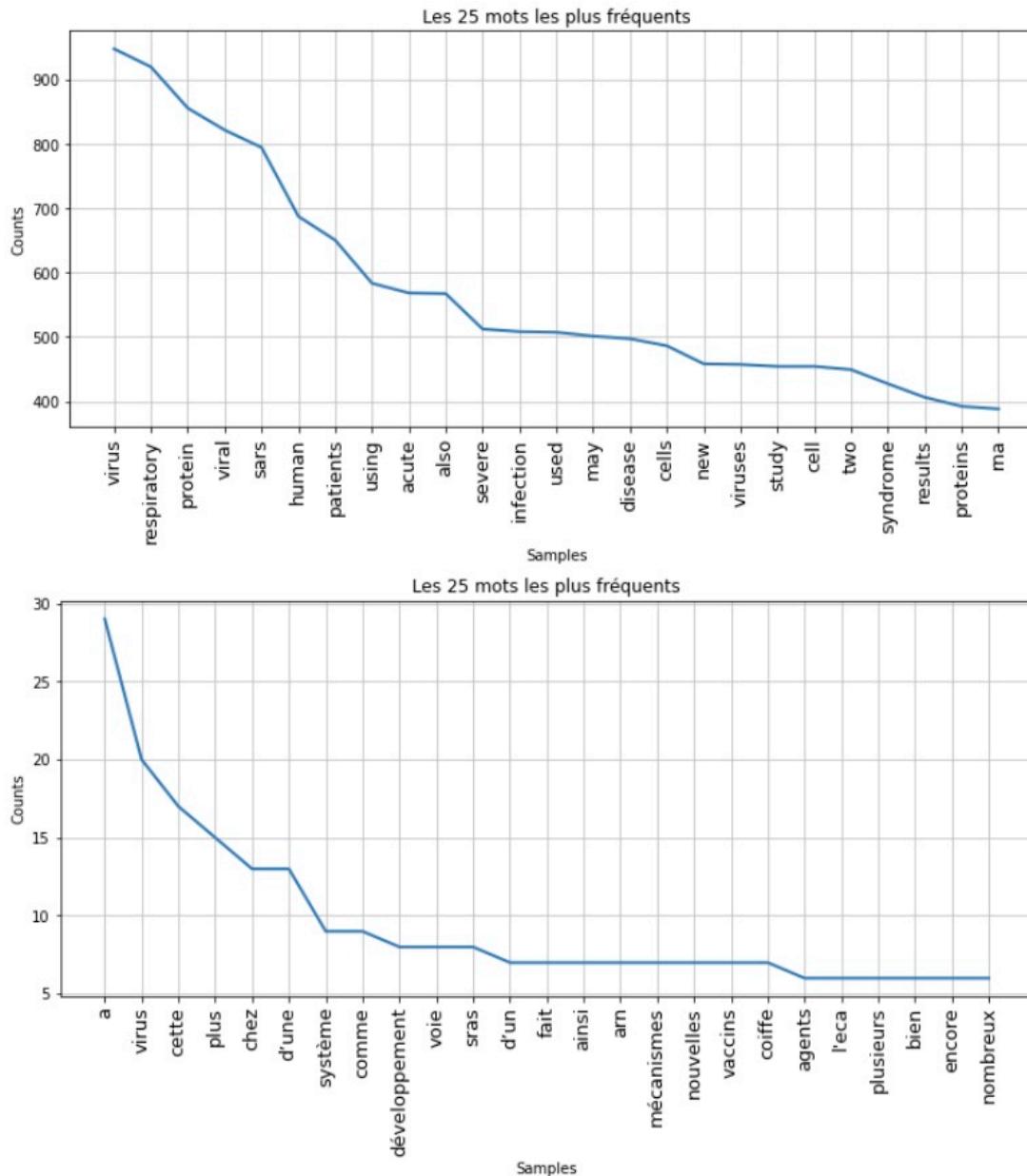
```
(base) [ Patrice Mac-Pro-de-Patrice-4 ~/nltk_data ] ls corpora/stopwords
README azerbaijani dutch finnish german hungarian italian nepali portuguese russian spanish tajik
arabic danish english french greek indonesian kazakh norwegian romanian slovene swedish turkish
(base) [ Patrice Mac-Pro-de-Patrice-4 ~/nltk_data ]
```

au	i	de	aber	إذن	и
aux	me	la	alle	إذن	в
avec	my	que	allem	إذن	во
ce	myself	el	allen	إذن	не
ces	we	en	aller	أذن	что
dans	our	y	alles	أذن	он
de	ours	a	als	أكتر	на
des	ourselv	los	also	أكتر	я
du	you	del	am	إلا	с
elle	you're	se	an	التي	со
en	you've	las	ander	اللذي	как
et	you'll	por	andere	الذين	а
eux	you'd	un	anderem	اللاتي	то
il	your	para	anderen	اللاتي	все
ils	yours	con	anderer	اللاتي	она
je	yoursel	no	anderes	اللاتي	так
la	yoursel	una	anderm	اللاتي	его
le	he	su	andern	اللاتي	но
les	him	al	anderr	اللاتي	да
leur	his	lo	anders	اللواتي	ты
lui	himself	como	auch	إلى	к
ma	she	más	auf	إليك	у
mais	she's	pero	aus	إليكم	же
me	her	sus	bei	إليكم	вы
même	hers	le	bin	إليكن	за
mes	herself	ya	bis	أم	бы
moi	it	o	bist	أنا	по
mon	it's	este	da	أنا	только
ne	its	sí	damit	أنا	ее
nos	itself	porque	dann	أنا	мне
notre	they	esta	der	أنا	было
nos	them	entre	den	أنا	пот

248	arabic
164	azerbaijani
94	danish
101	dutch
179	english
235	finnish
157	french
232	german
265	greek
199	hungarian
757	indonesian
279	italian
380	kazakh
254	nepali
176	norwegian
204	portuguese
355	romanian
151	russian
1784	slovene
313	spanish
114	swedish
162	tajik
53	turkish

```
import nltk
from nltk.corpus import stopwords

for langue in data['Langue'].unique():
    mots_des_resumes = []
    stop_words = []
    if langue.lower() == "anglais":
        stop_words = stopwords.words('english')
    elif langue.lower() == "français":
        stop_words = stopwords.words('french')
    else:
        stop_words = stopwords.words('german')
    for texte in data['Résumé'][data.Langue==langue]:
        if isinstance(texte,str):
            mots = texte.split()
            for mot in mots:
                if mot.lower() not in stop_words:
                    mots_des_resumes.append(mot.lower())
print("pour langue ", langue, " : ", Counter(mots_des_resumes).most_common(25))
```



Les mots les plus importants ? (TF-IDF)

```
from sklearn.feature_extraction.text import TfidfVectorizer

def tfidf(textes):
    tfidfVec = TfidfVectorizer(stop_words='english', use_idf=True, min_df=2,
token_pattern='[a-zA-Z]+')
    tfidfMatrice = tfidfVec.fit_transform(textes)
    print("Les 10 premiers mots : ", tfidfVec.get_feature_names()[:10])
    return tfidfMatrice

tfidfMatrice = tfidf(resumesClasses['Resume'])

print("Taille matrice : ", tfidfMatrice.shape)
print("Représentation 1er résumé : ", tfidfMatrice[0,:])
```

1276 documents, 7455 mots différents retenus, 1 vecteur par doc.

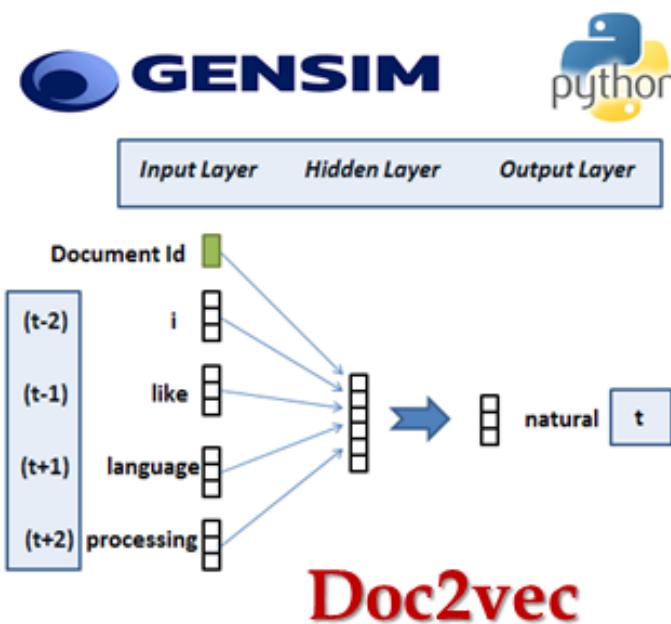
```
--  
Les 10 premiers mots : ['aa', 'ab', 'aberrant', 'abi', 'ability', 'able', 'abnormal', 'abnormalities', 'abo', 'abolish']  
Taille matrice : (1276, 7455)  
Représentation 1er résumé : (0, 2868) 0.2704313250945887  
(0, 6132) 0.24806386876009914  
(0, 3041) 0.23801900472661866  
(0, 1947) 0.09655372213525518  
(0, 1341) 0.2819297787345764  
(0, 3595) 0.1971843103507599  
(0, 1943) 0.17111132343868535  
(0, 5843) 0.17369088353073145  
(0, 533) 0.24272671127061593  
(0, 1949) 0.12103925328022304  
(0, 3202) 0.08852133361972166  
(0, 7069) 0.18858645715511513  
(0, 4125) 0.29376898845165716  
(0, 1403) 0.25422516491060365  
(0, 4331) 0.12637641076970627  
(0, 7072) 0.1328561891548727  
(0, 1457) 0.23801900472661866  
(0, 6230) 0.1291923149739438  
(0, 4024) 0.20415309475214133  
(0, 737) 0.35101596641375654  
(0, 6505) 0.2859850323557227
```

scores tf.idf

n° mot

Représentation en espaces réduits

Espace initial en très grande dimension (la taille du vocabulaire) :
 — réunir les mots similaires = projeter les documents sur un espace réduit



```
from gensim.models import Doc2Vec

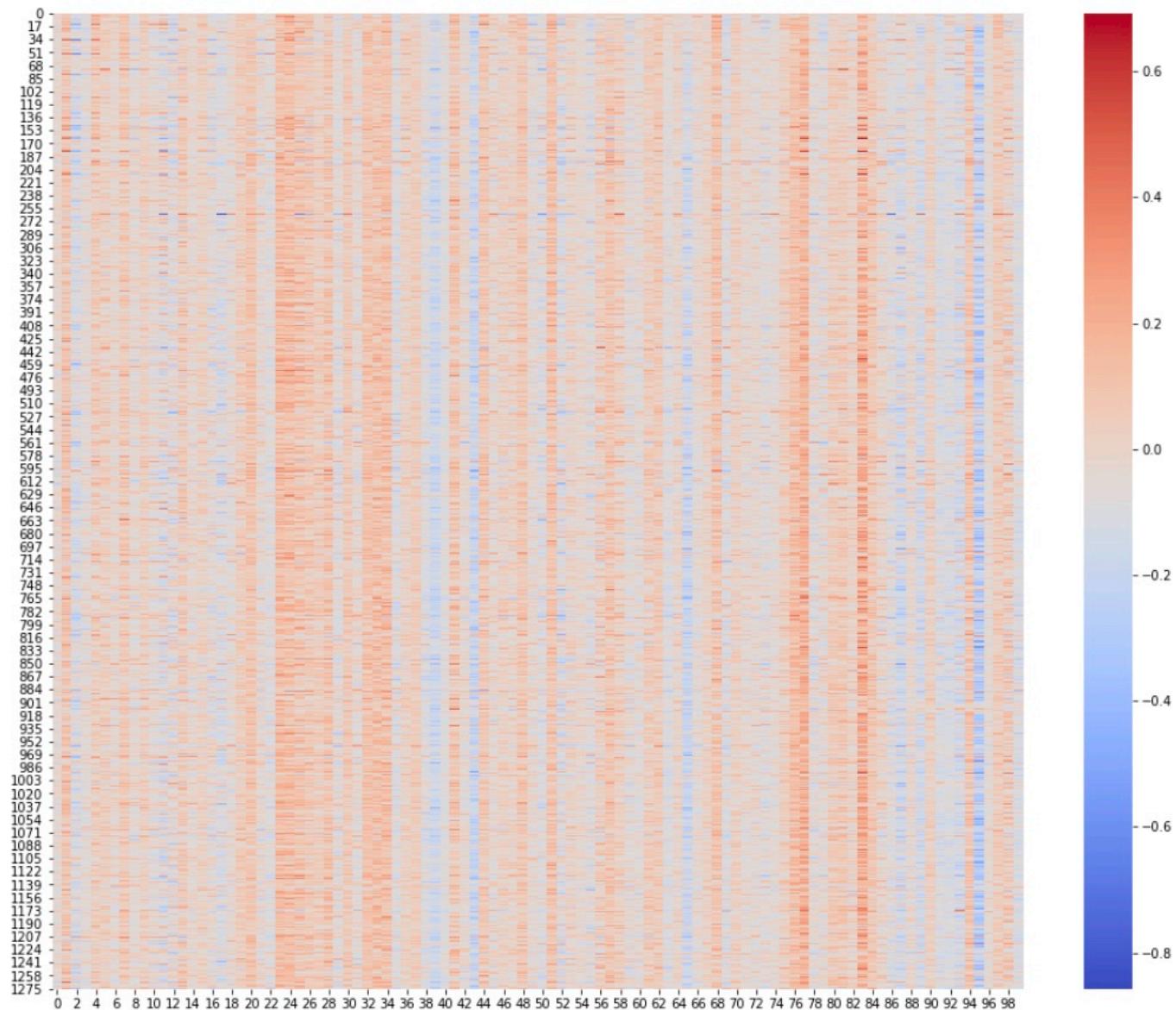
def doc2vect():
    document_tagged = []
    tagged_count = 0
    for _ in resumesClasses['Resume'].values:
        document_tagged.append(gensim.models.doc2vec.TaggedDocument(_, [tagged_count]))
        tagged_count += 1
    d2v = Doc2Vec(document_tagged)
    return d2v.docvecs.vectors_docs

%time doc2vec = doc2vect()

import seaborn as sn
plt.figure(figsize=(17,14))
sn.heatmap(doc2vec, cmap = "coolwarm")
```

CPU times: user 6.05 s, sys: 252 ms, total: 6.3 s
Wall time: 3.01 s

6 <AxesSubplot:>



Classification non supervisée (k-moyennes)

```
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.cm
import numpy as np

kmean_model = KMeans(n_clusters=3, n_jobs=-1)
#time km = kmean_model.fit_predict(doc2vec)
print ("intertie intra-classes : ", kmean_model.inertia_)
```

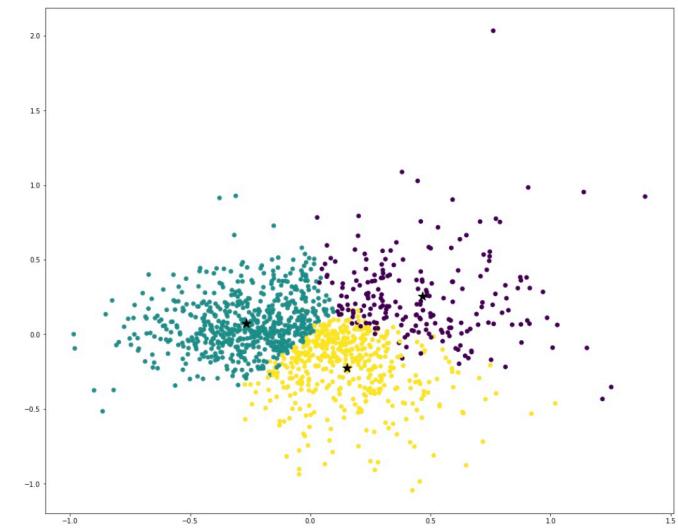
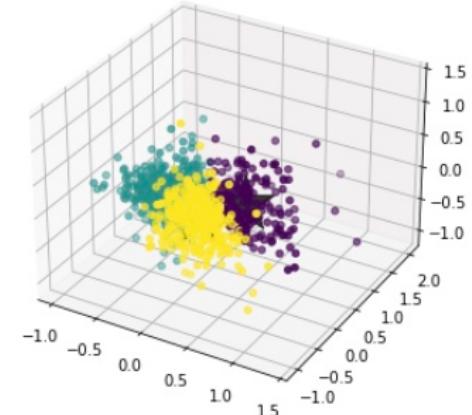
9 print(kmean_model.labels_)

[0 1 2 ... 1 1 0]

```
pca = PCA(n_components=3).fit(doc2vec)
data = pca.transform(doc2vec)
centroids = pca.transform(kmean_model.cluster_centers_)

#color = matplotlib.cm.rainbow(np.linspace(0, 1,
len(kmean_model.labels_)))
color = kmean_model.labels_
plt.figure(figsize=(50,20))
axis = Axes3D(plt.figure())
axis.scatter(data[:, 0], data[:, 1], data[:, 2], c =
color)
axis.scatter(centroids[:, 0], centroids[:, 1],
centroids[:, 2], marker='*', s=1500, c='#000000')
plt.show()

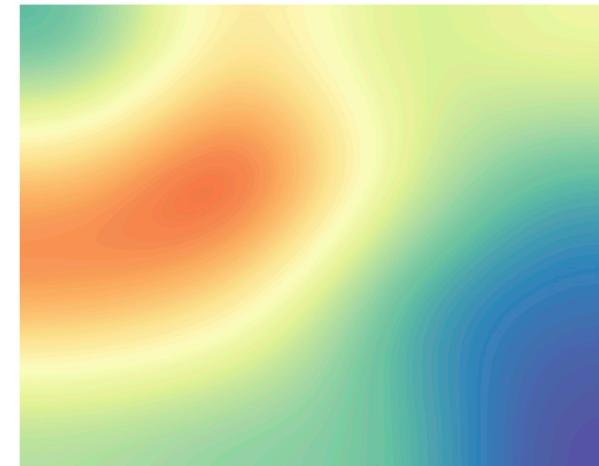
lt.figure(figsize=(17,14))
plt.scatter(data[:, 0], data[:, 1], c = color)
plt.scatter(centroids[:, 0], centroids[:, 1],
marker='*', s=200, c='#000000')
plt.show()
```



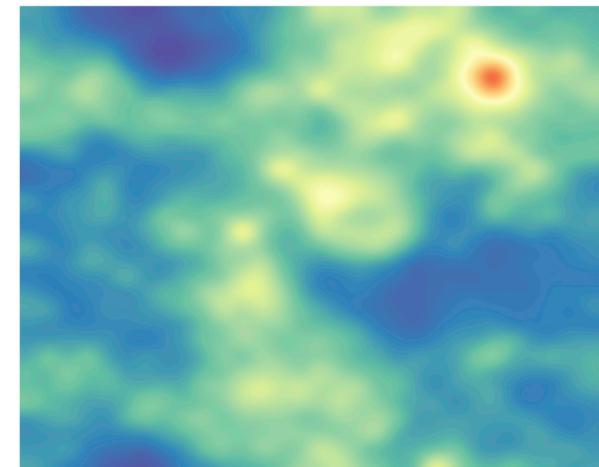
Cartographie documentaire (cartes auto-organisées)

```
import somoclu

som = somoclu.Somoclu(4, 4, maptype="toroid")
som.train(doc2vec)
som.view_umatrix(bestmatches=False,
figsize=(17,14))
```



```
som = somoclu.Somoclu(30, 30, maptype="toroid")
som.train(doc2vec)
som.view_umatrix(bestmatches=False,
figsize=(17,14))
```



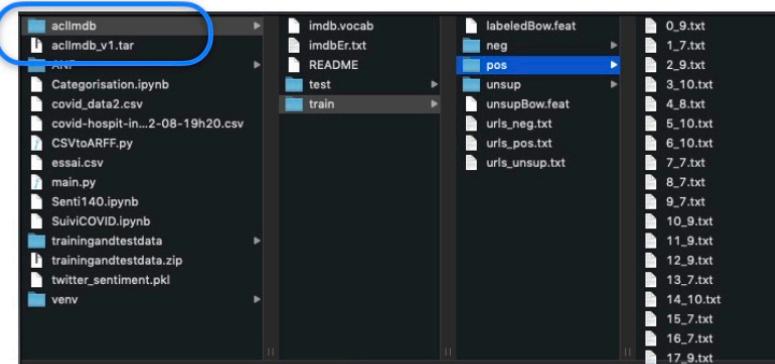


Exemple 2

Vers une analyse de sentiments automatisée

Large Movie Review Dataset

<http://ai.stanford.edu/~amaas/data/sentiment/>



Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 142-150).

4.3.2 IMDB Review Dataset

We constructed a collection of 50,000 reviews from IMDB, allowing no more than 30 reviews per movie. The constructed dataset contains an even number of positive and negative reviews, so randomly guessing yields 50% accuracy. Following previous work on polarity classification, we consider only highly polarized reviews. A negative review has a score ≤ 4 out of 10, and a positive review has a score ≥ 7 out of 10. Neutral reviews are not included in the dataset. In the interest of providing a benchmark for future work in this area, we release this dataset to the public.²

Corpus d'entraînement (train) : 12 500 positives, 12 500 négatives

This is a complex film that explores the effects of Fordist and Taylorist modes of industrial capitalist production on human relations. There are constant references to assembly line production, where workers are treated as cogs in a machine, overseen by managers wielding clipboards, controlling how much time the workers leave exposed, and firing workers (Stanley) who meet all criteria (as his supervisor says, are always on time, are hard workers, do good work) but who may in some unspecified future make a mistake. This system destroys families - Stanley has to send his father to a nursing home (here he quickly dies) after Stanley loses his job. Iris' daughter is a single teen mother who drops out of high school to take a job in the plant. References are made to the fact that now, with declining wages, both partners need to work, the implication being that there's nobody left at home to care for the kids. Iris' husband is dead from an illness, and with the multiple references in the film about the costs of medical care, the viewer must wonder if he might have lived with better and more costly care. Iris' brother in law gets abusive after yet another unsuccessful day at the unemployment office when his wife yells at him for buying a beer with her savings instead of leaving it for her face lift and/or teeth job (even the working class with no stake in conventional bourgeois notions of perfection and beauty buy into them). The one reference to race in the film is through a black factory line worker whose husband is in jail (presumably, he's also black, and black men suffer disproportionately high incarceration rates). She remarks that he, like her, "is doing time" - her family is composed of a prisoner and a wage slave. Stanley, however, still believes in human relations and is therefore fond of the film outside of the system of Fordist capitalism. He cares for his father in spite of the fact that it was his father's traveling salesman job that resulted in his illiteracy - he has not yet reduced human relations to a purely instrumental contract, as Iris' brother in law does (suggesting that he married the wrong sister). He does not, as Iris says, conform to the work-eat-sleep routine of everyone else; rather, he uses technology and the techniques of industrial production in an artisanal and creative way, in a sort of Bauhaus ideal. This was the dream of early modernists and 1920's socialists.

Pré-traitements du corpus

La première étape consiste à intégrer l'ensemble des critiques annotées (polarité négative ou positive) en un seul fichier au format CSV qui pourra être stocké en mémoire par un DataFrame (extension Pandas de Python).

taille du fichier movie_dataset.csv :
65,9 Mo (50 000 lignes, 14 millions
de *tokens*, 194 758 mots différents)

```
1 # Conversion du corpus d'origine en un fichier .csv

import pandas as pd
import os

repertoire_depart = '/Users/Patrice/PycharmProjects/ANF2021/aclImdb'

labels = {'pos':1, 'neg' : 0}
df = pd.DataFrame()
for f in ('test', 'train'):
    for l in ('pos', 'neg'):
        path = os.path.join(repertoire_depart, f, l)
        for fichier in os.listdir(path):
            with open(os.path.join(path, fichier), 'r', encoding='utf-8') as infile:
                txt = infile.read()
            df = df.append([[txt, labels[l]]], ignore_index=True)
df.columns=['review', 'polarity']

df.to_csv('movie_data.csv', index=False, encoding='utf-8')
df.head()
```

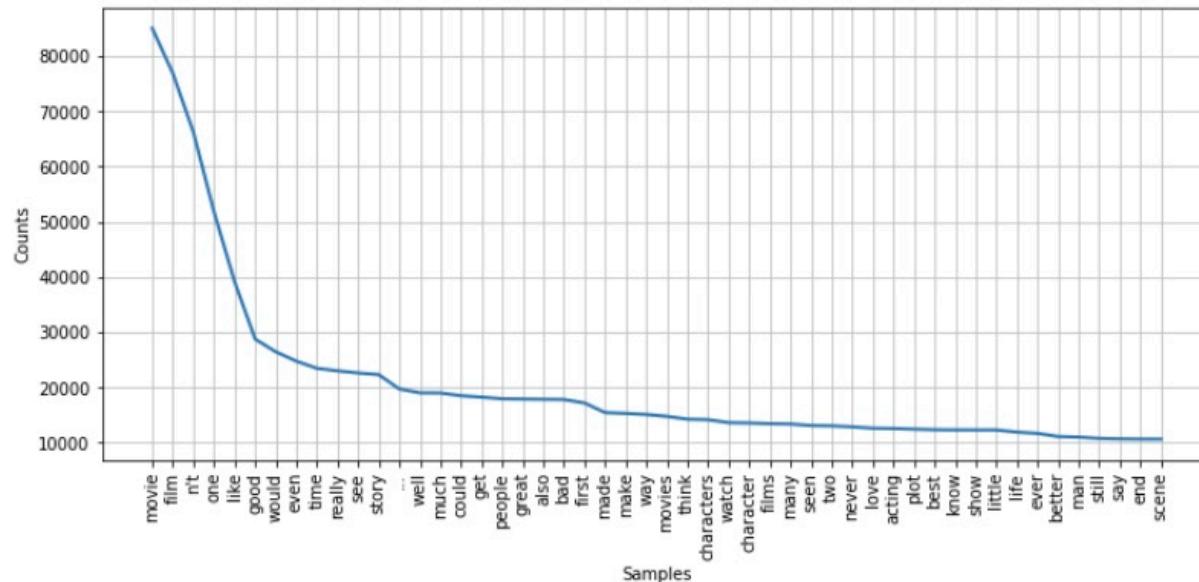
	review	polarity
0	Based on an actual story, John Boorman shows t...	1
1	This is a gem. As a Film Four production - the...	1
2	I really like this show. It has drama, romance...	1
3	This is the best 3-D experience Disney has at ...	1
4	Of the Korean movies I've seen, only three had...	1

```

import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords

#Concatenate strings in the Series/Index with given separator.
reviews = df.review.str.cat(sep=' ')
#function to split text into word
tokens = word_tokenize(reviews)
stop_words = set(stopwords.words('english'))
tokens = [w.lower() for w in tokens if not w.lower() in stop_words and len(w)>2]
vocabulary = set(tokens)
print("Taille du vocabulaire : ", len(vocabulary))
frequency_dist = nltk.FreqDist(tokens)
sorted(frequency_dist, key=frequency_dist.__getitem__, reverse=True)[0:50]
plt.figure(figsize=(12,5))
frequency_dist.plot(50)

```

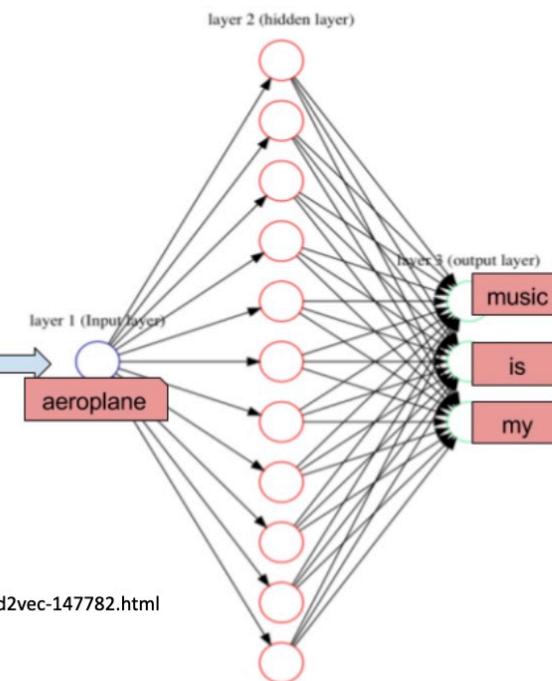
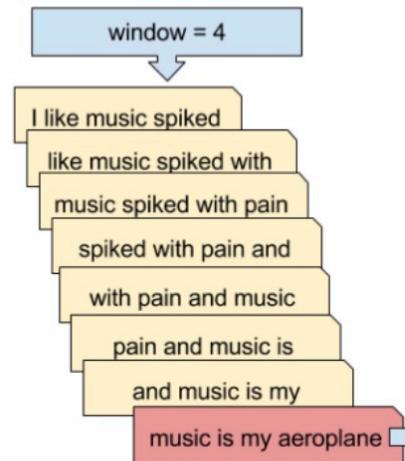




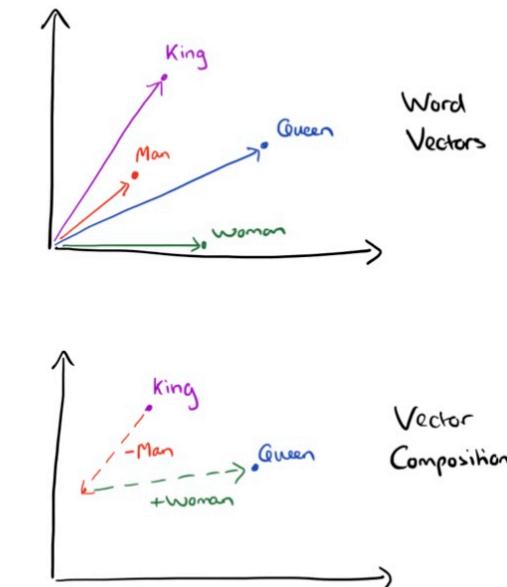
Réduction de la dimension (projection)

Les plongements de mots (word embeddings)

I like music spiked with pain and music is my aeroplane ...



Tommaso Teofili
<https://jaxenter.com/deep-learning-search-word2vec-147782.html>



Adrian Colyer
<https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vector/>

Distributed Representations of Words and Phrases and their Compositionality – Mikolov et al. 2013

Efficient Estimation of Word Representations in Vector Space – Mikolov et al. 2013

Apprentissage de représentation Word2Vec

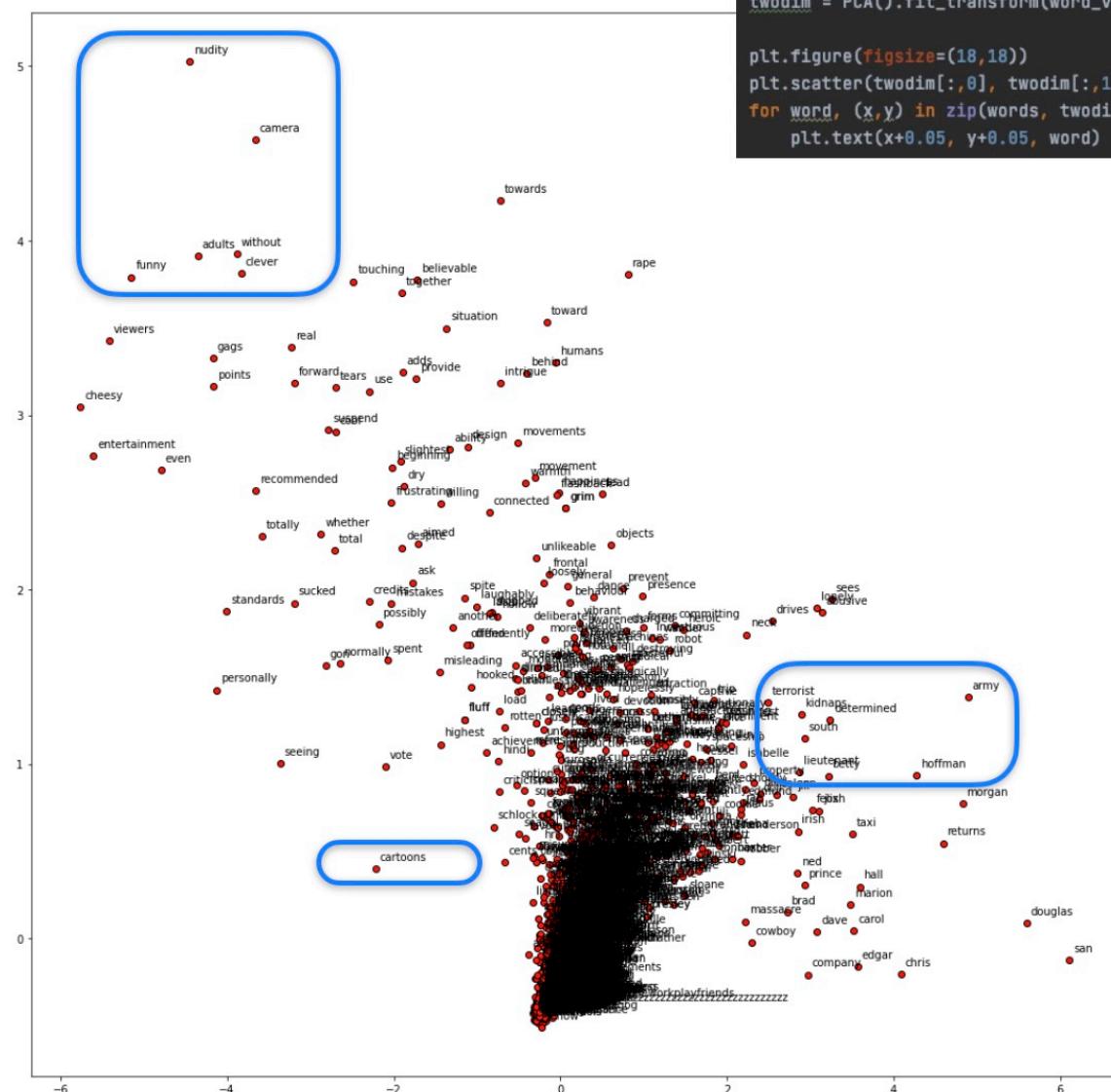
```
#L'espace de représentation est appris sur l'ensemble du corpus
for line in df['review']:
    tokens = word_tokenize(line)
    stop_words = set(stopwords.words('english'))
    tokens = [w.lower() for w in tokens if w.isalpha() and len(w)>1 and not w.lower() in stop_words]
    review_lines.append(tokens)
```

```
import gensim
model = gensim.models.Word2Vec(sentences=review_lines, size=200, window=5, workers=4, min_count=1)
motsComplet = list(model.wv.vocab)
```

```
movie -1.3216566 0.36635584 -0.28186616 -1.0511837 -1.0501945 -1.7482823 -0.42692444 0.16830114 -1.073119 -1.5651205 -0.96654 -0.54516864 1.2929311 0.49605948 1.1482662 0.38361785 -0.30000296 0.78807664 -0.62371856 -1.5082116 -0.13787036 74925745 0.41954425 0.35796735 0.3195898 -0.20374134 -0.25748256 -0.90302813 -0.44684523 -0.46419883 0.43331063 0.38018 -0.23262957 -0.57022005 -0.6890808 0.29229978 -0.06665888 -0.045591816 -0.31439704 -0.44238204 -1.19862 0.12611166 0.921796 0.17370766 0.20563798 0.8580158 0.8143437 -0.026487244 -0.12953776 1.6001002 0.2723402 0.053601284 0.440381 0.05817 0.001299 -0.14719109 -0.3533582 -1.16035 1.0383319 0.3641711 -0.29797938 -0.041548226 0.35354558 -0.7025537 0.17819 -1.3149184 0.21495351 -0.7291604 0.18647747 -1.2000268 -0.51228637 0.36612657 -0.25129464 -0.746 -0.1836730 0.6096396 0.34609687 0.40593633 0.7030198 0.023112642 -0.9067271 0.43155307 0.4280309 -0.049969178 -0.679059 0.6962609 0.16017178 0.66016424 -0.5926901 -0.013376136 -0.22369754 -1.0953285 -0.56589377 -0.42723322 0.71673262 0.8491248 -0.484025 -0.31997883 0.18664318 -0.5761222 0.33220634 -1.0463667 -0.009183551 0.5471651 0.6037895 -1.0772457 -0.646116 1.1264194 -0.9413773 0.08854891 -0.122176886 -0.056594223 0.5072317 1.13529 -0.088807 0.37230954 -0.61006385 -1.1492089 -1.5274029 -0.037806857 -0.19853547 0.2762417 -0.9356259 -0.377374 film -1.1342325 0.5424572 0.0140855415 -0.54681146 -1.0229077 -2.3149817 -0.3617721 0.08117554 -0.69557166 -1.0018283 -0.25 0.1879876 1.2258501 0.53333026 0.71119124 -1.3764403 -0.69352823 0.67989963 0.049601056 -1.0814724 -0.17875 -0.29950 0.46457902 0.110982075 0.07333746 -1.321096 -0.19277126 0.023522813 -0.31523454 -0.23818257 0.4992599 0.20365019 0.2108204 0.43208042 0.03197141 0.19413853 -0.32528928 -0.14852582 0.12936993 0.068569176 -0.36599588 0.116247706 0.68026376 0.314871 -0.30278912 0.69517577 0.4294458 -0.3990693 -0.76446646 1.5112543 0.3708154 0.11746891 0.701029 -0.7823005 1.6288 -0.26889926 1.0239882 -1.2052739 -0.047914516 0.9869529 -0.46331605 -0.07111113 0.079658456 0.37919065 -0.006453751 0.45773625 -0.58498067 0.45197055 -0.49910277 0.317274 -0.90511173 0.42767948 0.22158863 -0.068598926 0.58532935 -0.010821 2.0317261 0.7017311 0.12857646 1.0322477 0.30594614 0.5822884 -1.2792618 -0.27707702 0.5073626 0.5156112 -0.7731857 0.63963 -0.25596482 0.66147095 -0.007577596 -1.0135919 -0.37657994 0.21909198 -1.2694278 -0.758413 -0.9453872 -0.23568347 0.5475771 0.36981234 0.29823944 -0.37622204 0.22047852 0.2637362 -1.1235323 0.12577608 -0.56808615 -0.49570698 0.2905903 0.37622717 0.11014897 -1.2906862 0.10878839 0.9532716 -0.9014037 -0.41337353 0.57484233 -0.76305604 0.26593128 0.29173
```

pour chaque mot,
un vecteur en
200 dimensions

ACP
de 3000
mots pris
au hasard

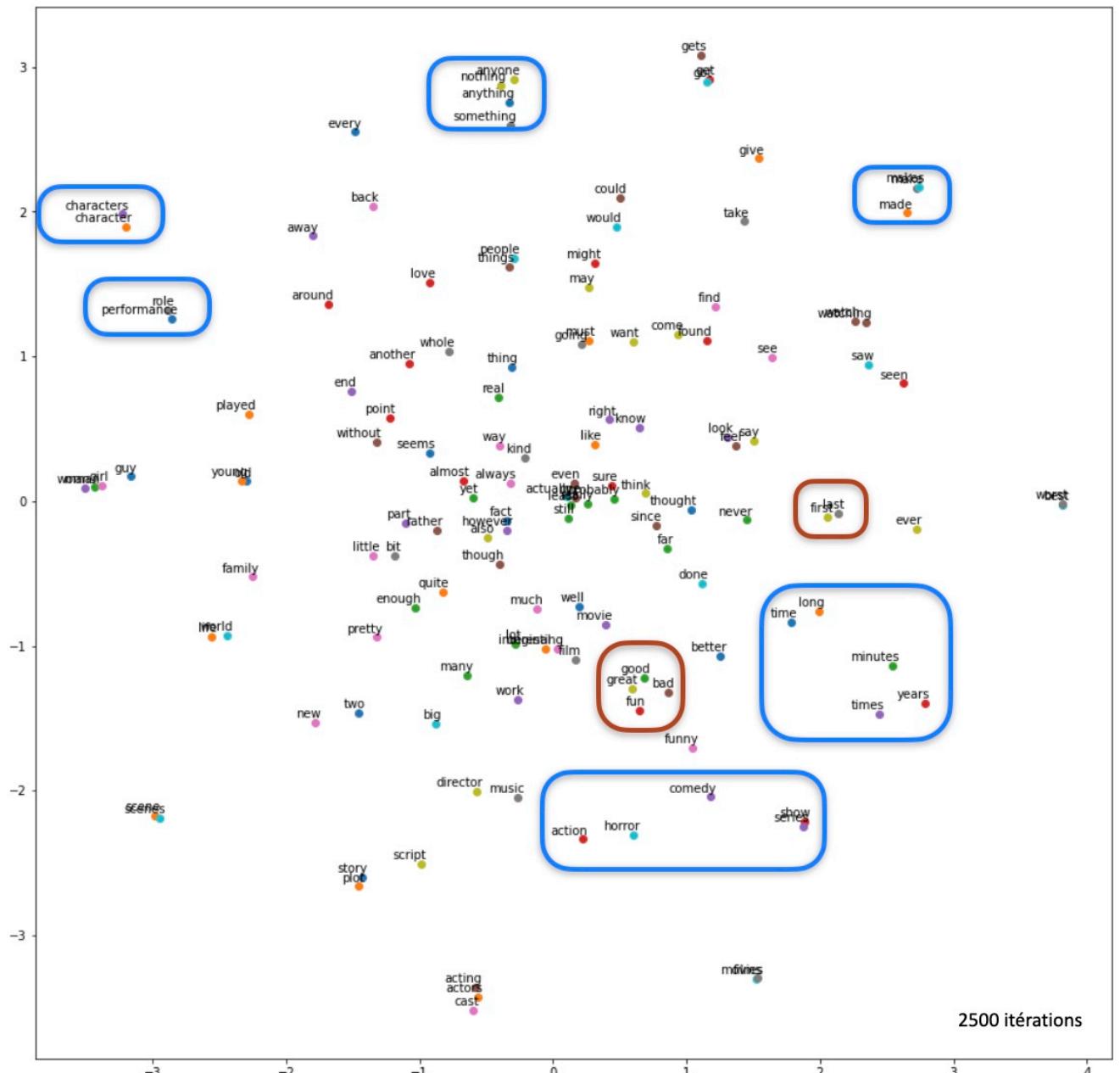


<https://web.stanford.edu/class/cs224n/materials/Gensim%20word%20vector%20visualization.html>

```
twodim = PCA().fit_transform(word_vectors)[:,2]

plt.figure(figsize=(18,18))
plt.scatter(twodim[:,0], twodim[:,1], edgecolors='k', c='r')
for word, (x,y) in zip(words, twodim):
    plt.text(x+0.05, y+0.05, word)
```

t-SNE des mots dont
occurrences > 5000
en limitant
Word2Vec
aux mots ayant
au moins
100 occurrences



Démarche générale (pré-traitements)

```
new_df = pd.DataFrame({'document':df})

# removing everything except alphabets
new_df['clean_doc'] = new_df['document'].str.replace("[^a-zA-Z#]", " ")

# removing short words
new_df['clean_doc'] = new_df['clean_doc'].apply(lambda x: ' '.join([w for w in x.split() if len(w)>3]))

# make all text lowercase
new_df['clean_doc'] = new_df['clean_doc'].apply(lambda x: x.lower())
```

```
from nltk.corpus import stopwords
stopwords = stopwords.words('english')

# tokenization
tokenized_doc = new_df['clean_doc'].apply(lambda x: x.split())

# remove stop-words
tokenized_doc = tokenized_doc.apply(lambda x: [item for item in x if item not in stopwords])

# de-tokenization
detokenized_doc = []
for i in range(len(news_df)):
    t = ' '.join(tokenized_doc[i])
    detokenized_doc.append(t)

new_df['clean_doc'] = detokenized_doc
```

<https://github.com/pbellot/ANFTDM2021>

Atelier "Apprentissage automatique pour la classification textuelle"

ANF CNRS "Exploration documentaire" 2021

Cet atelier permet d'explorer des tâches de classification automatique non supervisée (k-moyennes et cartes auto-organisées) ou supervisée (classification bayésienne, arbres de décision, réseaux de neurones profonds et plongements lexicaux). Les données sont d'une part les métadonnées de documents issus d'ISTEX concernant le mot clé "covid" (catégorisation en domaines à partir des résumés puis partitionnement) et une collection de critiques IMDB de films pour l'analyse de sentiment. Les environnements sont Jupyter Notebook et Weka. Le langage est Python.

Pour réaliser les exemples de l'atelier, vous devez :

(le plus simple) Pour une exécution du code Python dans l'environnement distant Google Colab (<http://colab.research.google.com>) :

<https://anf-tdm-2020.sciencesconf.org/329939>

ANF TDM 2020 - Exploration documentaire et extraction d'information
27-28 Janv. 2021 Paris (France)

L'apprentissage automatique pour la classification textuelle
Patrice Bellot 1

Programme en ligne

Intervenant(e)s

SUPPORT

Personnes connectées

Présentation de l'atelier

L'objectif de l'atelier est de présenter, sous forme de démonstrations et de notebooks partagés, deux environnements logiciels permettant la classification automatique de données textuelles selon des approches d'apprentissage automatique incluant les réseaux neuronaux. Cela permettra aux auditeurs de comprendre la nature et les objectifs des traitements mis en œuvre et d'estimer l'effort nécessaire pour expérimenter les approches les plus actuelles sur ses propres données.

Programme détaillé

- Manipulation des outils d'apprentissage automatique de classification supervisée (catégorisation) ou non supervisée (partitionnement en classes proches) à partir de textes.
- Classification supervisée de document (des textes annotés et des représentations vectorielles des mots et des documents) - Expérimentation avec l'environnement Weka.
- La mise en œuvre d'une approche neuronale pour la classification automatique - Expérimentation à partir d'un Notebook en Python.
- De nombreuses tâches de la fouille de textes vues comme des problèmes de classification.

www.youtube.com › watch
ANF TDM 2020 - Introduction à la fouille de texte et ... - YouTube
Présentation de la conférenceLa fouille de données textuelles informatisée met en jeu un certain nombre de ...
16 mars 2021 · Ajouté par ANF CNRS - Fouille de textes et de données

www.youtube.com › watch
ANF TDM 2020 - L'apprentissage automatique ... - YouTube
Présentation de l'atelierL'objectif de l'atelier est de présenter, sous forme de démonstrations et de Notebooks ...
16 mars 2021 · Ajouté par ANF CNRS - Fouille de textes et de données

https://www.youtube.com/channel/UCFyKvV_dkY2ww2ejT4pcF1g

Des exemples de .csv

<https://people.math.sc.edu/Burkardt/datasets/csv/csv.html>

<https://www.opendataphilly.org/dataset/parking-violations>

<https://www.kdnuggets.com/datasets/index.html>

<https://www.kaggle.com/datasets>

<https://perso.telecom-paristech.fr/eagan/class/igr204/datasets>

<https://www.dataquest.io/blog/free-datasets-for-projects/>

The screenshot shows the Kaggle homepage. On the left, a sidebar menu includes 'Home', 'Competitions', 'Datasets' (which is selected and highlighted in grey), 'Code', 'Discussions', 'Courses', and 'More'. Below this, 'Recently Viewed' datasets are listed: 'PLOTLY CHEAT SHEET...', 'E-Commerce Data', 'Customer Segmentation', 'Sentiment Lexicons for...', and 'Basic Visualization and...'. The main content area features sections for 'Datasets', 'Trending Datasets', and 'Popular Datasets'. The 'Datasets' section has a search bar and filters for 'Datasets', 'Tasks', 'Computer Science', 'Education', and 'Classification'. The 'Trending Datasets' section shows cards for 'GoodReads 100k books' (by Manav Dhamani) and 'Roller Coaster Accidents' (by steven lasch). The 'Popular Datasets' section shows cards for 'Heart Attack Analysis &' (with a heart diagram) and 'World Happiness Report' (with a landscape photo).