

APPRENTISSAGE AUTOMATIQUE POUR LA CLASSIFICATION TEXTUELLE

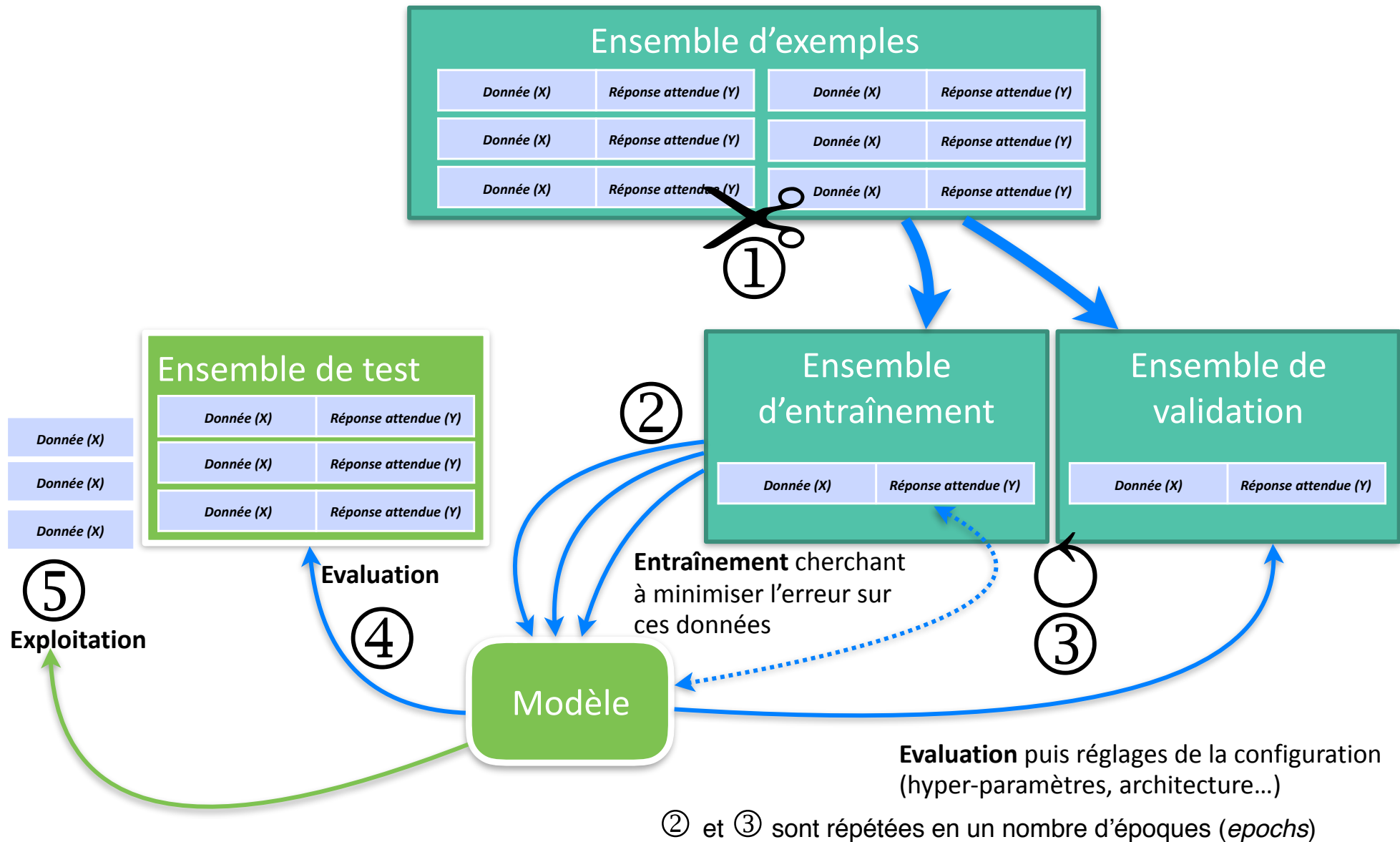
Atelier ANF CNRS - Exploration documentaire
février 2021

Patrice Bellot
Aix-Marseille Université - CNRS (LIS-INS2I)

patrice.bellot@univ-amu.fr



Apprentissage automatique, modèle, évaluation



CATEGORISATION DE TEXTES ET CLASSIFICATION NON SUPERVISÉE

Nom de fichier;Titre;Auteur(s);Affiliation(s);Revue ou monographie;ISSN;e-ISSN;ISBN;e-ISBN;Éditeur;Type de publication;Type de document;Date de r;Catégories WoS;Catégories Science-Metrix;Catégories Scopus;Catégories INIST;Score qualité;Version PDF;XML structuré;Identifiant ISTEX;ARK;DOI s_00002;Structures and diseases;"K Ulrich Wendt ¹ ; Manfred S Weiss ² ; Patrick Cramer ³ ; Dirk W Heinz ⁴ ; Sanofi-Aventis, Frankfurt, D-65926, Germany ; European Molecular Biology Laboratory, c/o DESY, Hamburg, D-22603, Germany ; Gene Centre, Ludwig of Structural Biology, Helmholtz Centre for Infection Research, Braunschweig, D-38124, Germany";Nature Structural & Molecular Biology;1545-9996
tural biology is making significant contributions toward an understanding of molecular constituents and mechanisms underlying human diseases a rnau Conference on Structural Biology of Disease Mechanisms held in September 2007 in Murnau, Germany."; "1 - science ; 2 - cell biology ; 2 - lth 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; C20103CC68E46DEBF1A30871D342FC7F50B87

Corpus SARS-MERS-Export																			
Accueil Insertion Dessin Mise en page Formules Données Révision Affichage Acrobat Dites-le nous Couper Copier Coller Mise en forme Calibri (Corps) 12																			

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

Calibri (Corps) 12

Standard

Mise en forme conditionnelle Mettre sous forme de tableau Styles de cellule Insérer Supprimer Mise en forme

Somme automatique Remplissage Trier et filtrer Rechercher et sélectionner

Idées Créer et partager un PDF Adobe

	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 m	ISSN	e-ISSN	ISBN	e-ISBN	Éditeur	Type de publi	Type de docu	Date de publ	Langue(s) du	Résumé	Mots-clés d'	Catégories V	Catégories S	Catégories S	Catégories S	Score qualit	Version PDF	XML structur	Identifiant IS	ARK	DOI	PMID
2	sars-mers_00002	Structures and diseases	K Ulrich Wer	Department	Nature Struc	1545-9993	1545-9985			Nature	book	conference	2008	Anglais	Structural biology is makir	1 - science ;	1 - health sci	1 - Life Scien	1 - sciences	5.26	1.4	Absent	C20103C68	ark:/67375/	10.1038/n	18250627	
3	sars-mers_00003	Evaluating Euro-Mediterranean Relatio	Stephen C. Calleja	Evaluating E	iro-Mediterr	nean Relatio		9,780E+12	9,7802E+12	taylor-franci	book	book	2005	Anglais	What are the prospects for the future of the Euro-Mediterranean area and wha	8.92	1.6	Absent	836466E2B2	ark:/67375/	10.4324/978	10230317647					
4	sars-mers_00005	Emerging pathogens and their implica	Roger Y. Doc	Research an	British Journ	0007-1048	1365-2141			Wiley	journal	review-articl	2012	Anglais	The threat of blood transf	1 - science ;	1 - health sci	1 - Health Sc	1 - sciences	7.792	1.3	Absent	04101E0D45	ark:/67375/	10.1111/bjh	2292410	
5	sars-mers_00006	Pandémie grippale A/H5N1 et niveau	C. 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	888F9B4332	ark:/67375/	10.1007/s13	149-011-0179					
6	sars-mers_00007	Planetary science: Mission to Earth's c	David J. Stevenson	California In	Nature	0028-0836				Nature	journal	research-art	2003	Anglais	Not science fiction, but a t	1 - science ;	1 - general ;	1 - General ;	1 - sciences	4.012	1.4	Absent	4AA98A1CAI	ark:/67375/	10.1038/423	12748631	
7	sars-mers_00008	Première étude sur le dépiage et la	A. J. Rémy	Fédération d	Journal Afric	1954-3204				Lavoisier	journal	research-art	2008	Français							2.77	1.3	Absent	D2F268A1F8	ark:/67375/	10.1007/s12	157-008-0044
8	sars-mers_00395	RNA aptamer-based sensitive detectio	Dae-Gyun Al	Department	The Analyst	0003-2654	1364-5528			RSC [journal	journal	other	2009	Anglais	Severe acute respiratory s	1 - science ;	1 - natural sc	1 - Physical S	1 - sciences	7.925	1.6	Absent	A9D0C024AE	ark:/67375/	10.1039/b9c	19648916	
9	sars-mers_00009	Short burst oxygen therapy for relief o	C M Roberts	Corresponde	Thorax	0040-6376	1468-3296			BMJ	journal	editorial	2004	Anglais	asthma ; alle	1 - science ;	1 - health sci	1 - Health Sciences ;	2 - M	7.012	1.2	Absent	0E3A17F898	ark:/67375/	10.1136/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	severe acute	1 - science ;	1 - health sci	1 - Health Sciences ;	2 - M	7.012	1.2	Absent	CD157E2382	ark:/67375/	10.1136/thx	15282380	
11	sars-mers_00011	Antiviral agents and corticosteroids in	W C Yu <sup	Department	Thorax	0040-6376	1468-3296			BMJ	journal	editorial	2004	Anglais		1 - science ;	1 - natural sc	1 - Physical Sciences ;	2 - M	7.012	1.2	Absent	6B109A2C02	ark:/67375/	10.1136/thx	15282381	
12	sars-mers_00506	Contents and Highlights in Chemical Technology			The Analyst	0003-2654	1364-5528			RSC [journal	journal	other	2009	Anglais		1 - science ;	1 - natural sc	1 - Physical Sciences ;	2 - M	7.012	1.6	Absent	8432ACB434	ark:/67375/	10.1039/b9j	19161	
13	sars-mers_00012	An initial investigation of the associati	Jianguo Tan	Shanghai Ur	Journal of Ep	0143-005X	1474-2738			BMJ	journal	other	2005	Anglais	Objective: To severe acute	1 - social sci	1 - health sci	1 - Physical Sciences ;	2 - M	7.012	1.3	Absent	8F5A9A4509	ark:/67375/	10.1136/je	15709076	
14	sars-mers_00250	Carrier-resolved technology for homoge	Lian U <sup	School of Ph	The Analyst	0003-2654	1364-5528			RSC [journal	journal	other	2008	Anglais	For clinical diagnosis, a sn	1 - science ;	1 - natural sc	1 - Physical Sciences ;	2 - M	7.012	1.2	Absent	60363FC0F1	ark:/67375/	10.1039/b8c	18709199	
15	sars-mers_00013	Advantages to being different	Lucy Bird		Nature Rev	1474-1731	1474-1741			Nature	journal	article	2004	Anglais		1 - science ;	1 - health sci	1 - Life Sciences ;	2 - Imm	2.776	1.5	Absent	72F9D64632	ark:/67375/	10.1038/nrl	427	
16	sars-mers_00396	Synthesis, properties and uses of bact	Nicholas Th	Cavendish La	Soft Matter	1744-683X	1474-6848			RSC [journal	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/b92	17559b	
17	sars-mers_00014	Stratégies, espaces et territoires	Jacques Lauer	ESC Rouen ;	Revue Franç	0338-4551	1777-5663			Lavoisier	journal	other	2008	Français							7.012	1.5	Absent	49F88715D1	ark:/67375/	10.1039/b9c	184.91-103
18	sars-mers_00065	Nucleotides and Nucleic Acids; Oligo-	David Laake	a Medical R	Organophos	0305-9804	1465-1882	978-1-84755	978-1-84973	RSC [e-book	book-series	other	2010	Anglais							7.012	1.3	Absent	32795E4D8F	ark:/67375/	10.1039/978	1849730839-0
19	sars-mers_00066	NMR of proteins and nucleic acids	P. J. Simps	a Cross-Fac	Nuclear Mag	0305-9804	1465-1882	978-1-84755	978-1-84973	RSC [e-book	book-series	other	2010	Anglais							7.012	1.3	Absent	898D03F013	ark:/67375/	10.1039/978	1849730846-0
20	sars-mers_00434	New development of glycan arrays	Chung-Yi W	The Genomi	Organic & Bi	1477-0520	1477-0539			RSC [journal	journal	other	2009	Anglais	The development of glyca	1 - science ;	1 - natural sc	1 - Physical S	1 - sciences	8.032	1.2	Absent	1C575A86D0	ark:/67375/	10.1039/b9c	19462030	
21	sars-mers_00017	Matière préliminaire	Alain Pellet		Recueil des			978-90-04-1	619-6	Brill HACCO	reference-w	collected-co	2007	Français							7.012	1.6	Absent	85A01AA81	ark:/67375/	10.1163/ej	9789004166155
22	sars-mers_00210	Association of ICAM3 Genetic Variant	Kelvin Y. K.	C Department	The Journal	1002-1899	1537-6613			OUP	journal	research-art	2007	Indéterminé	Genetic polymorphisms h	1 - science ;	1 - health sci	1 - Health Sc	1 - sciences	6.32	1.2	Absent	74C96F6E66	ark:/67375/	10.1086/518	17570115	
23	sars-mers_00211	Open Reading Frame 8a of the Human	Chia-Yen Ch	Institute of Ph	The Journal	1002-1899	1537-6613			OUP	journal	research-art	2007	Indéterminé	Background. A unique gen	1 - science ;	1 - health sci	1 - Health Sc	1 - sciences	8.079	1.4	Absent	9E1EC49675	ark:/67375/	10.1086/519	17597455	
24	sars-mers_00018	Host factors and disease severity in tw	Volker Ricke	Medizinisch	Laboratoriu	0342-3026	0025-8466			Degruyter [j	journal	research-art	2006	Anglais	Infection wit comorbidity	1 - science ;	1 - health sci	1 - Health Sciences ;	2 - M	5.676	1.3	Absent	646368E6F8	ark:/67375/	10.1515/IJLM	2006.003	
25	sars-mers_00019	Viral lower respiratory tract infection	J B M van W	Emma Child	BMJ	0959-8138	1468-5833			BMJ	journal	other	2003	Anglais							7.012	1.6	Absent	2C40E38946	ark:/67375/	10.1136/bm	12842956
26	sars-mers_00020	Chapter 1 - Preliminary issues	A.V.M. Struyken	Recueil des				9,789E+12		Brill HACCO	reference-w	book	2004	Anglais							7.012	1.6	Absent	69148E6729	ark:/67375/	10.1163/ej	9789004145553
27	sars-mers_00021	Survey of the year 2004 commercial or	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 Scien	1 - sciences	9.184	1.3	Absent	432A8D0D1J	ark:/67375/	10.1002/jmr	16252250	
28	sars-mers_00022	L'enzyme de conversion de l'angiotensin	Guillaume R	UMR M100	Journal de la	1295-0661	1760-6128			EDP Science	journal	research-art	2010	Français	L'Enzyme de Angiotensin-converting en	1 - health sci	1 - Life Sciences ;	2 - Bioc	9.892	1.3	Absent	CS54647A86	ark:/67375/	10.1051/bic	2009032		
29	sars-mers_00023	Les déterminants de l'orientation Yelc	Imen Zrelli	Institut Supé	Revue Franç	0338-4551	1777-5663			Lavoisier	journal	research-art	2010	Français	Cet article fait le point sur l'état d'avancement des travaux traitant le Yield Ma	7.756	1.4	Absent	EC156A5D49	ark:/67375/	10.3166/rfg	107.63-82					
30	sars-mers_00001	CHAPTER 9 - Virus-coded Ion Channels	Stephen Grac	a Leeds Insti	Successful S	2041-3203	2041-3211			RSC [e-book	book-series	research-art	2013	Anglais	Ion channels constitute effective drug targets for myriad human d	1 - sciences	9.352	1.3	Absent	88E946452F	ark:/67375/	10.1039/978	1849737814-0				
31	sars-mers_00025	Investigation d'une épidémie hospitali	C. Aumeran	Service d'hy	Réanimator	1624-0693	1951-6959			Lavoisier	journal	research-art	2011	Français	Résumé: Les Épidémie; Enquête; Rougeole; Méningococcémie; Hôpital; Out	7.816	1.3	Absent	0FE1F5DA7E	ark:/67375/	10.1007/s13	146-011-0345					
32	sars-mers_01744	Comparative sequence analysis of full-	J. E. Phillips	Department	Virus Genes	0920-8569	1572-994X			Springer [jo	journal	research-art	2013	Anglais	Feline infect Feline infectious peritonitis virus ; Feline enteric coronavirus ; Pat	8.07	1.4	Absent	86C960457B	ark:/67375/	10.1007/s11	162-013-0972					
33	sars-mers_00363	An inexpensive and portable microchip	Gowind V. Ka	Applied Mini	The Analyst	0003-2654	1364-5528			RSC [journal	journal	other	2008	Anglais	We present an inexpensive	1 - science ;	1 - natural sc	1 - Physical S	1 - sciences	9.28	1.6	Absent	742E4F899B	ark:/67375/	10.1039/b7i	18299747	
34	sars-mers_00364	Contents			New Journal	1144-0546	1369-9261			RSC [journal	journal	other	2008	Anglais		1 - science ;	1 - natural sc	1 - Physical Sciences ;	2 - M	7.012	1.6	Absent	0732A97D8C	ark:/67375/	10.1039/b8c	2827N	
35	sars-mers_00004	Chapter 3 - Pharmacophore-based Virt	Christian Lag	a Departme	Chemoinform	atics Appro	ches to Virtu	978-0-85404	978-1-84755	RSC [e-book	book	other	2008	Anglais							7.012	1.3	Absent	4A0A2FA34I	ark:/67375/	10.1039/978	1847558879-0
36	sars-mers_00027	Introduction au 8e Congrès International Y. Buisson	Institut de l		Bulletin de l	0037-9085				Lavoisier	journal	editorial	2010	Français							2.82	1.3	Absent	6A8B152B05	ark:/67375/	10.1007/s13	149-010-0058
37	sars-mers_00028	Huitième congrès international de la Société de pathologie exoti			Bulletin de l	0037-9085				Lavoisier	journal	abstract	2010	Français							7.012	1.3	Absent	7AFEEA16C4	ark:/67375/	10.1007/s13	149-010-0065
38	sars-mers_00029	Virology: SARS virus infection of cats a	Byron E. E. N	Institute of	Nature	0028-0836	1476-4679			Nature	journal	research-art	2003	Anglais	There is now a choice of a	1 - science ;	1 - general ;	1 - General ;	1 - sciences	3.009	1.4	Absent	822D6508FE	ark:/67375/	10.1038/425	14586458	
39	sars-mers_00069	Subject Index			Structural Vi	1757-7152	1757-7160	978-0-85404	978-1-84973	RSC [e-book	book-series	other	2011	Anglais							4.4	1.3	Absent	240E5CA120	ark:/67375/	10.1039/978	1849732239-0
40	sars-mers_00031	Dix-huitième réunion du Comité local	c. B. - A. Gaüzé	CHR de La Ri	Bulletin de l	0037-9085	1961-9049			Lavoisier	journal	research-art	2011	Français							7.012	1.3	Absent	0BA06A9BAI	ark:/67375/	10.1007/s13	149-011-0164
41	sars-mers_00032	Faster drugs for unknown bugs	Alexandra Fl	mming	Nature Rev	1474-1776	1474-1784			Nature	journal	article	2005	Anglais		1 - science ;	1 - health sci	1 - Life Sciences ;	2 - Phar	3.334	1.4	Absent	837397B28C	ark:/67375/	10.1038/nrd	858	
42	sars-mers_00033	Structural genomics of infectious dise	Robin Stacey	Acta Crystall	1744-3091					Wiley	journal	article	2011	Anglais	The Seattle SSGCID ; str	1 - science ;	2 - crystallogr	1 - Physical S	1 - sciences	8.241	1.3	Absent	4089C7BF8B	ark:/67375/	10.1107/S17	21904037	
43	sars-mers_00044	NMR of carbohydrates, lipids and men	Ewa Śwież	a Institute of	Nuclear Mag	0305-9804	1465-1882	978-1-84973	978-1-84973	RSC [e-book	book-series	other	2011	Anglais							7.012	1.3	Absent	8D58ED152	ark:/67375/	10.1039/978	1849732796-0
44	sars-mers_00035	CONTENTS			International	1955-1339	2191-0294			Degruyter [j	journal	research-art	2005	Anglais		1 - science ;	1 - natural sc	1 - Physical Sciences ;	2 - M	9.952	1.4	Absent	195FOCA28C	ark:/67375/	10.1039/978	1849732796-0	
45	sars-mers_00036	Angiotensin-converting enzyme 2 is a	Wenhui Li	Partners AID	Nature	0028-0836	1476-4679			Nature	journal	other	2003	Anglais		1 - science ;	1 - general ;	1 - General ;	2 - Multidis	6.478	1.4	Absent	88F9C21567	ark:/67375/	10.1038/nat	14647384	
46	sars-mers_00016	Chapter 3 - Antisense Morpholino Olig	Hong M. Li	1 AVI BioPh	Therapeutic	1757-7160	1757-7160	978-0-85404	978-1-84755	RSC [e-book	book-series	chapter	2008	Anglais							7.012	1.3	Absent	402D000A3	ark:/67375/	10.1039/978	1847558275-0
47	sars-mers_00468	Contents and Chemical Technology			Chemical Co	1359-7548	1364-548X			RSC [journal	journal	other	2007	Anglais		1 - science ;	1 - natural sc	1 - Physical Sciences ;	2 - M	7.012	1.6	Absent	5C9A181E4E	ark:/67375/	10.1039/b7i	1471M	
48	sars-mers_00361																										

Lire le fichier .csv en Python avec le module Pandas

```
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", usecols=[0,1,13,14,15,16])
data.info()
data.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9 entries, 0 to 8
Data columns (total 6 columns):
#   Column                Non-Null Count  Dtype
---  ---
0   Nom de fichier         9 non-null     object
1   Titre                  9 non-null     object
2   Langue(s) du document  9 non-null     object
3   Résumé                 6 non-null     object
4   Mots-clés d'auteur     4 non-null     object
5   Catégories WoS         6 non-null     object
dtypes: object(6)
memory usage: 560.0+ bytes
```

	Nom de fichier	Titre	Langue(s) du document	Résumé	Mots-clés d'auteur	Catégories WoS
0	sars-mers_00002	Structures and diseases	Anglais	Structural biology is making significant contr...	NaN	1 - science ; 2 - cell biology ; 2 - biophysic...
1	sars-mers_00003	Evaluating Euro-Mediterranean Relations	Anglais	What are the prospects for the future of the E...	NaN	NaN
2	sars-mers_00005	Emerging pathogens and their implications for ...	Anglais	The threat of infection by conventional transf...	blood transfusion ; safety ; emerging infections	1 - science ; 2 - hematology
3	sars-mers_00006	Pandémie grippale A/H5N1 et niveau de préparat...	Français	Résumé: Dans les pays industrialisés, l'émerge...	Pandémie grippale ; Professionnels de santé ; ...	NaN
4	sars-mers_00007	Planetary science: Mission to Earth's core — a...	Anglais	Not science fiction, but a technically feasi...	NaN	1 - science ; 2 - multidisciplinary sciences
5	sars-mers_00008	Première étude sur le dépistage et la prise en...	Français	NaN	NaN	NaN
6	sars-mers_00395	RNA aptamer-based sensitive detection of SARS ...	Anglais	Severe acute respiratory syndrome coronavirus .	NaN	1 - science ; 2 - chemistry, analytical
7	sars-mers_00009	Short burst oxygen therapy for relief of breat...	Anglais	NaN	oxygen ; breathlessness ; chronic obstructive ...	1 - science ; 2 - respiratory system
8	sars-mers_00010	Regulation: the art of control? Regulatory T c...	Anglais	NaN	asthma ; allergy ; immunotherapy ; T cells	1 - science ; 2 - respiratory system

Explorer les données

Données Corpus SARS-MERS-Export.csv

Mise en forme en .csv pour Weka

Lecture du fichier de départ Corpus SARS-MERS-Export.csv

```
Entrée [ ]: 1 import pandas as pd
2 import csv
3 from collections import Counter
4 import matplotlib.pyplot as plt
5 import nltk
6 from nltk.corpus import stopwords
7
8 fichierCSVEntree = "/Users/Patrice/PycharmProjects/ANF2021/ANF/CorpusCovid.csv"
9 fichierSortie = "/Users/Patrice/PycharmProjects/ANF2021/ANF/CorpusWeka.csv"
```

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 :

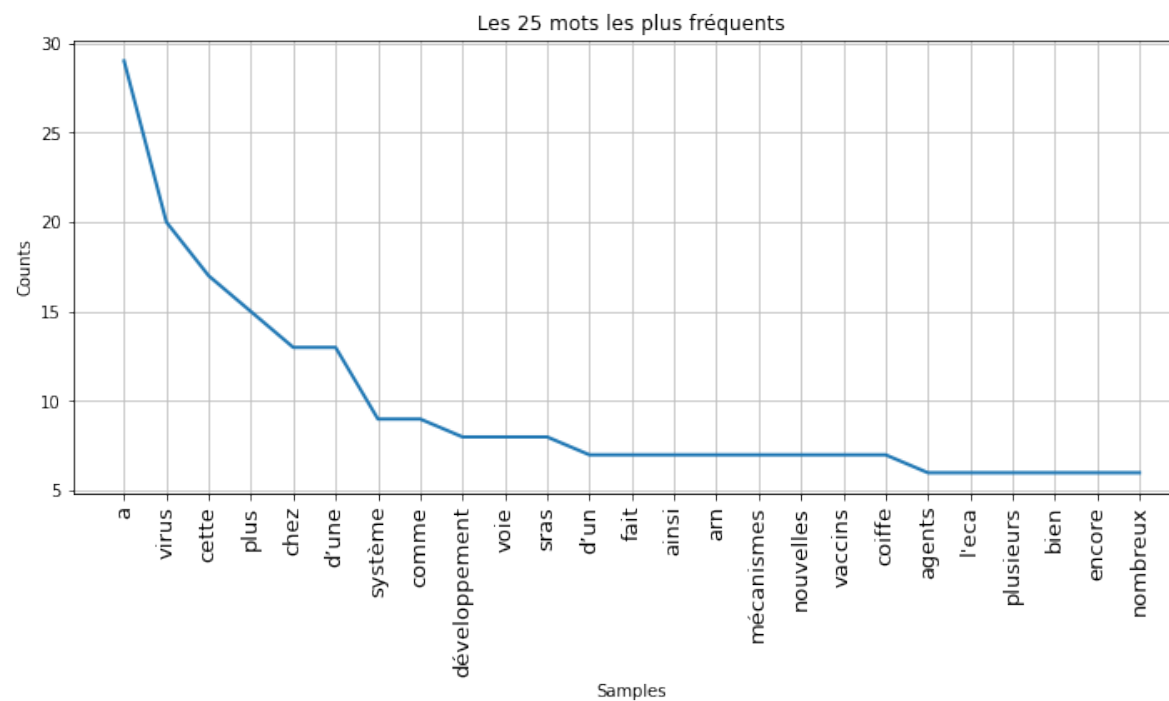
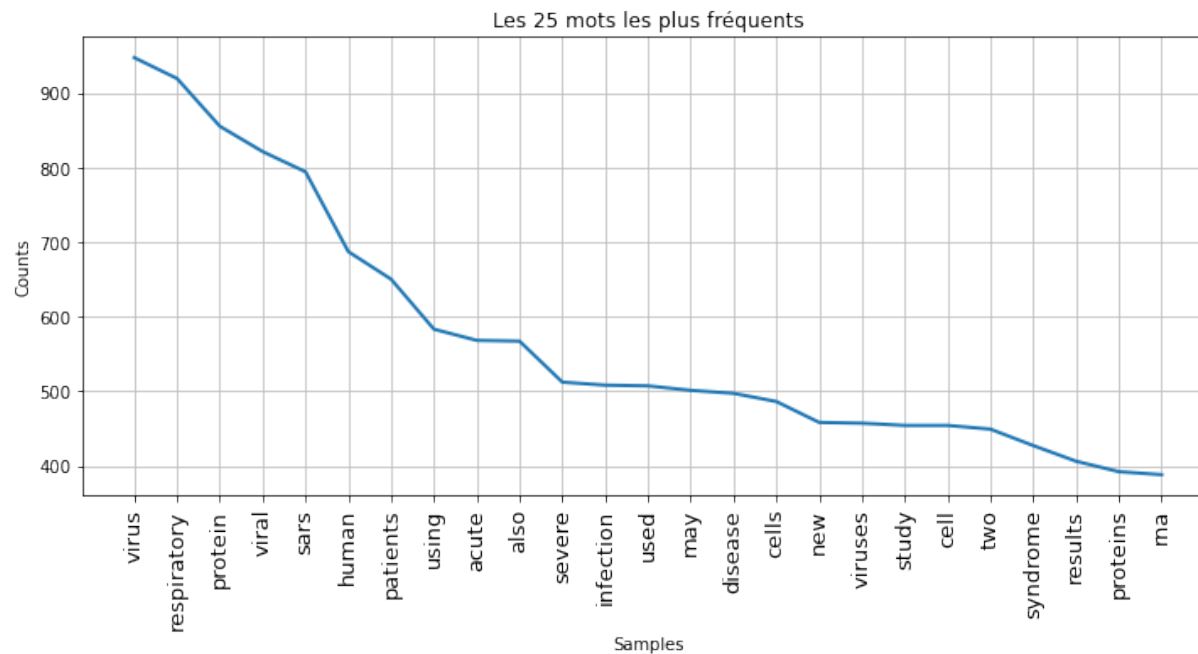
pour langue Anglais : [('of', 1543), ('and', 973), ('in', 852), ('the', 774), ('a', 451), ('for', 338), ('respiratory', 281), ('acute', 229), ('to', 217), ('with', 217), ('sars', 204), ('severe', 204), ('coronavirus', 202), ('virus', 198), ('syndrome', 179), ('on', 164), ('by', 162), ('from', 150), ('human', 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), (':', 10), ('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)]

```
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, 6))
        mots = titre.split() titre: Trends in der Impfstoffentwicklung. DNA- und 2
        for mot in mots: mots: ['Für', 'eine', 'Reihe', 'von', 'Infektionskrankhei
        mots_des_titres.append(mot.lower()) mots_des_titres: ['wirksamkeit', '
    print("pour Langue ", langue, " : ", Counter(mots_des_titres).most_common(25))
```



Mise en forme du .csv pour Weka

- Weka utilise normalement le format .ARFF mais peut importer les .CSV
- Il est plus facile de respecter les paramètres CSV de Weka avant l'importation...

```
resumesClasses = data[['Resume', 'Categories']][data['Resume'].notnull()]
resumesClasses = resumesClasses[resumesClasses['Categories'].notnull()]
```

On extrait les
colonnes
Résumés et
Catégories

```
for index, row in resumesClasses.iterrows():
    cat = str(row['Categories'])
    catRetenue = re.search("2 - ([a-z]+)", cat)
    if catRetenue:
        row['Categories'] = catRetenue[1]
    else:
        catRetenue = re.search("1 - ([a-z]+)", cat)
        if catRetenue:
            row['Categories'] = catRetenue[1]
```

On ne conserve qu'une seule catégorie

1 - science ; 2 immunology

On enregistre en .csv « Weka »

```
resumesClasses.to_csv(fichierSortie, sep=',', na_rep='?', index = False, header=True,
quoting=csv.QUOTE_NONNUMERIC, quotechar='"', doublequote=False, escapechar='\\')
```

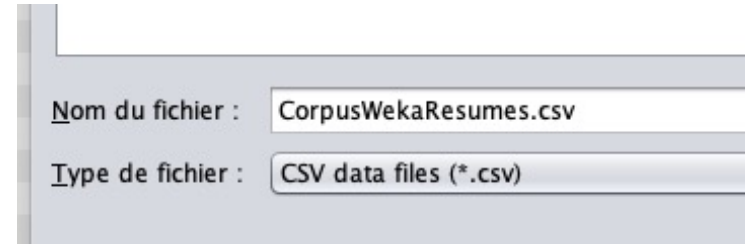
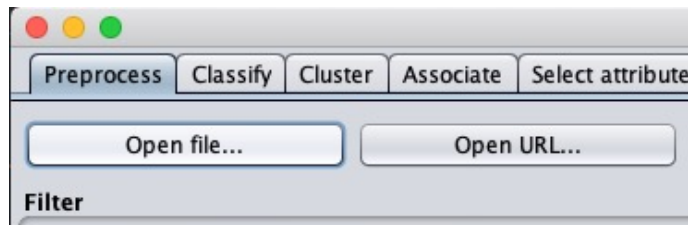

ça ne suffit pas... les types ne sont pas spécifiés

Three attribute types are supported:

- numeric: This type of attribute represents a floating-point number.
- nominal: This type of attribute represents a fixed set of nominal values.
- string: This type of attribute represents a dynamically expanding set of nominal values.

The image shows two screenshots from the Weka data mining software. The top screenshot is the ARFF-Viewer window, displaying a CSV file named 'CorpusWekaResumes.csv'. The window shows a table with two columns: 'No.' and '1: Resume Nominal'. The '1: Resume' column contains text snippets from resumes, and the '2: Categorie' column contains category names like 'cell', 'hematology', 'multidisciplinary', etc. The '1: Resume' column header is circled in red. The bottom screenshot is the Weka GUI Chooser window, showing the 'Tools' menu with options like 'Package manager', 'ArffViewer', 'SqlViewer', and 'Bayes net editor'. The 'ArffViewer' option is highlighted. The 'Applications' section shows buttons for 'Explorer', 'Experimenter', 'KnowledgeFlow', 'Workbench', and 'Simple CLI'. The Weka logo and version information (3.8.5) are also visible.

Ouverture du .csv dans Weka « Explorer »



Current relation

Relation: CorpusWekaResumes
Instances: 1276
Attributes: 2
Sum of weights: 1276

Attributes

All None Invert Pattern

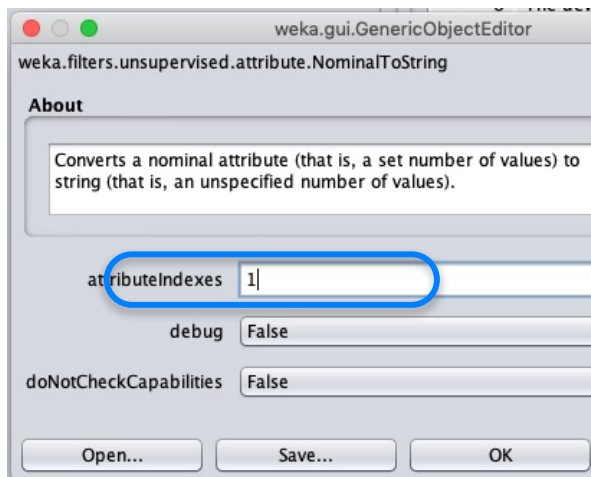
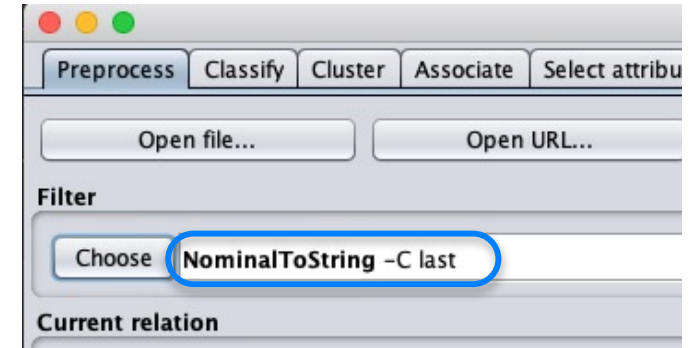
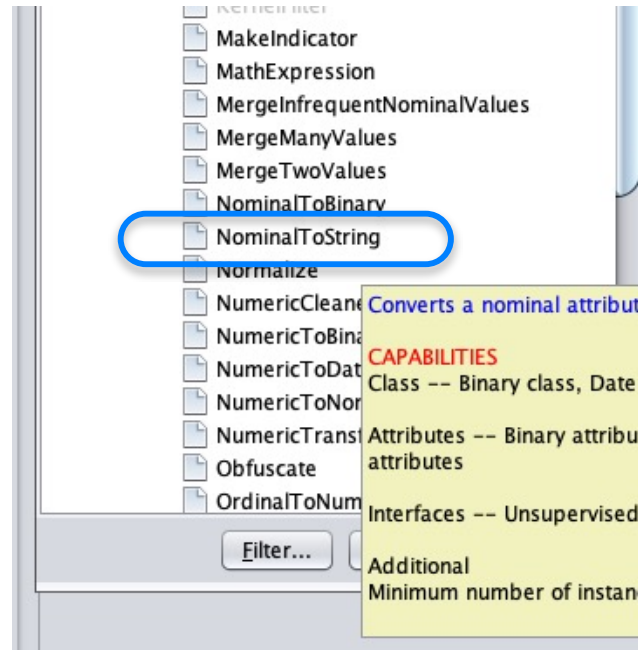
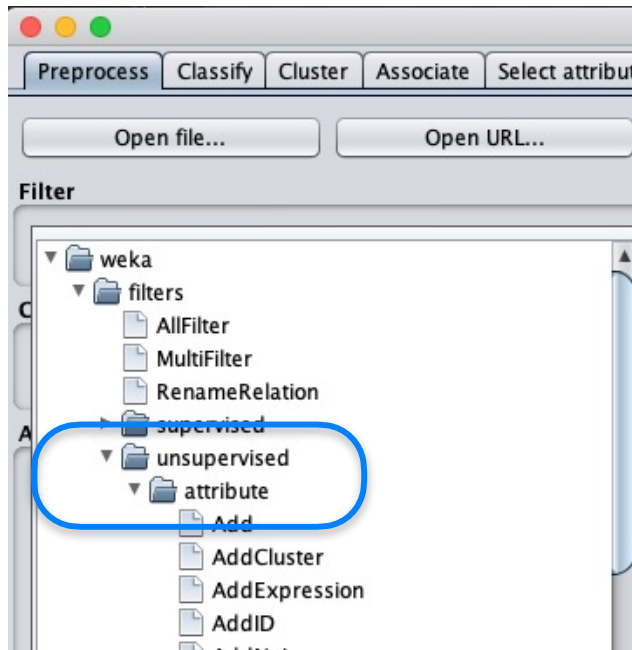
No.	Name
1	☒ Resume
2	☐ Categories

Selected attribute

Name: Resume
Missing: 0 (0%)
Distinct: 1275
Type: Nominal
Unique: 1274 (100%)

No.	Label	Count	Weight
1	Structural biology is maki...	1	1.0
2	The threat of infection by ...	1	1.0
3	Not science fiction, but a ...	1	1.0
4	Severe acute respiratory ...	1	1.0
5	Objective: To understand...	1	1.0
6	For clinical diagnosis, a s...	1	1.0
7	Bacterial storage lipids in...	1	1.0
8	The development of glyca...	1	1.0
9	Genetic polymorphisms h...	1	1.0
10	Background. A unique ge...	1	1.0
11	Infection with the SARS (S...	1	1.0

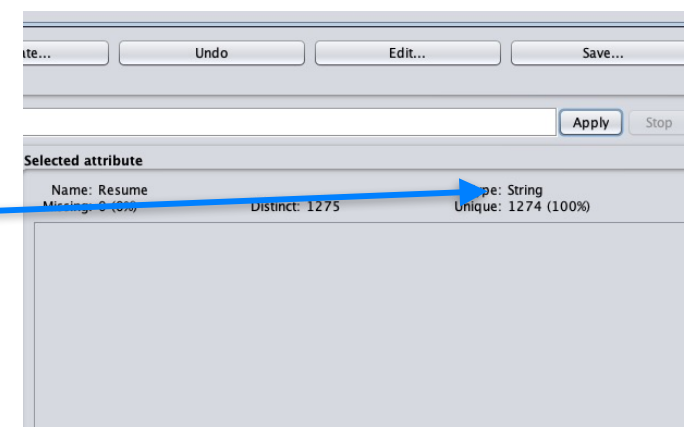
Conversion de l'attribut Résumé en « string »



Weka GUI - Selected attribute details.

Name: Resume
Missing: 0 (0%)
Distinct: 1275
Type: Nominal
Unique: 1274 (100%)

No.	Label	Count	Weight
1	Structural biology is maki...	1	1.0
2	The threat of infection by ...	1	1.0
3	Not science fiction, but a ...	1	1.0
4	Severe acute respiratory ...	1	1.0
5	Objective: To understand...	1	1.0
6	For clinical diagnosis, a s...	1	1.0
7	Bacterial storage lipids in...	1	1.0
8	The development of glyca...	1	1.0
9	Genetic polymorphisms h...	1	1.0



Reste à vectoriser les résumés avec un filtre adapté

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize DI4j Inference

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **NominalToString -C 1** Apply Stop

Current relation

Relation: CorpusWekaResumes-weka.filters.unsupervised.attribute.NominalToString... Attributes: 2
Instances: 1276 Sum of weights: 1276

Attributes

All None Invert Pattern

No.	Name
1	Resume
2	Categories

Remove

Selected attribute

Name: Categories
Missing: 0 (0%)
Distinct: 79
Type: Nominal
Unique: 19 (1%)

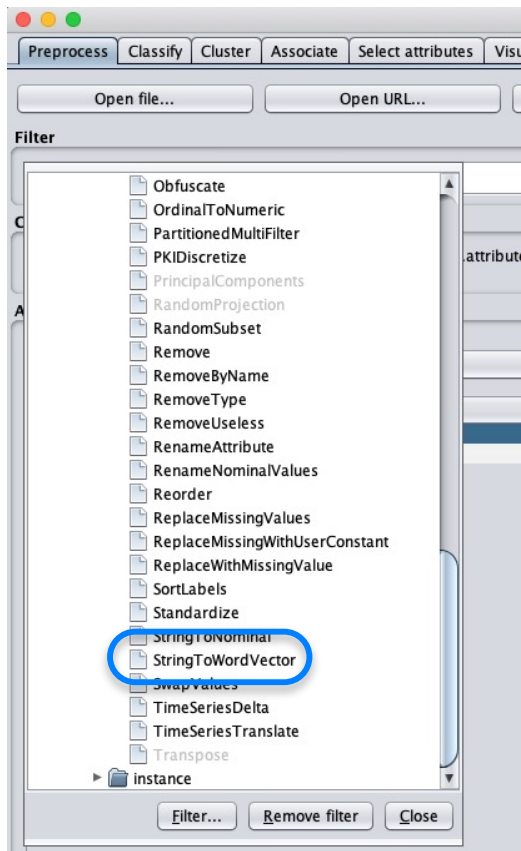
No.	Label	Count	Weight
1	cell	49	49.0
2	hematology	29	29.0
3	multidisciplinary	42	42.0
4	chemistry	117	117.0
5	public	65	65.0
6	polymer	1	1.0
7	microbiology	135	135.0
8	medical	5	5.0
9	biophysics	40	40.0
10	crystallography	32	32.0
11	biochemistry	104	104.0
12	evolutionary	23	23.0
13	medicine	38	38.0
14	psychology	3	3.0
15	emergency	4	4.0
16	environmental	5	5.0

Class: Categories (Nom) Visualize All

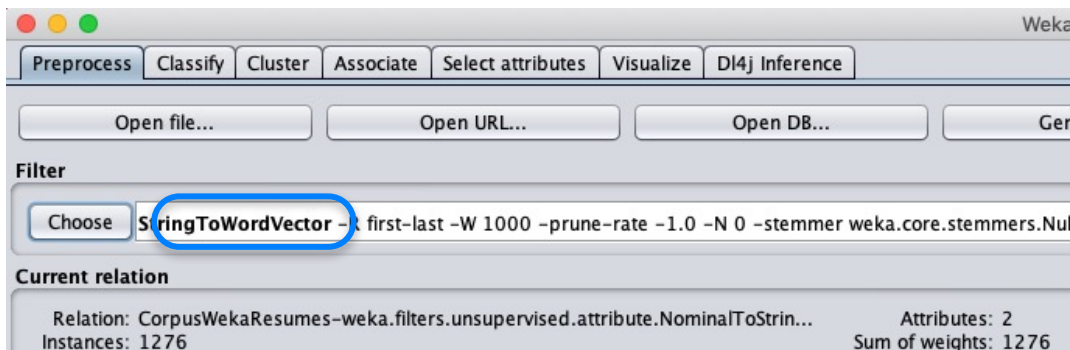
Status

OK Log x 0

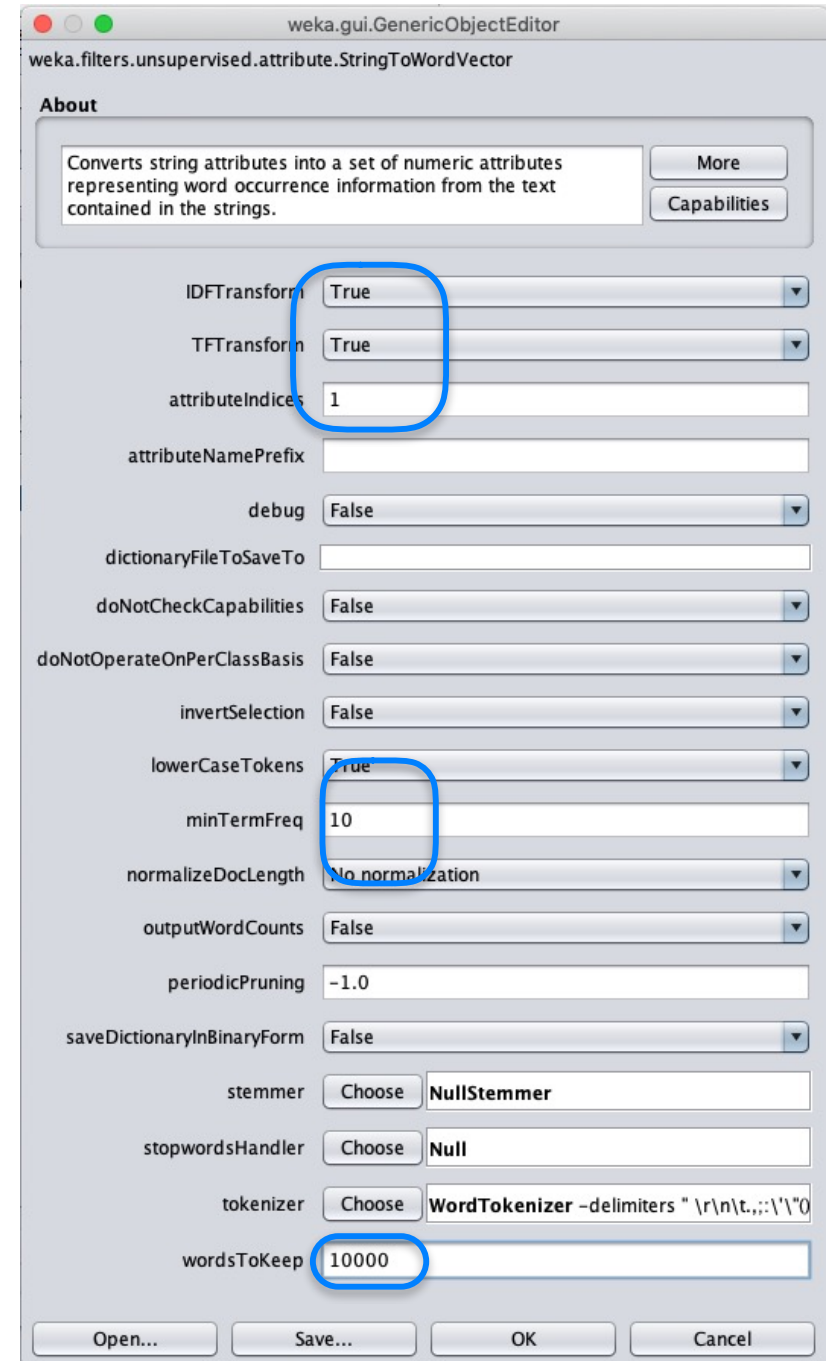
1



2



3



On applique StringToWordVector : il y a autant d'attributs que de mots

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize DL4j Inference

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter

Choose **StringToWordVector** -R 1 -W 10000 -prune-rate -1.0 -T -I -N 0 -L -stemmer weka.core.stemmers.NullStemmer -stopwords-handler weka.core.stopwords.Null -M 10 -tokenizer "weka.core.tokenizers.Null" **Apply** Stop

Current relation

Relation: CorpusWekaResumes-weka.filters.unsupervised.attribute.NominalToString...
Instances: 1276
Attributes: 951
Sum of weights: 1276

Attributes

All None Invert Pattern

No.	Name
1	☒ Categories
2	☐ 1
3	☐ 10
4	☐ 1002/wrna
5	☐ a
6	☐ ace2
7	☐ activity
8	☐ also
9	☐ an
10	☐ analysis
11	☐ and
12	☐ antiviral
13	☐ are
14	☐ article
15	☐ as
16	☐ at
17	☐ bag3
18	☐ be
19	☐ been
20	☐ between
21	☐ binding
22	☐ but
23	☐ by
24	☐ can
25	☐ cell
26	☐ cells
27	☐ cellular
28	☐ chloroquine
29	☐ complex
30	☐ development

Remove

Selected attribute

Name: Categories
Missing: 0 (0%)
Distinct: 79
Type: Nominal
Unique: 19 (1%)

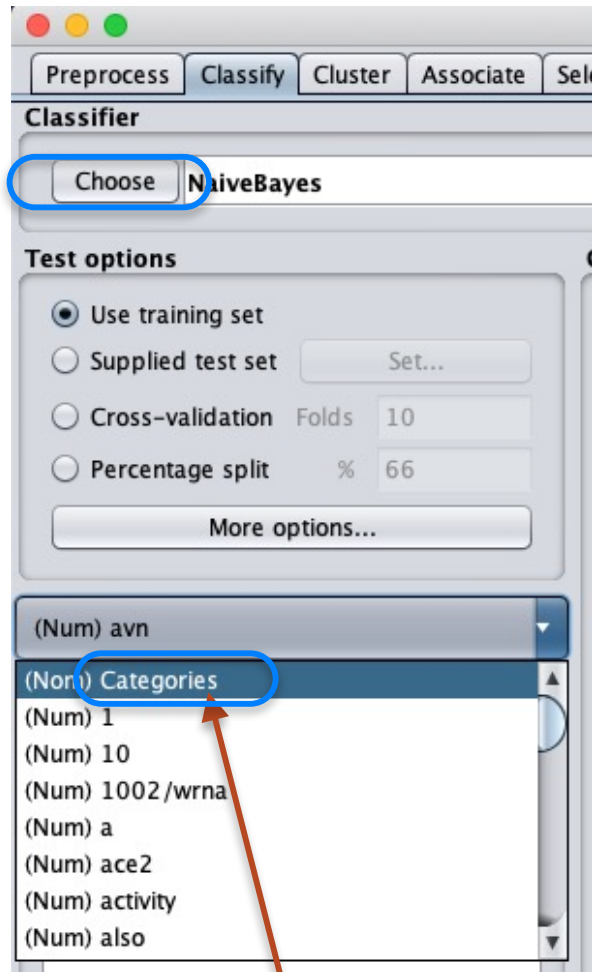
No.	Label	Count	Weight
1	cell	49	49.0
2	hematology	29	29.0
3	multidisciplinary	42	42.0
4	chemistry	117	117.0
5	public	65	65.0
6	polymer	1	1.0
7	microbiology	135	135.0
8	medical	5	5.0
9	biophysics	40	40.0
10	crystallography	32	32.0
11	biochemistry	104	104.0
12	evolutionary	23	23.0
13	medicine	38	38.0
14	psychology	3	3.0
15	emergency	4	4.0
16	environmental	5	5.0

Class: Categories (Nom) Visualize All

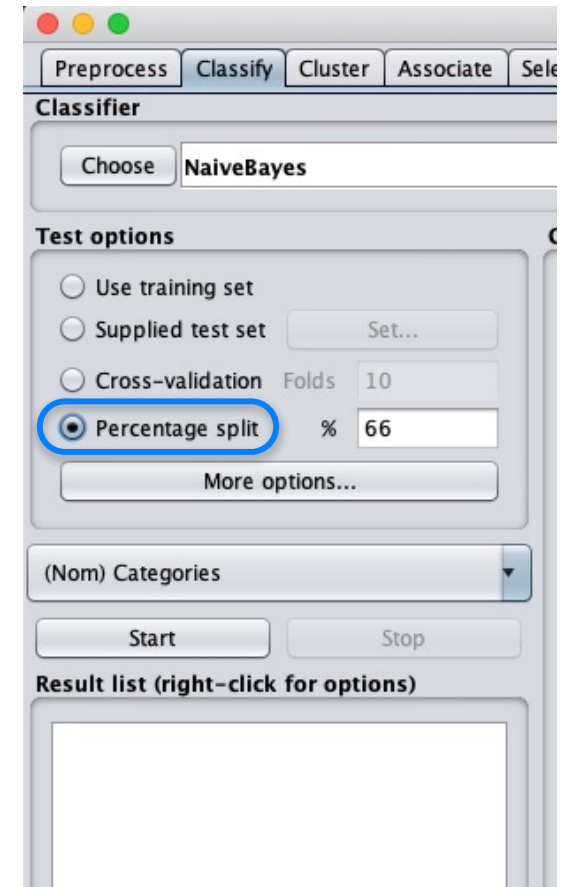
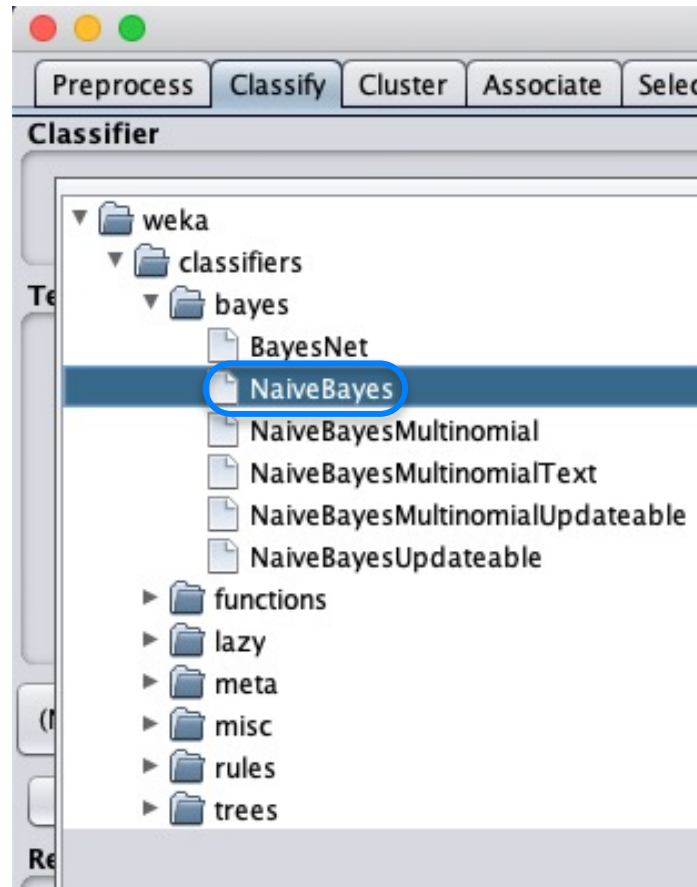
Status

OK Log x 0

Utilisation d'un classifieur bayésien pour apprendre à prédire les catégories



Attention à bien choisir l'attribut à prédire



$$X = (x_1, x_2, \dots, x_n) \longrightarrow \text{classe}$$

$$X = (x_1, x_2, \dots, x_n) \longrightarrow \text{classe}$$

$$X = (x_1, x_2, \dots, x_n) \longrightarrow \text{classe}$$

$$X = (x_1, x_2, \dots, x_n) \longrightarrow \text{classe}$$

Données
d'apprentissage

Max ? $P(\text{classe}|X) = \frac{P(X|\text{classe})P(\text{classe})}{P(X)}$

La probabilité de chaque classe candidate
Autant de scores que de classes

connaissance
a priori

inutile pour comparer les $P(\text{classe})$

Classifieur
bayésien
(règle de Bayes)

avec :

les x sont les descripteurs (features) de l'individu à classer

$$P(X|\text{classe}) = \prod_i P(x_1|\text{classe}) \times \dots \times P(x_i|\text{classe}) \times \dots \times P(x_n|\text{classe})$$

le modèle
appris

{ la fréquence avec laquelle on observe x_1 dans la classe parmi les exemples (données d'apprentissage) la fréquence avec laquelle on observe x_n dans la classe parmi les exemples (données d'apprentissage)

Classifier output

Time taken to build model: 0.36 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 6.46 seconds

=== Summary ===

Correctly Classified Instances	209	48.1567 %
Incorrectly Classified Instances	225	51.8433 %
Kappa statistic	0.4467	
Mean absolute error	0.013	
Root mean squared error	0.111	
Relative absolute error	54.1205 %	
Root relative squared error	101.4383 %	
Total Number of Instances	434	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,357	0,031	0,278	0,357	0,313	0,289	0,883	0,307	cell
0,667	0,002	0,857	0,667	0,750	0,751	0,923	0,746	hematology
0,300	0,042	0,143	0,300	0,194	0,180	0,815	0,172	multidisciplinary
0,476	0,043	0,541	0,476	0,506	0,458	0,886	0,620	chemistry
0,633	0,037	0,559	0,633	0,594	0,563	0,912	0,578	public
0,000	0,007	0,000	0,000	0,000	-0,004	0,894	0,021	polymer
0,585	0,055	0,596	0,585	0,590	0,534	0,889	0,625	microbiology
0,000	0,000	?	0,000	?	?	0,365	0,005	medical
0,706	0,010	0,750	0,706	0,727	0,717	0,935	0,780	biophysics
0,500	0,000	1,000	0,500	0,667	0,702	0,985	0,892	crystallography
0,750	0,068	0,500	0,750	0,600	0,570	0,950	0,753	biochemistry
0,600	0,021	0,250	0,600	0,353	0,377	0,983	0,684	evolutionary
0,263	0,000	1,000	0,263	0,417	0,505	0,688	0,327	medicine
0,000	0,000	?	0,000	?	?	0,866	0,092	psychology
0,000	0,000	?	0,000	?	?	0,891	0,031	emergency
0,000	0,000	?	0,000	?	?	0,995	0,333	environmental

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize DL4j Inference

Classifier

Choose NaiveBayes

Test options

☒ Use training set
☐ Supplied test set Set...
☐ Cross-validation Folds 10
☐ Percentage split % 66
More options...

(Nom) Categories

Start Stop

Result list (right-click for options)

19:12:44 - bayes.NaiveBayes
19:15:35 - bayes.NaiveBayes

Classifier output

Time taken to build model: 0.38 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 19.13 seconds

=== Summary ===

Correctly Classified Instances	1079	84.5611 %
Incorrectly Classified Instances	197	15.4389 %
Kappa Statistic	0.8374	
Mean absolute error	0.0039	
Root mean squared error	0.0597	
Relative absolute error	16.1586 %	
Root relative squared error	54.4993 %	
Total Number of Instances	1276	

à comparer avec 48% avec 2/3 — 1/3 (test)

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,735	0,012	0,706	0,735	0,720	0,709	0,982	0,803	cell
0,966	0,000	1,000	0,966	0,982	0,982	1,000	0,989	hematology
0,714	0,024	0,508	0,714	0,594	0,587	0,979	0,702	multidisciplinary
0,735	0,010	0,878	0,735	0,800	0,785	0,978	0,881	chemistry
0,892	0,010	0,829	0,892	0,859	0,852	0,989	0,920	public
1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	polymer
0,807	0,027	0,779	0,807	0,793	0,768	0,974	0,866	microbiology
0,800	0,000	1,000	0,800	0,889	0,894	1,000	1,000	medical
0,900	0,004	0,878	0,900	0,889	0,885	0,998	0,969	biophysics
0,906	0,000	1,000	0,906	0,951	0,951	1,000	0,998	crystallography
0,865	0,023	0,769	0,865	0,814	0,799	0,991	0,895	biochemistry
1,000	0,002	0,920	1,000	0,958	0,958	1,000	0,994	evolutionary
0,605	0,001	0,958	0,605	0,742	0,756	0,976	0,830	medicine

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize DI4j Inference

Classifier

Choose RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) Categories

Start Stop

Result list (right-click for options)

19:20:59 - trees.RandomTree

Classifier output

=== Run information ===

Scheme: weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Relation: CorpusWekaResumes-weka.filters.unsupervised.attribute.NominalToString-C1-weka.filters.unsupervised.attribute.StringToWordVector-R1-W10000-p

Instances: 1276

Attributes: 951

[list of attributes omitted]

Test mode: split 66.0% train, remainder test

=== Classifier model (full training set) ===

RandomTree

```

coronavirus < 0.57
|
| disease < 0.57
| |
| | can < 0.61
| | |
| | | high < 0.92
| | | |
| | | | study < 0.54
| | | | |
| | | | | we < 0.36
| | | | | |
| | | | | | its < 0.82
| | | | | | |
| | | | | | | proteins < 0.73
| | | | | | | |
| | | | | | | | most < 0.82
| | | | | | | | |
| | | | | | | | | a < 0.05
| | | | | | | | | |
| | | | | | | | | | parameters < 1.56
| | | | | | | | | | |
| | | | | | | | | | | an < 0.33
| | | | | | | | | | | |
| | | | | | | | | | | | viral < 0.5
| | | | | | | | | | | | |
| | | | | | | | | | | | | two < 0.71
| | | | | | | | | | | | | |
| | | | | | | | | | | | | | en < 1.56
| | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | vitro < 1.39
| | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | symptoms < 1.23
| | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | some < 1.05
| | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | derivatives < 1.46
| | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | virus < 0.41
| | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | also < 0.59
| | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | sont < 1.8
| | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | knowledge < 1.8
| | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | because < 1.62 : chemistry (2/0)
| | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | because >= 1.62 : electrochemistry (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | knowledge >= 1.8 : chemistry (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | sont >= 1.8 : medicine (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | also >= 0.59 : spectroscopy (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | virus >= 0.41 : multidisciplinary (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | derivatives >= 1.46 : chemistry (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | some >= 1.05
| | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | diseases < 0.93 : chemistry (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | diseases >= 0.93 : multidisciplinary (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | symptoms >= 1.23 : pathology (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | vitro >= 1.39 : pharmacology (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | en >= 1.56
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | est < 1.62 : water (1/0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | est >= 1.62 : medicine (5/0)

```

Et avec un arbre de décision ?

Status

OK Log x 0

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize DI4j Inference

Classifier

Choose RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) Categories

Start Stop

Result list (right-click for options)

19:20:59 - trees.RandomTree

Classifier output

=== Summary ===

Correctly Classified Instances	65	14.977 %
Incorrectly Classified Instances	369	85.023 %
Kappa statistic	0.0992	
Mean absolute error	0.0215	
Root mean squared error	0.1467	
Relative absolute error	89.3359 %	
Root relative squared error	134.0334 %	
Total Number of Instances	434	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,143	0,040	0,105	0,143	0,121	0,088	0,551	0,043	cell
0,333	0,009	0,429	0,333	0,375	0,366	0,662	0,157	hematology
0,200	0,038	0,111	0,200	0,143	0,122	0,581	0,041	multidisciplinary
0,190	0,064	0,242	0,190	0,213	0,141	0,563	0,125	chemistry
0,100	0,032	0,188	0,100	0,130	0,091	0,534	0,081	public
0,000	0,000	?	0,000	?	?	0,500	0,002	polymer
0,283	0,131	0,231	0,283	0,254	0,139	0,576	0,153	microbiology
0,000	0,014	0,000	0,000	0,000	-0,008	0,493	0,005	medical
0,118	0,017	0,222	0,118	0,154	0,137	0,550	0,061	biophysics
0,000	0,002	0,000	0,000	0,000	-0,008	0,499	0,028	crystallography
0,222	0,108	0,157	0,222	0,184	0,098	0,557	0,099	biochemistry
0,000	0,014	0,000	0,000	0,000	-0,013	0,493	0,012	evolutionary
0,158	0,002	0,750	0,158	0,261	0,333	0,578	0,155	medicine
0,000	0,000	?	0,000	?	?	0,500	0,005	psychology
0,000	0,000	?	0,000	?	?	0,500	0,005	emergency
0,000	0,000	?	0,000	?	?	0,500	0,002	environmental
0,125	0,035	0,063	0,125	0,083	0,064	0,545	0,024	immunology
0,000	0,014	0,000	0,000	0,000	-0,025	0,493	0,041	pathology
?	0,000	?	?	?	?	?	?	history
0,000	0,007	0,000	0,000	0,000	-0,006	0,497	0,005	nanoscience
0,000	0,002	0,000	0,000	0,000	-0,003	0,499	0,005	physics
0,087	0,056	0,080	0,087	0,083	0,030	0,515	0,055	pharmacology
0,000	0,002	0,000	0,000	0,000	-0,002	0,499	0,002	pediatrics
0,000	0,009	0,000	0,000	0,000	-0,009	0,495	0,009	food
0,000	0,000	?	0,000	?	?	0,500	0,002	ophthalmology
?	0,000	?	?	?	?	?	?	engineering
0,125	0,014	0,143	0,125	0,133	0,118	0,555	0,034	biology
?	0,007	0,000	?	?	?	?	?	endocrinology
?	0,002	0,000	?	?	?	?	?	otorhinolaryngology
0,000	0,000	?	0,000	?	?	0,500	0,002	parasitology
0,000	0,026	0,000	0,000	0,000	-0,026	0,487	0,025	veterinary
0,000	0,005	0,000	0,000	0,000	-0,005	0,498	0,005	genetics
0,100	0,042	0,053	0,100	0,069	0,042	0,529	0,026	respiratory
0,000	0,012	0,000	0,000	0,000	-0,010	0,494	0,009	mathematical
0,333	0,053	0,382	0,333	0,356	0,298	0,640	0,187	virology
0,000	0,007	0,000	0,000	0,000	-0,011	0,496	0,016	biotechnology
0,000	0,000	?	0,000	?	?	0,500	0,002	oncology
?	0,000	?	?	?	?	?	?	instruments
0,000	0,009	0,000	0,000	0,000	-0,007	0,495	0,005	surgery

Status

OK Log

The screenshot displays the Orange3 data mining software interface. At the top, there are tabs for 'Preprocess', 'Classify', 'Cluster', 'Associate', 'Select attributes', 'Visualize', and 'Data'. The 'Classify' tab is active, showing a 'Classifier' widget configuration window.

Classifier Configuration:

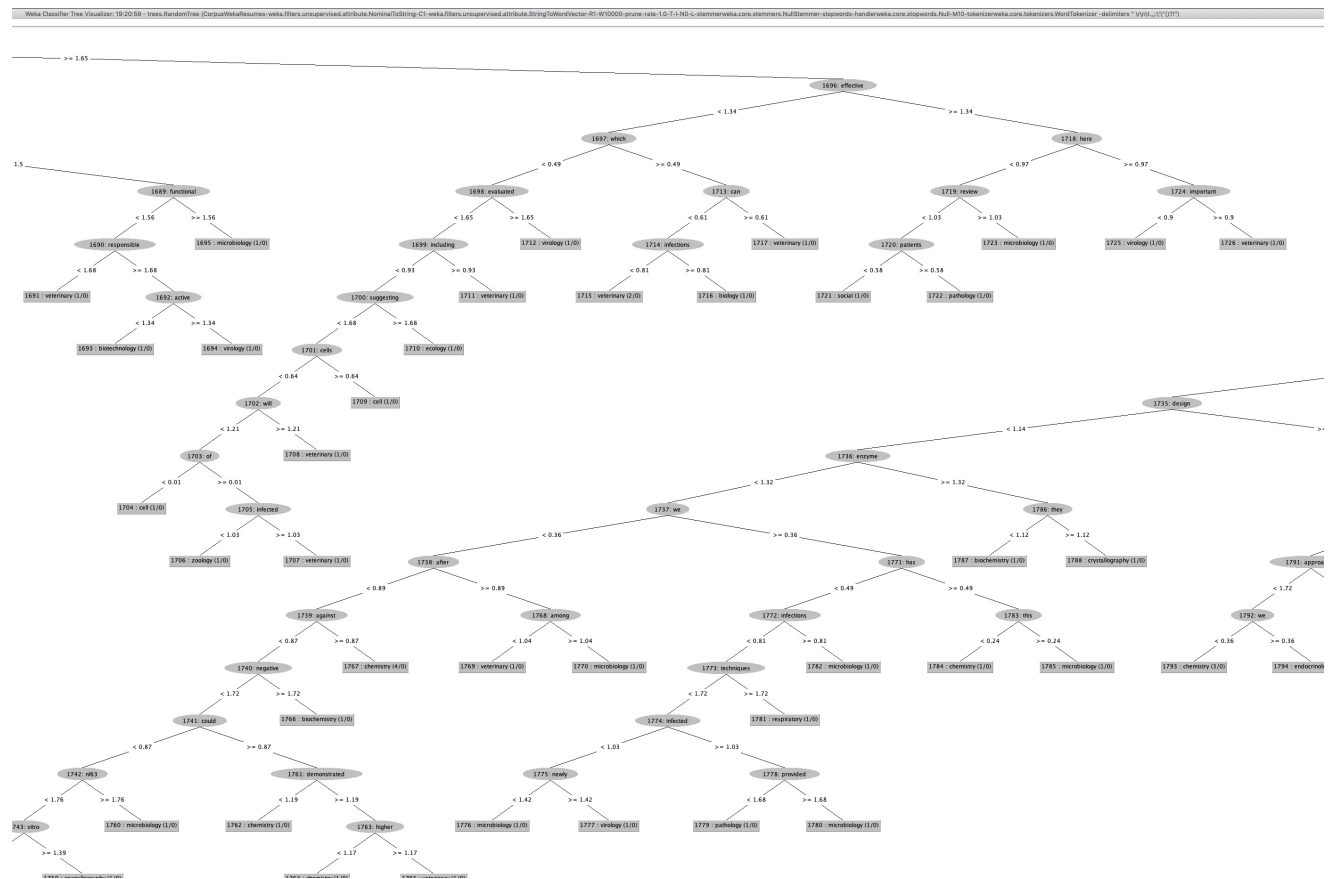
- Classifier:** RandomTree -K 0 -M 1.0 -V 0.001 -S 1
- Test options:**
 - ☐ Use training set
 - ☐ Supplied test set (Set... button)
 - ☐ Cross-validation (Folds: 10)
 - ☒ Percentage split (%: 66)
 - More options... button
- Classifier output:**
 - Summary:
 - Correctly Classified Instances: 100
 - Incorrectly Classified Instances: 0
 - Kappa statistic: 1.0
 - Mean absolute error: 0.0
 - Root mean squared error: 0.0
 - Relative absolute error: 0.0
 - Root relative squared error: 0.0
 - Total Number of Instances: 100
 - Detailed Accuracy By Class:

Class	TP	Rate
0	143	0.143
1	333	0.333
2	200	0.200
3	100	0.100

Result list (right-click for options):

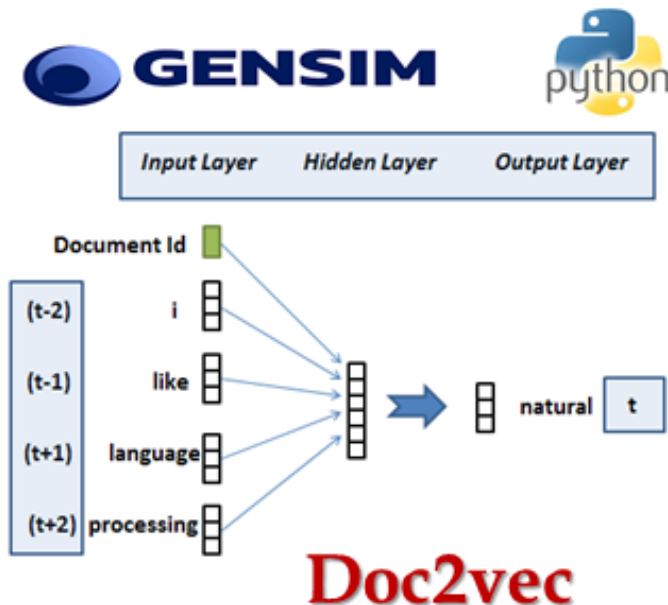
The result list shows a single entry: 19:20:59 - trees.RandomTree. A context menu is open over this entry, listing the following actions:

- View in main window
- View in separate window
- Save result buffer
- Delete result buffer(s)
- Load model
- Save model
- Re-evaluate model on current test set
- Re-apply this model's configuration
- Visualize classifier errors
- Visualize tree** (highlighted)
- Visualize margin curve
- Visualize threshold curve
- Cost/Benefit analysis
- Visualize cost curve

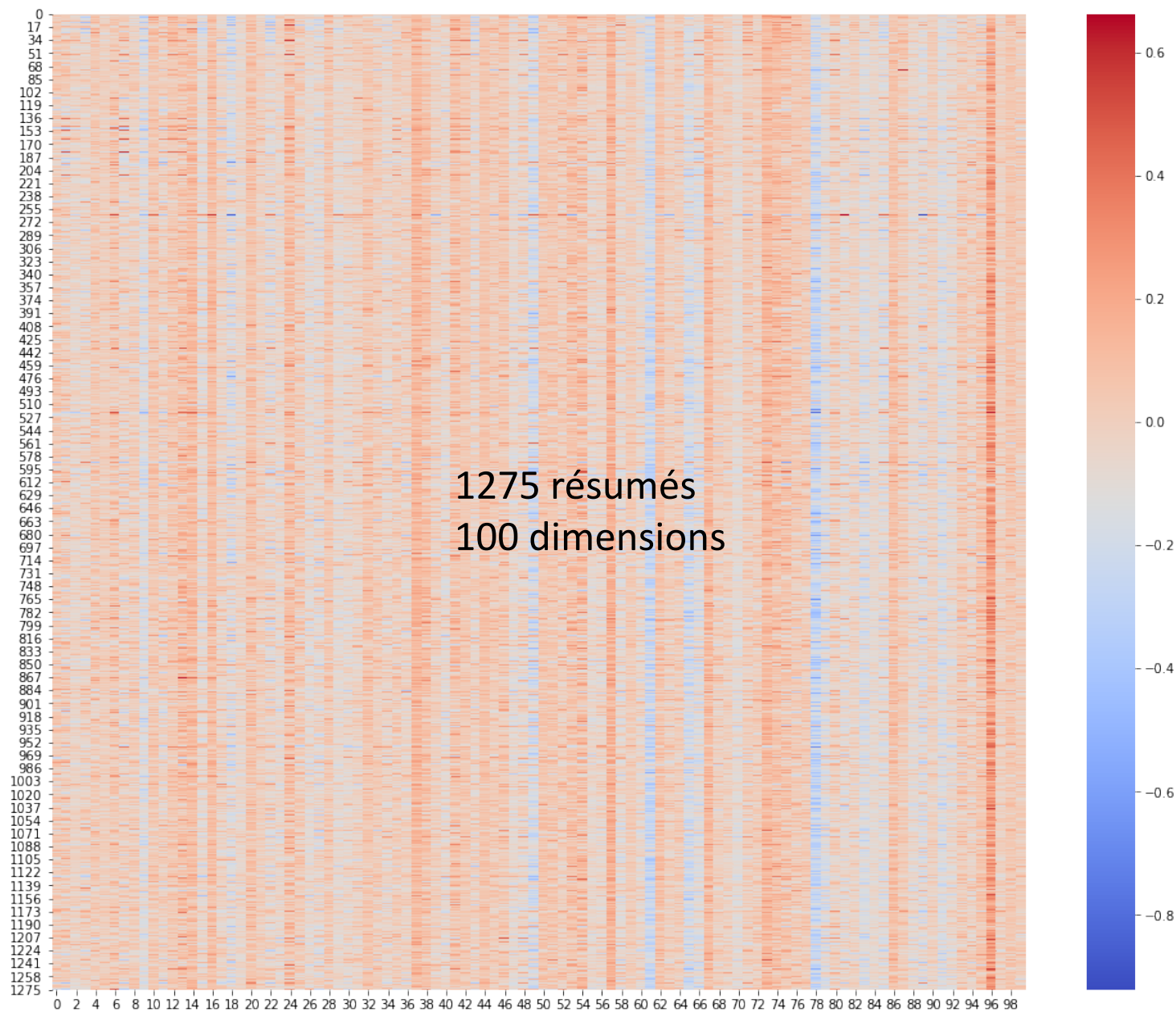


Classification non supervisée (avec Python)

- **Objectif** : réunir les documents en fonction de leurs similarités et visualiser les classes obtenues
- **Moyens** :
 - algorithmes de partitionnement tels que les k-Moyennes ou les cartes auto-organisées
 - visualisation par ACP
- Espace initial en très grande dimension (la taille du vocabulaire) :
 - réunir les mots similaires = projeter les documents sur un espace réduit



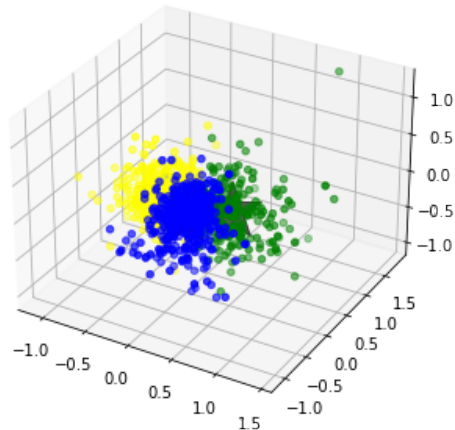
```
def doc2vect():
    document_tagged = []
    tagged_count = 0
    for _ in resumesClasses['Resume'].values:
        document_tagged.append(gensim.models.doc2vec.TaggedDocument(_, [tagged_count]))
        tagged_count += 1
    d2v = Doc2Vec(document_tagged)
    return d2v.docvecs.vectors_docs
```



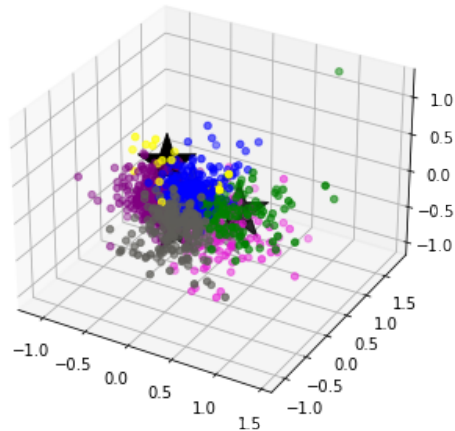
```
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA PCA: <cla

kmean_model = KMeans(n_clusters=15, n_jobs=-1)
%time km = kmean_model.fit_predict(doc2vec)
```

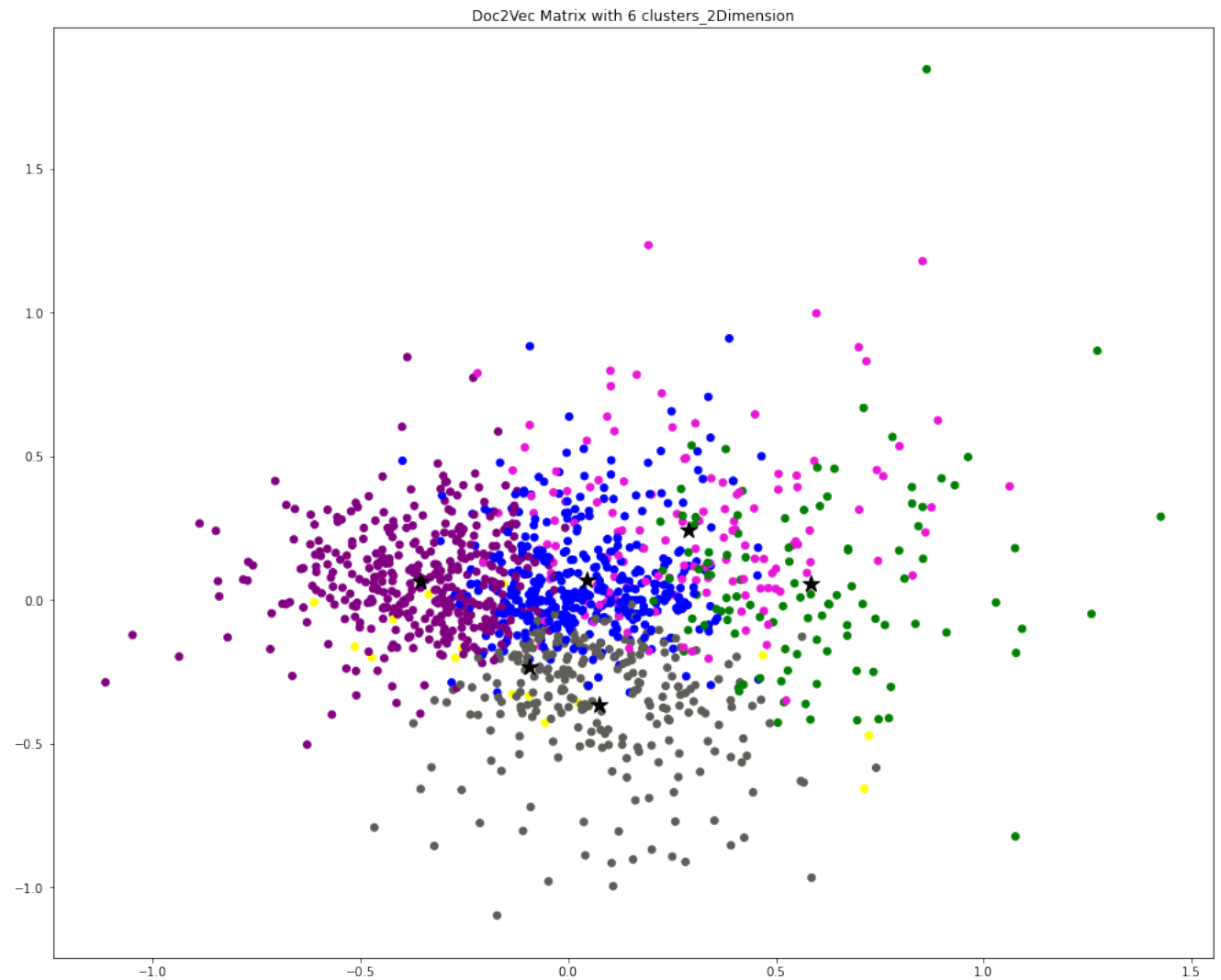
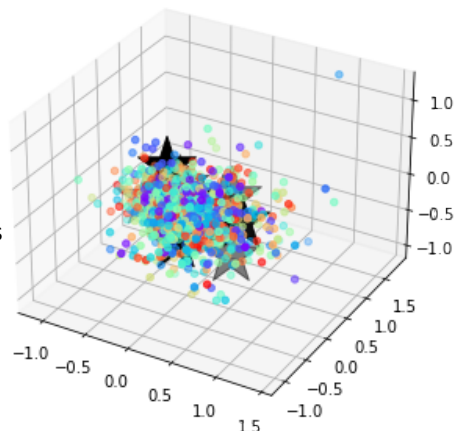
3 classes



6 classes



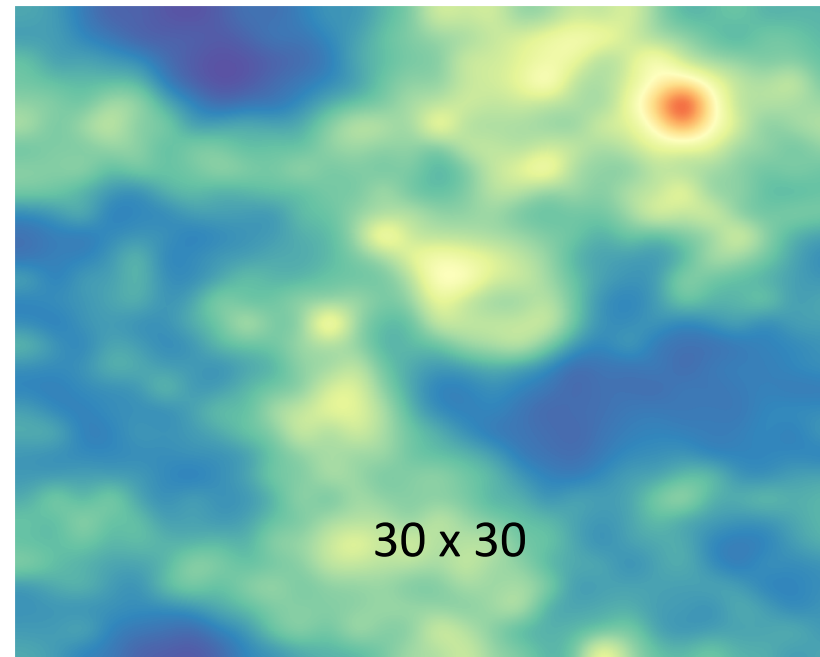
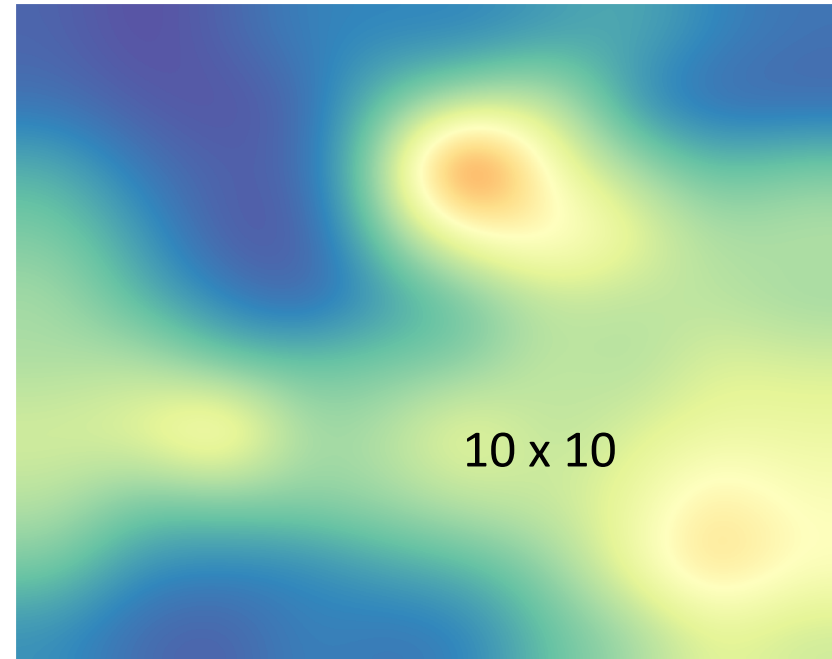
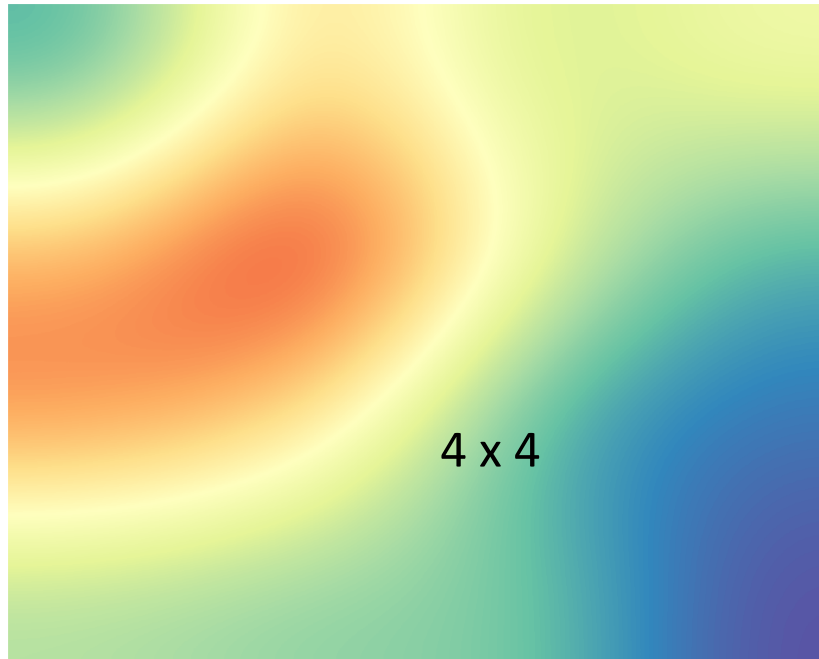
15 classes



k-Means puis ACP pour visualiser les classes


```
som = somoclu.Somoclu(4, 4, maptype="toroid")
```

```
som.train(doc2vec)
```

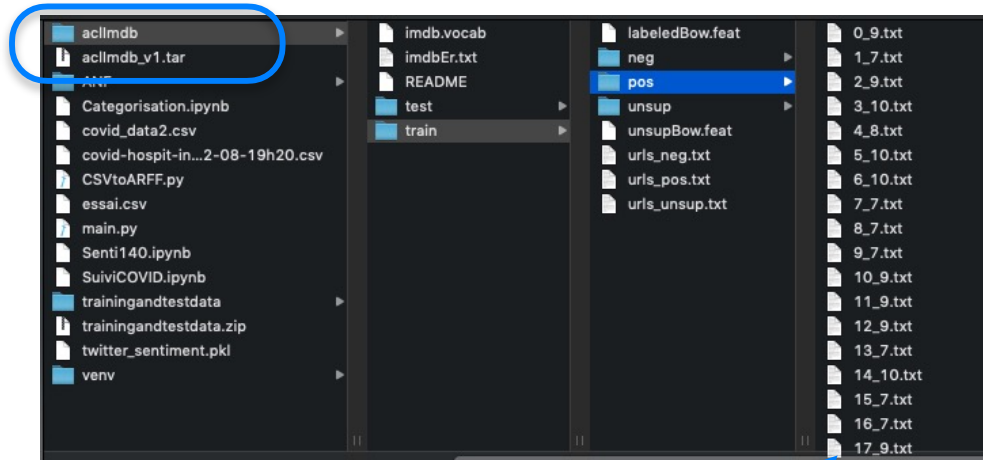


Cartes auto-organisées
(SOM)

ANALYSE DE SENTIMENT (POLARITE) SUR DES CRITIQUES DE FILMS

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 air 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 where 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 for most 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

Le plus simple : utiliser des modules existants

```
from textblob import TextBlob
```

```
print(TextBlob("I hate that movie").sentiment.polarity)
```

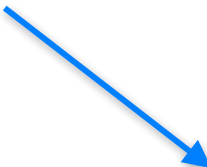
- 0,8



```
texte = "This is a gem. As a Film Four production - the anticipated quality was indeed delivered. Shot with great style that reminded me some Errol Morris films, well arranged and simply gripping. It's long yet horrifying to the point it's excruciating. We know something bad happened (one can guess by the lack of participation of a person in the interviews) but we are compelled to see it, a bit like a car accident in slow motion. The story spans most conceivable aspects and unlike some documentaries did not try and refrain from showing the grimmer sides of the stories, as also dealing with the guilt of the people Don left behind him, wondering why they didn't stop him in time. It took me a few hours to get out of the melancholy that gripped me after seeing this very-well made documentary."
```

```
print(TextBlob(texte).sentiment.polarity)
```

- 0,054



Mais.... comment ? quelle performance en moyenne ? comment l'améliorer ?

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).

```
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()
```

1

	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

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

les tokens les plus fréquents :

```
['the', ',', '.', 'a', 'and', 'of', 'to', 'is', '/', '>', '<', 'br', 'in', 'I', 'it', 'that', "'s", 'this', 'was', 'The', 'as', 'with', 'movie', 'for', 'film', ')', '(', 'but', '""', "n't", '`', 'on', 'you', 'are', 'not', 'have', 'his', 'be', '!', 'he', 'one', 'at', 'by', 'an', 'all', 'who', 'they', 'from', 'like', 'It']
```

100 749 mots n'apparaissent qu'une fois :

```
themeparks
Disney-MGM
artistically-inclined
conscience-less
monsieur
non-lonely
upsetting.
finger-sewing
boondoggling
Bathian
moneygrubbing
smarmy.
Olivier/Garson
cold-fish
highlife
Marchionesse
Udolphi
frazzled.
bunt
Lorelay
obsession
adverterous
Giraud
Schlater
MissCastaway.com
UNBELIEVABLE
Coober
Pedy
Docudrama
'Cobra
'Renegade
Farsape
mega-makeup
puppet/digital
Hynerian
Sebaceans
irreversibly
Crichton.
starburst
Hillarious
unfortunatley
dissapeared
```

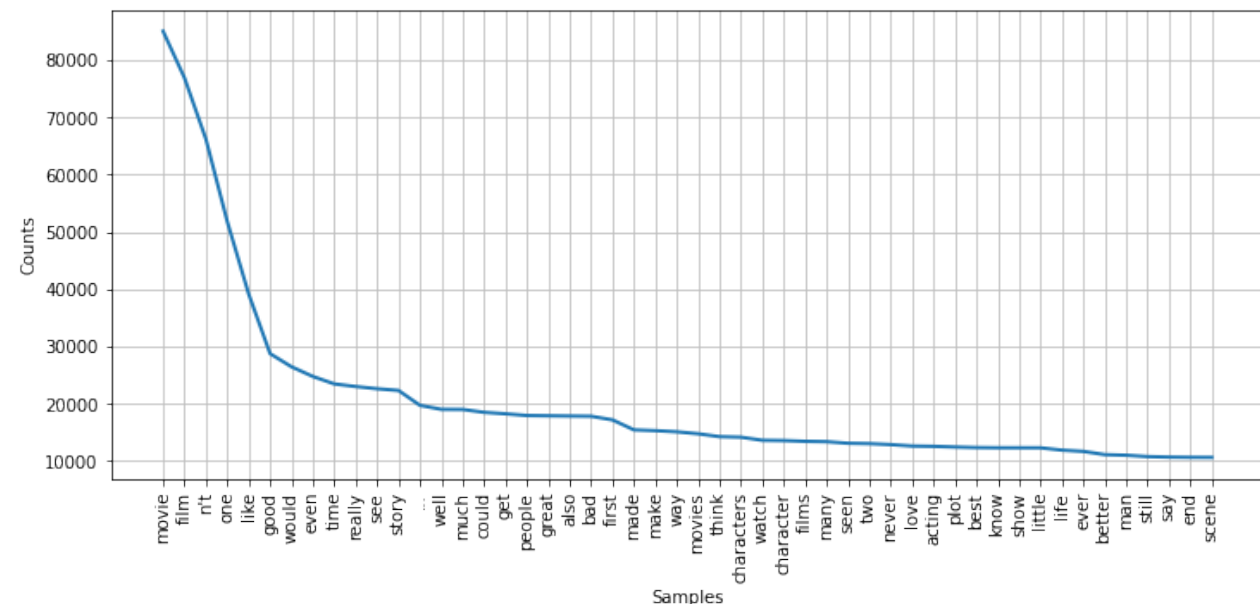
```
hapax = frequency_dist.hapaxes()
```

'the' apparait 573 397 fois

```
{'the': 573397, ',': 544031, '.': 467886, 'and': 309118, 'a': 309103, 'of': 285087, 'to': 263658, 'is': 214740}
```

les tokens les plus fréquents après suppressions des mots outils :

```
from nltk.corpus import stopwords
stop_words = set(stopwords.words('english'))
tokens = [w.lower() for w in tokens if not w.lower() in stop_words and len(w)>2]
```



[illegible]

Avec un classifieur bayésien naïf (NB)

```
###
X_train = df.loc[:24999, 'review'].to_numpy()
# Return a Numpy representation of the DataFrame
# Only the values in the DataFrame will be returned, the axes labels will be removed
y_train = df.loc[:24999, 'polarity'].to_numpy()
X_test = df.loc[25000:, 'review'].to_numpy()
y_test = df.loc[25000:, 'polarity'].to_numpy()
```

Division des exemples :
50 % entraînement (*train*)
50 % test

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(lowercase=False, max_features=10000)
train_vectors = vectorizer.fit_transform(X_train)
test_vectors = vectorizer.transform(X_test)
print(train_vectors.shape, test_vectors.shape)

###
from sklearn.naive_bayes import MultinomialNB
clf = MultinomialNB().fit(train_vectors, y_train)

###
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
predicted = clf.predict(test_vectors)
print("Global Accuracy :", accuracy_score(y_test, predicted))
print(classification_report(y_test, predicted))
```

Evaluation (classes 0 et 1)

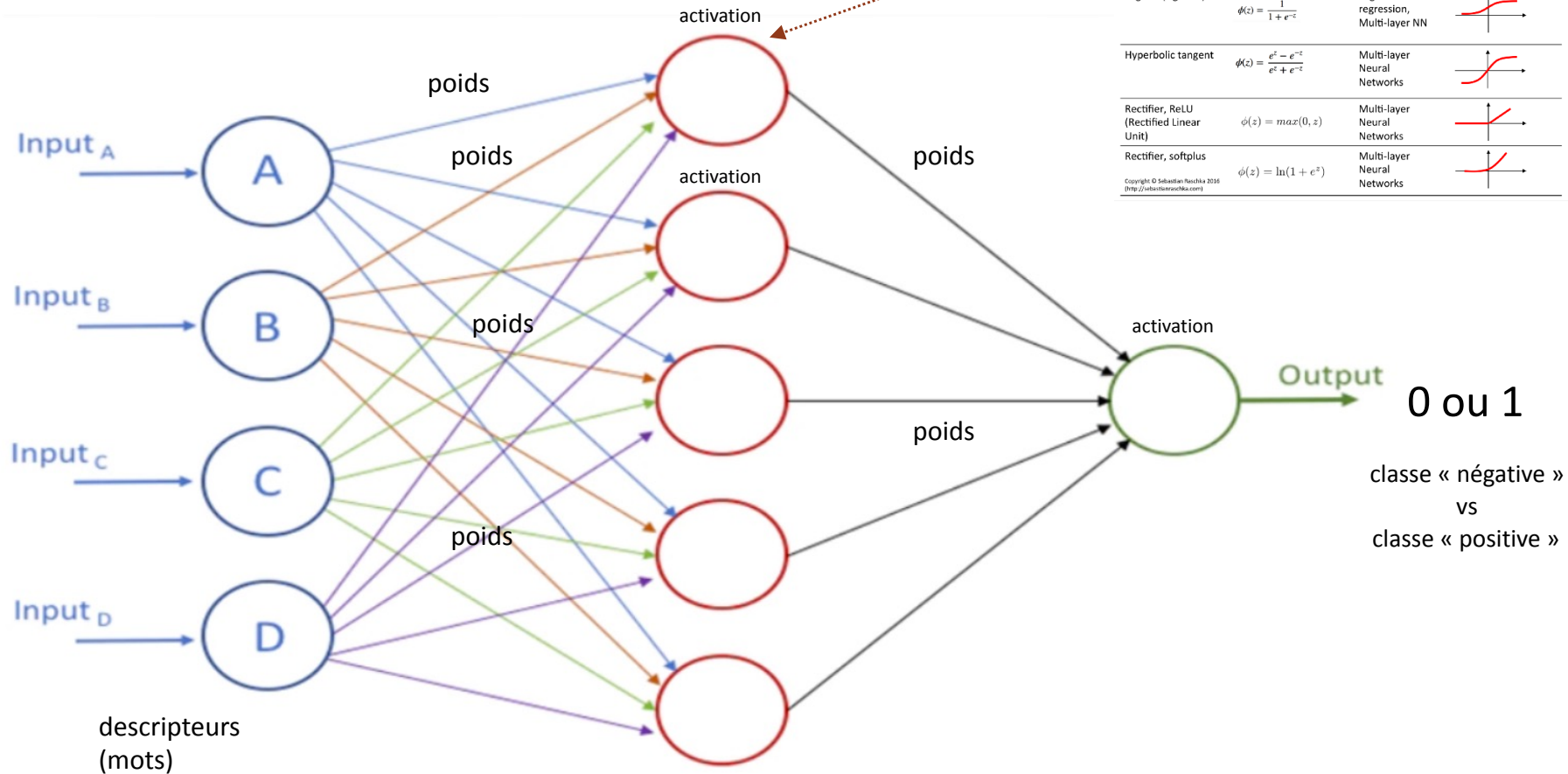
données de test

Global Accuracy : 0.8454					
	precision	recall	f1-score	support	
0	0.84	0.86	0.85	12500	
1	0.85	0.83	0.84	12500	
accuracy			0.85	25000	
macro avg	0.85	0.85	0.85	25000	
weighted avg	0.85	0.85	0.85	25000	

$$accuracy(Y, \hat{Y}) = \frac{1}{n_{exemples}} \sum_i 1(\hat{y}_i = y_i)$$

classes réelles classes prédites

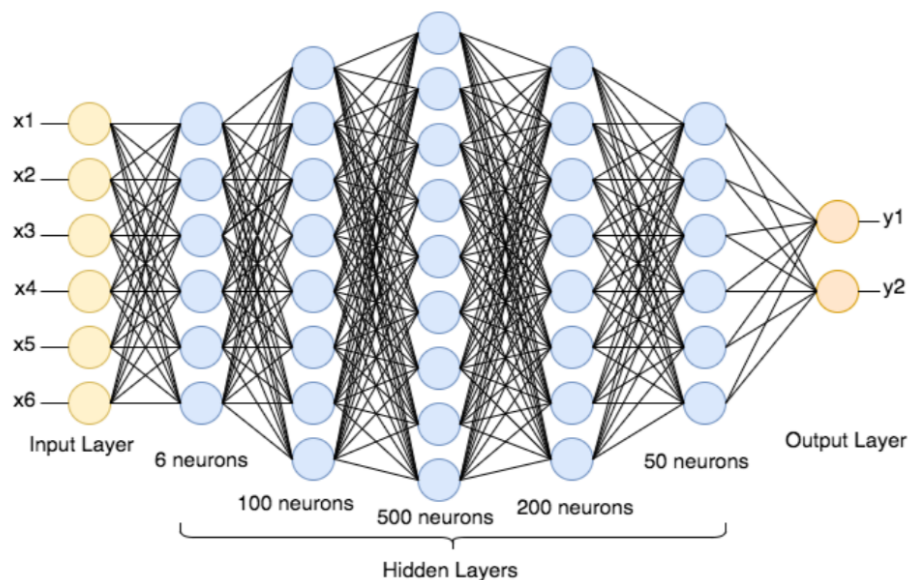
Réseau de neurones



Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	Multi-layer Neural Networks	
Rectifier, ReLU (Rectified Linear Unit)	$\phi(z) = \max(0, z)$	Multi-layer Neural Networks	
Rectifier, softplus	$\phi(z) = \ln(1 + e^z)$	Multi-layer Neural Networks	

Apprentissage de relations (non linéaires) entre les entrées et la sortie

Architectures neuronales

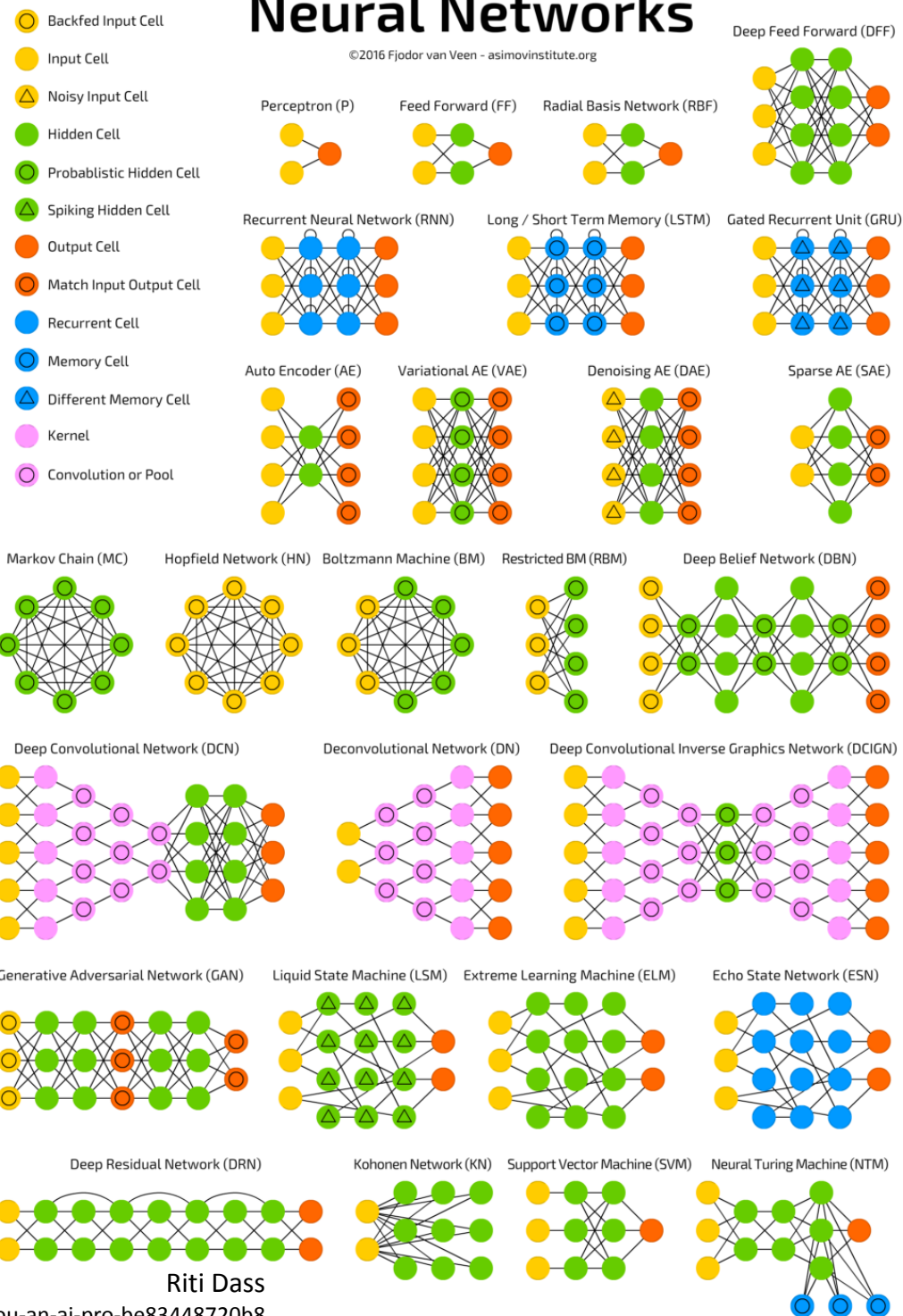


Deep Learning for Ligand-Based Virtual Screening in Drug Discovery
October 2018

DOI: 10.1109/PAIS.2018.8598488

Conference: 2018 3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS)

Meriem BahiMohamed Batouche

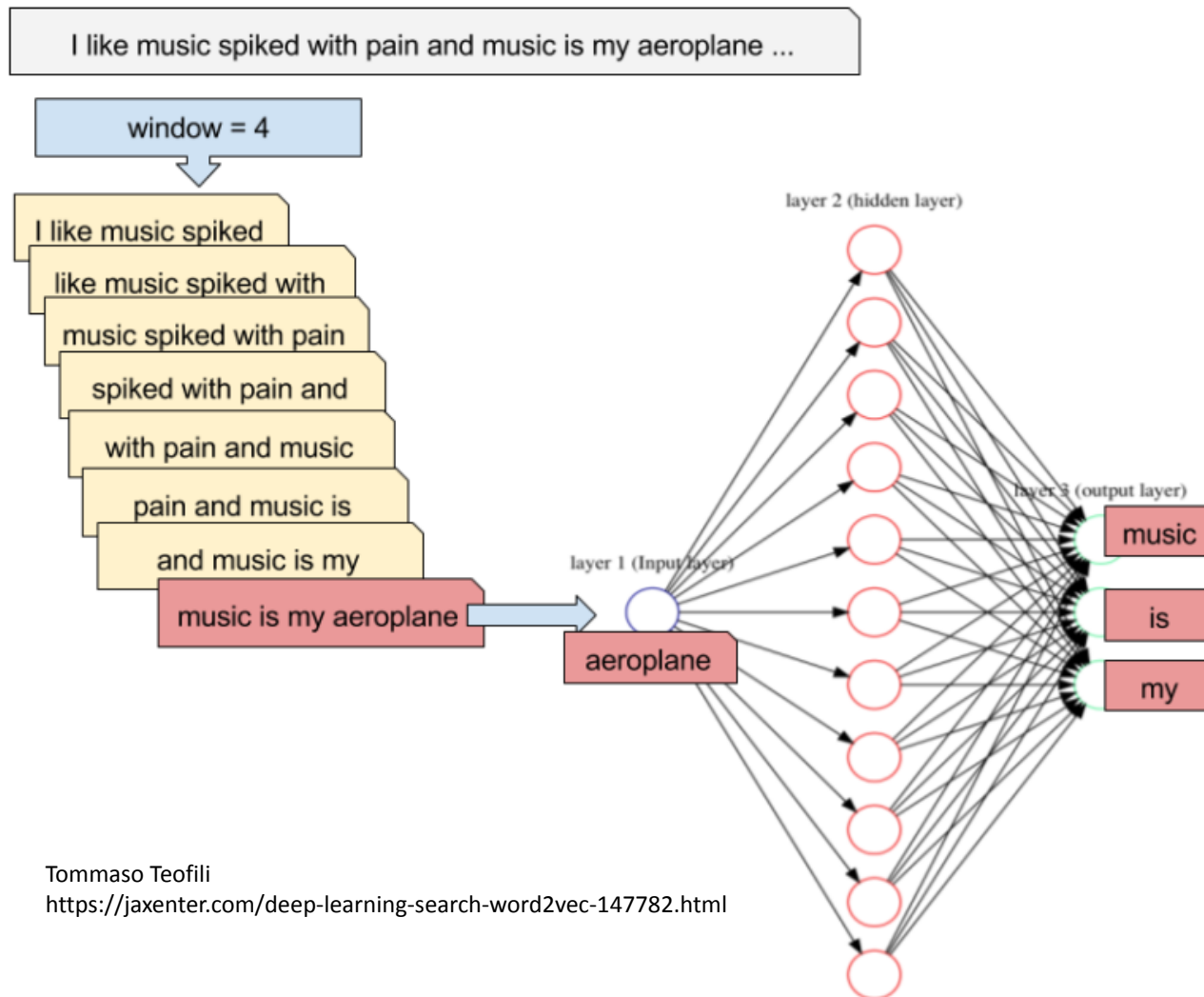


Riti Dass

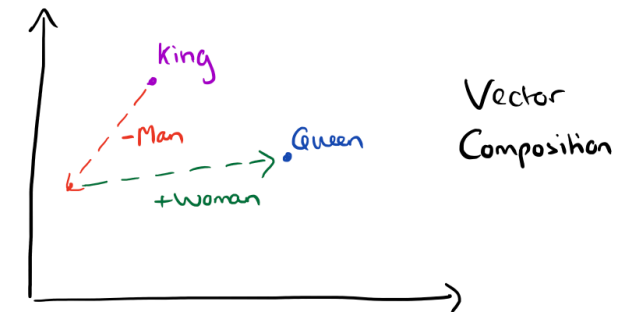
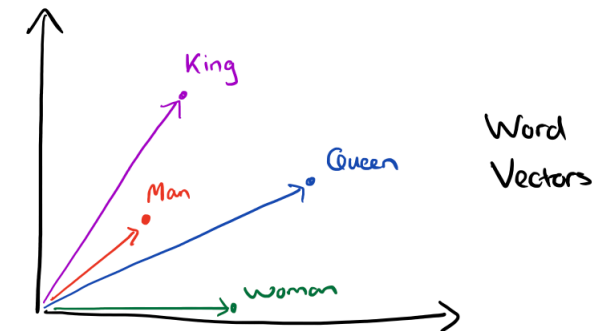
<https://medium.com/predict/the-complete-list-to-make-you-an-ai-pro-be83448720b8>

Réduction de la dimension (projection)

Les plongements de mots (word embeddings)



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-vectors/>

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)
motsComplets = list(model.wv.vocab)
```

pour chaque mot,
un vecteur en
200 dimensions

```
movie -1.3216566 0.36635584 -0.28186616 -1.0511837 -1.0501945 -1.7482823 -0.42692444 0.16830114 -1.073119 -1.5651205 -1.96654 -0.54516864 1.2929311 0.49605948 1.1482662 0.38361785 -0.30000296 0.78807664 -0.62371856 -1.5082116 -0.13787036 -0.74925745 0.41954425 0.35796735 0.3195898 -0.20374134 -0.25748256 -0.90302813 -0.44684523 -0.46419883 0.43331063 0.38016 -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 0.0581 -0.14719109 -0.3533582 -1.16035 1.0383319 0.3641711 -0.29797938 -0.041548226 0.35354558 -0.7025537 0.17913149184 0.21495351 -0.7291604 0.18647747 -1.2000268 -0.51228637 0.36612657 -0.25129464 -0.746 -0.1836736 0.096396 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 0.773262 0.8491248 -0.484025 -0.31997883 0.18664318 -0.5761222 0.33220634 -1.0463667 -0.009183551 0.54716516 0.37895 -1.0772457 -0.646116 1.1264194 -0.9413773 0.08854891 -0.122176886 -0.056594223 0.5072317 1.13529 -0.8807 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 -0.29959 0.46457902 0.110982075 0.07333746 -1.321096 -0.19277126 0.023522813 -0.31523454 -0.23818257 0.4992599 0.20365019 0.2108 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 1.628 -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 -0.01083 21 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 0.29059038 0.37622717 0.11014897 -1.2906862 0.10878839 0.9532716 -0.9014037 -0.41337353 0.57484233 -0.76305604 0.26593128 0.291733
```


Le modèle Word2Vec appris sur les critiques

```
#### Enregistrement du modèle appris
(ici format réduit : gain de place mais ne permet pas de continuer
l'entraînement avec de nouveaux textes -- pour enregistrer un modèle complet,
faire model.save à la place)
#%
```

```
nomEmbeddings = 'imdb_embeddings_word2vec_200_5_100'
model.wv.save_word2vec_format(nomEmbeddings, binary=False)
```

```
print("Taille du vocabulaire : ", len(motsComplet))
print("Les mots les plus proches de horrible sont :")
model.wv.most_similar('horrible')
#%
print("Les mots les plus proches de superb sont :")
model.wv.most_similar('superb')
```

```
Taille du vocabulaire : 96855
Les mots les plus proches de horrible sont :
[('terrible', 0.9239196181297302),
 ('awful', 0.8446447849273682),
 ('horrendous', 0.7841349840164185),
 ('pathetic', 0.7593107223510742),
 ('sucks', 0.7501437664031982),
 ('atrocious', 0.744674563407898),
 ('dreadful', 0.7377941012382507),
 ('horrid', 0.7361111640930176),
 ('lousy', 0.7139973640441895),
 ('ridiculous', 0.7078706622123718)]
```

```
Les mots les plus proches de superb sont :
[('outstanding', 0.8770316243171692),
 ('exceptional', 0.8604843616485596),
 ('excellent', 0.8578838109970093),
 ('terrific', 0.8461805582046509),
 ('fabulous', 0.8225735425949097),
 ('fantastic', 0.817896842956543),
 ('splendid', 0.8140406608581543),
 ('phenomenal', 0.8114627599716187),
 ('marvelous', 0.8058702945709229),
 ('impeccable', 0.788905680179596)]
```

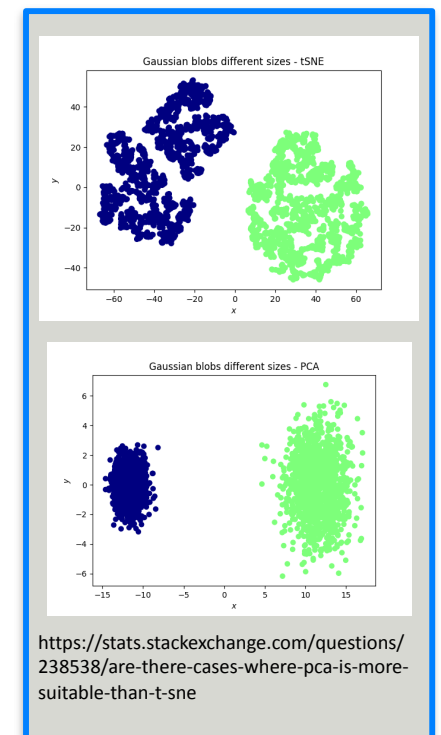
Visualisation des plongements (2 dimensions)

Table of Difference between PCA and t-SNE

allure générale
conservée (variance)

S.NO.	PCA	t-SNE
1.	It is a linear Dimensionality reduction technique.	It is a non-linear Dimensionality reduction technique.
2.	It tries to preserve the global structure of the data.	It tries to preserve the local structure (cluster) of data.
3.	It does not work well as compared to t-SNE.	It is one of the best dimensionality reduction technique.
4.	It does not involve Hyperparameters.	It involves Hyperparameters such as perplexity, learning rate and number of steps.
5.	It gets highly affected by outliers.	It can handle outliers.
6.	PCA is a deterministic algorithm.	It is a non-deterministic or randomised algorithm.
7.	It works by rotating the vectors for preserving variance.	It works by minimising the distance between the point in a gaussian.
8.	We can find decide on how much variance to preserve using eigen values.	We cannot preserve variance instead we can preserve distance using hyperparameters.

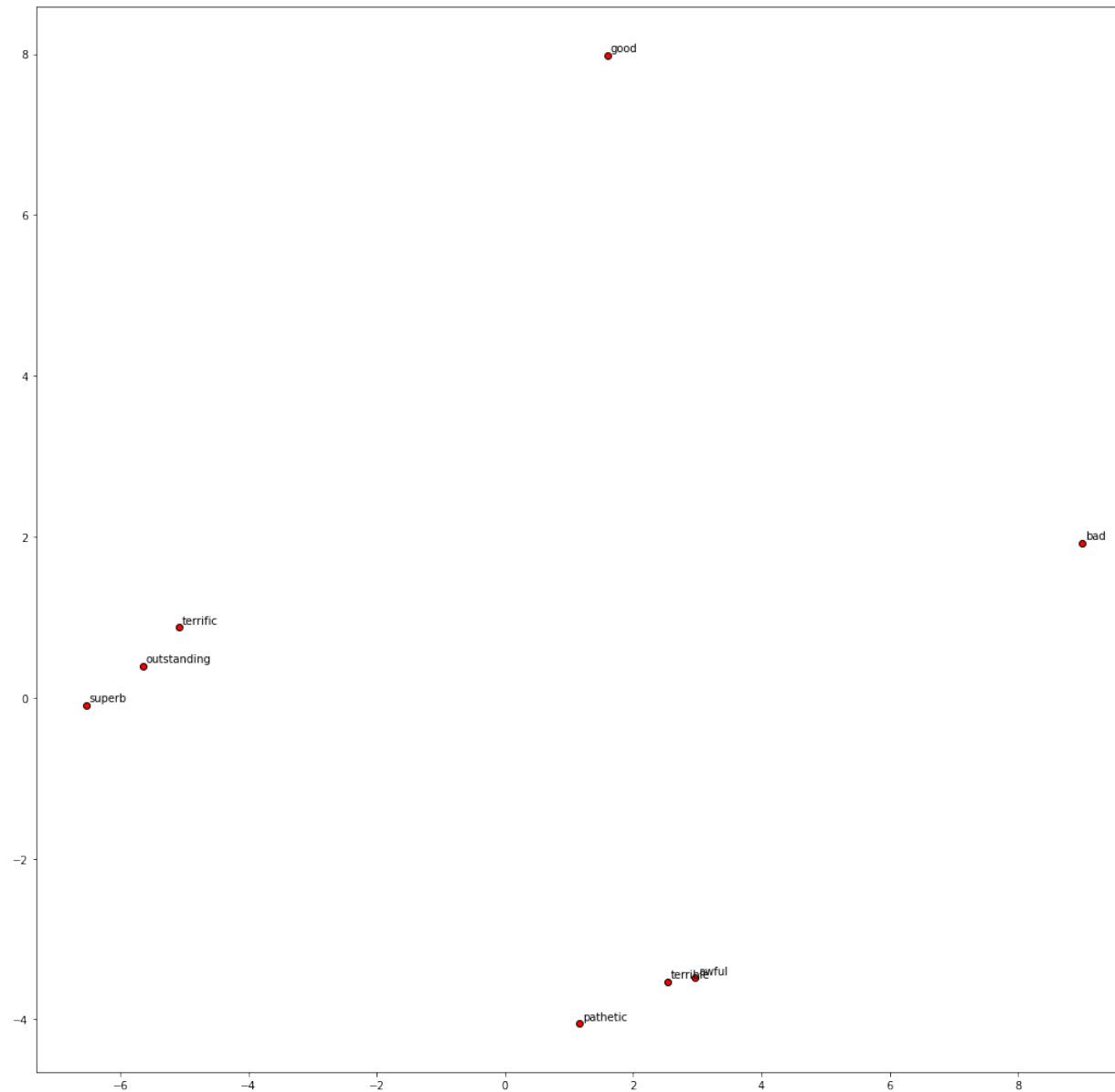
voisinage conservé
(distance)



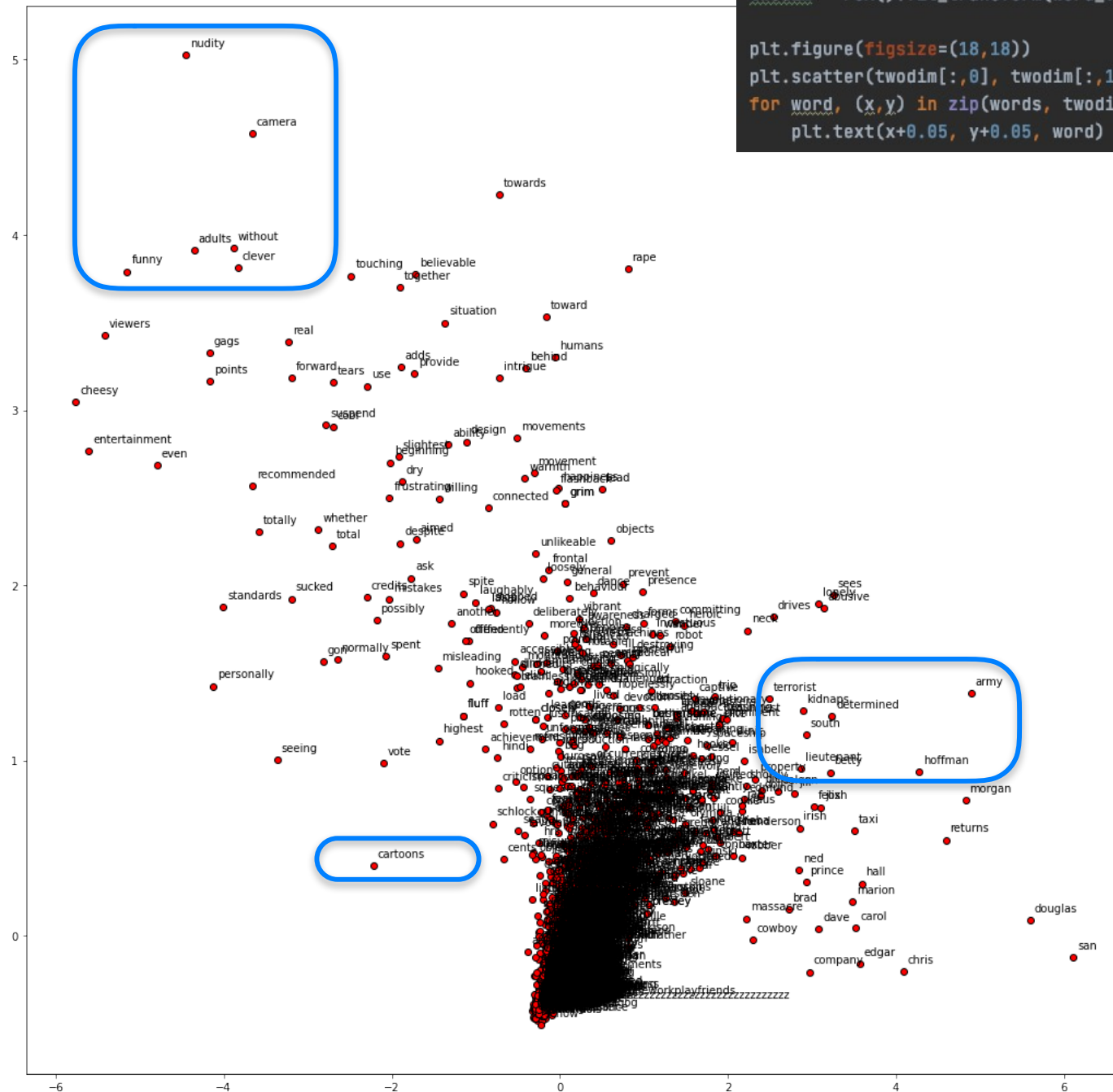
<https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/>

<https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b>


```
display_pca_scatterplot(model, ['superb', 'good', 'terrible', 'awful', 'pathetic', 'outstanding', 'terrific', 'bad'])
```



ACP
de 3000
mots pris
au hasard

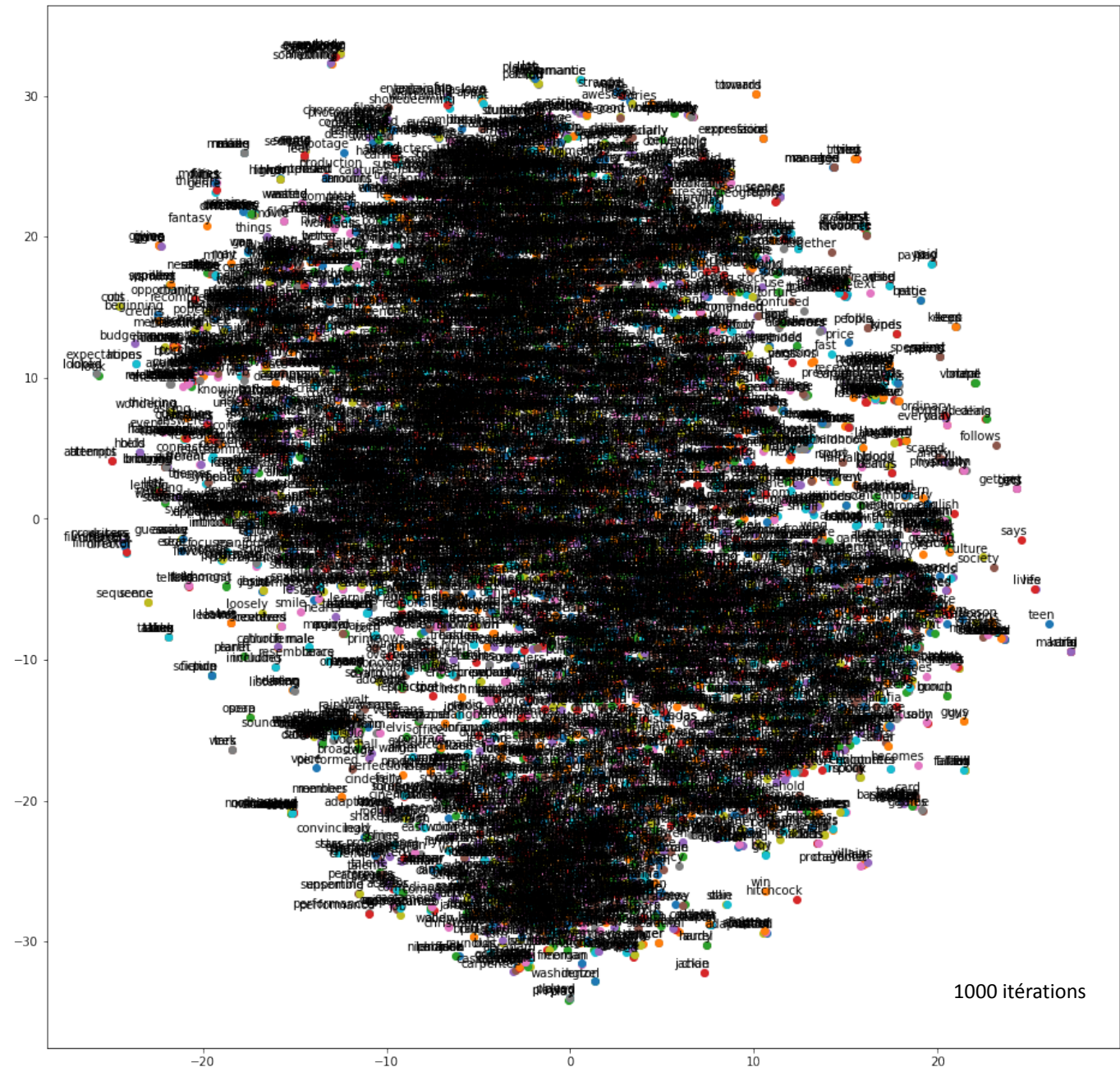
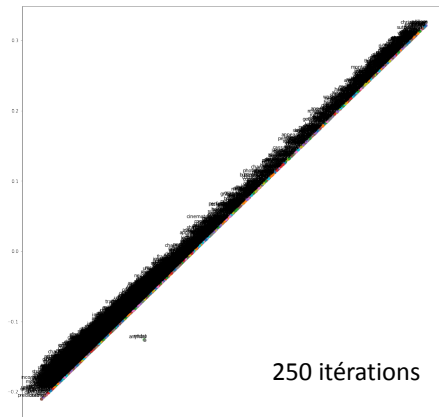


```
twodim = PCA().fit_transform(word_vectors)[:,:2]

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)
```

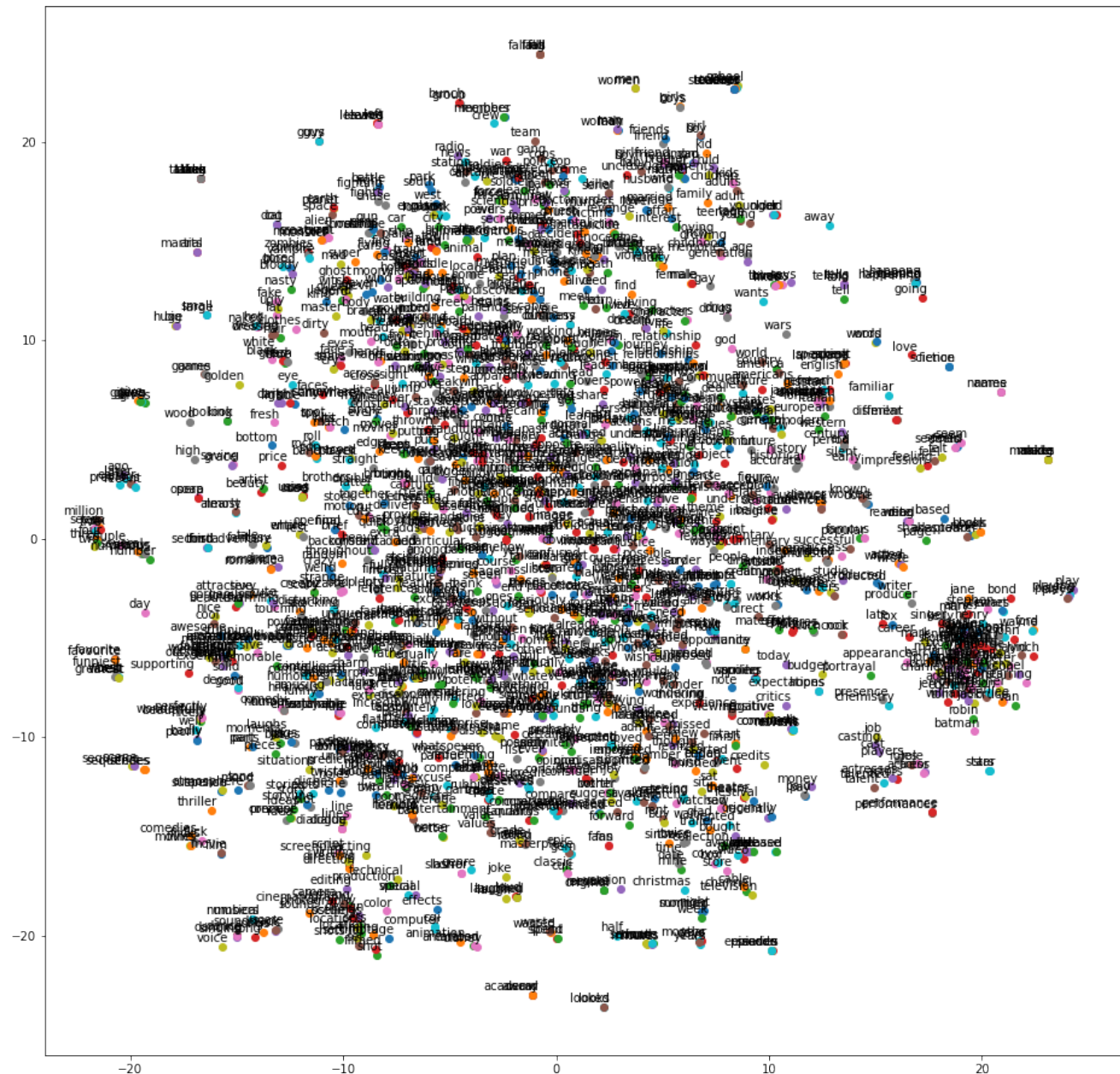


t-SNE
 en limitant
 Word2Vec
 aux mots ayant
 au moins
 100 occurrences
 (soit 6561 mots)



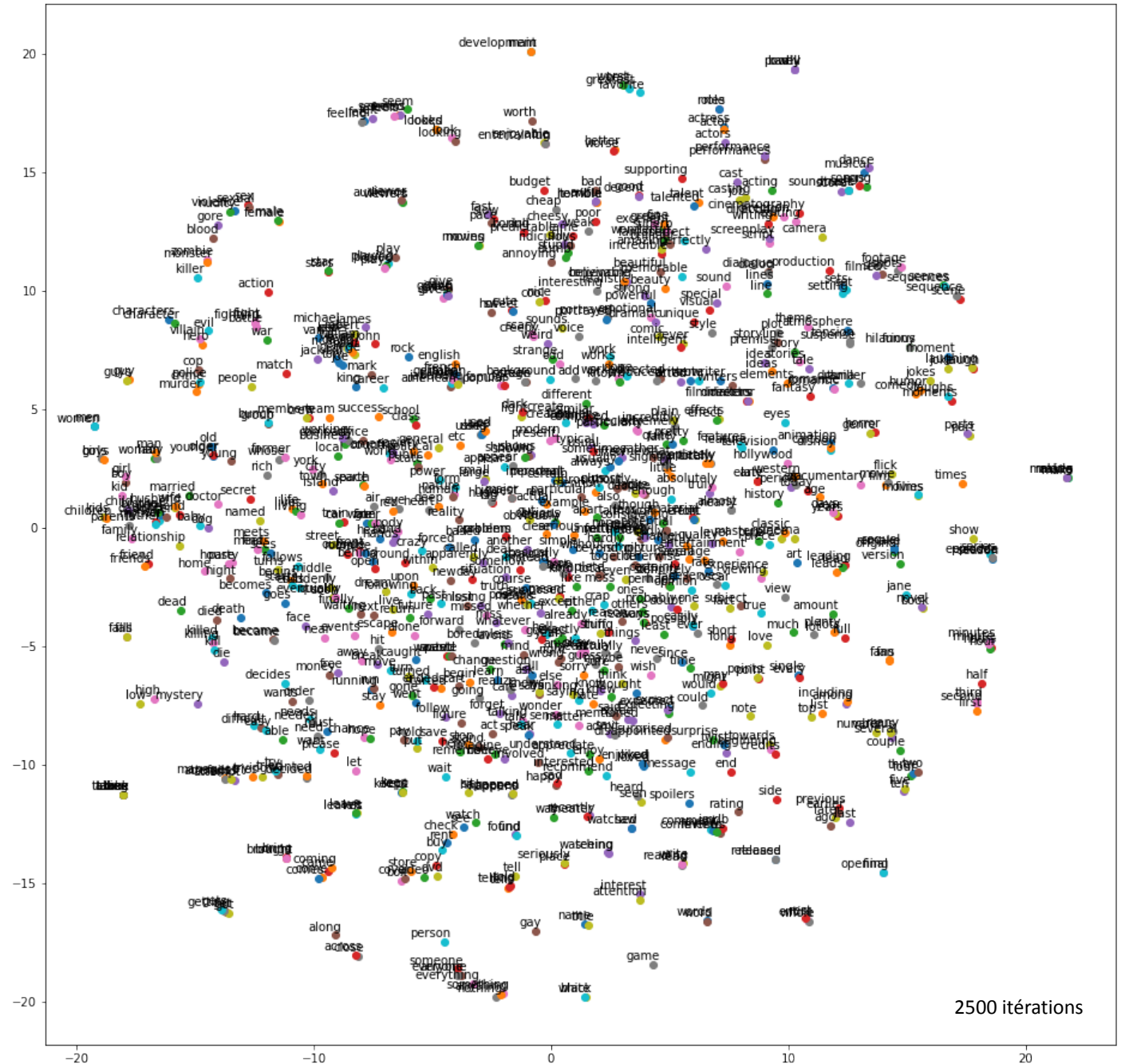
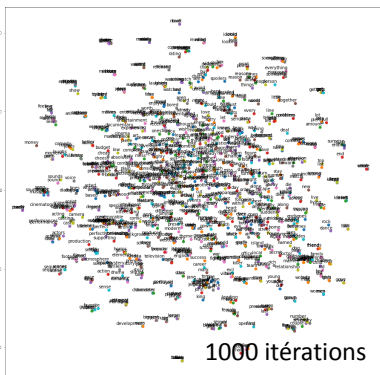
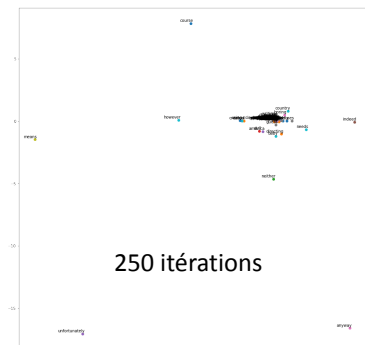
```
tsne_model = TSNE(perplexity=40, n_components=2, init='random', n_iter=1000, random_state=23, verbose=True, n_jobs=12)
```


t-SNE des mots dont
occurrences > 500
en limitant
Word2Vec
aux mots ayant
au moins
100 occurrences



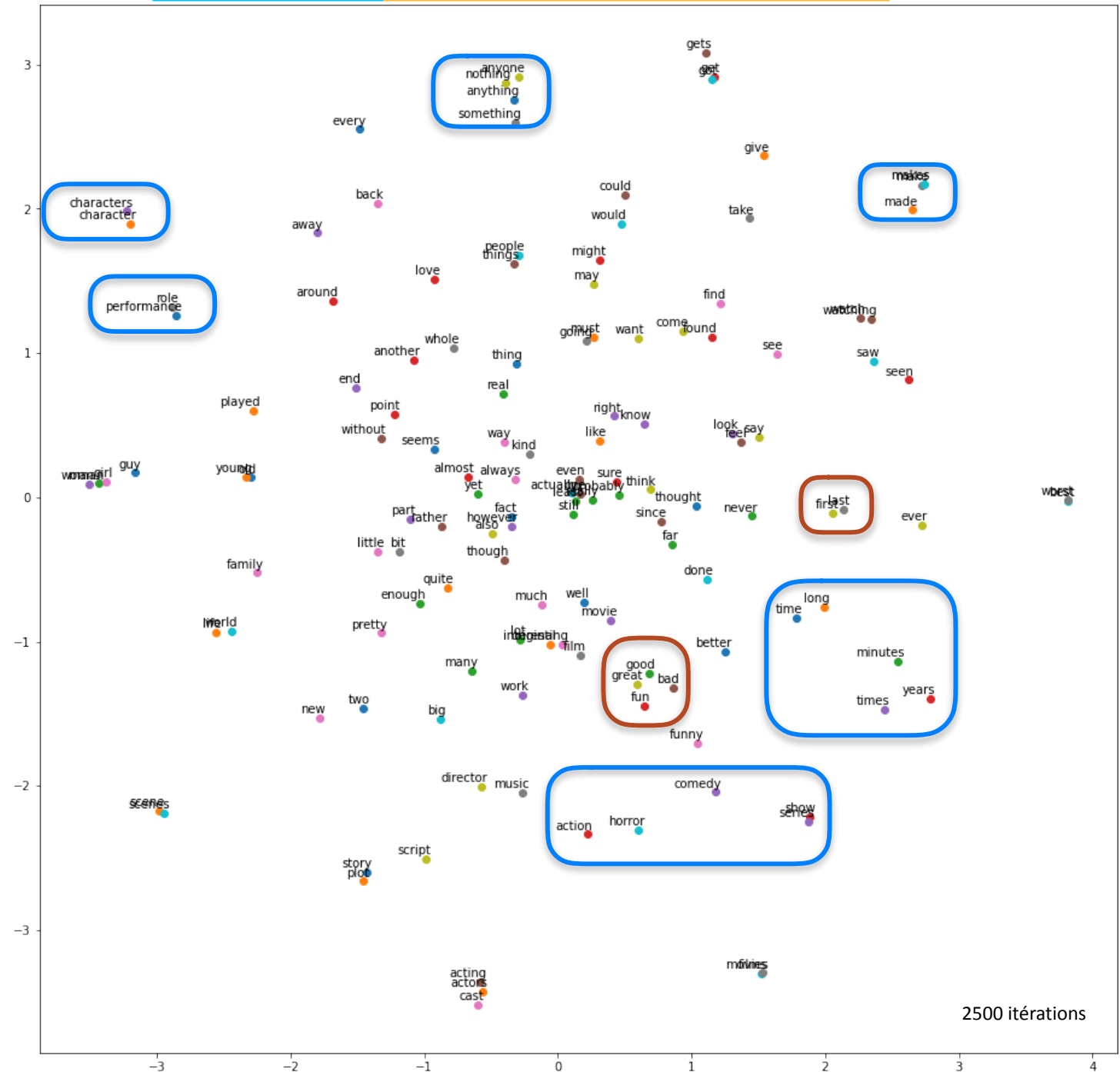
```
tsne_model = TSNE(perplexity=40, n_components=2, init='random', n_iter=1000, random_state=23, verbose=True, n_jobs=12)
```


t-SNE des mots dont
occurrences > 1000
en limitant
Word2Vec
aux mots ayant
au moins
100 occurrences



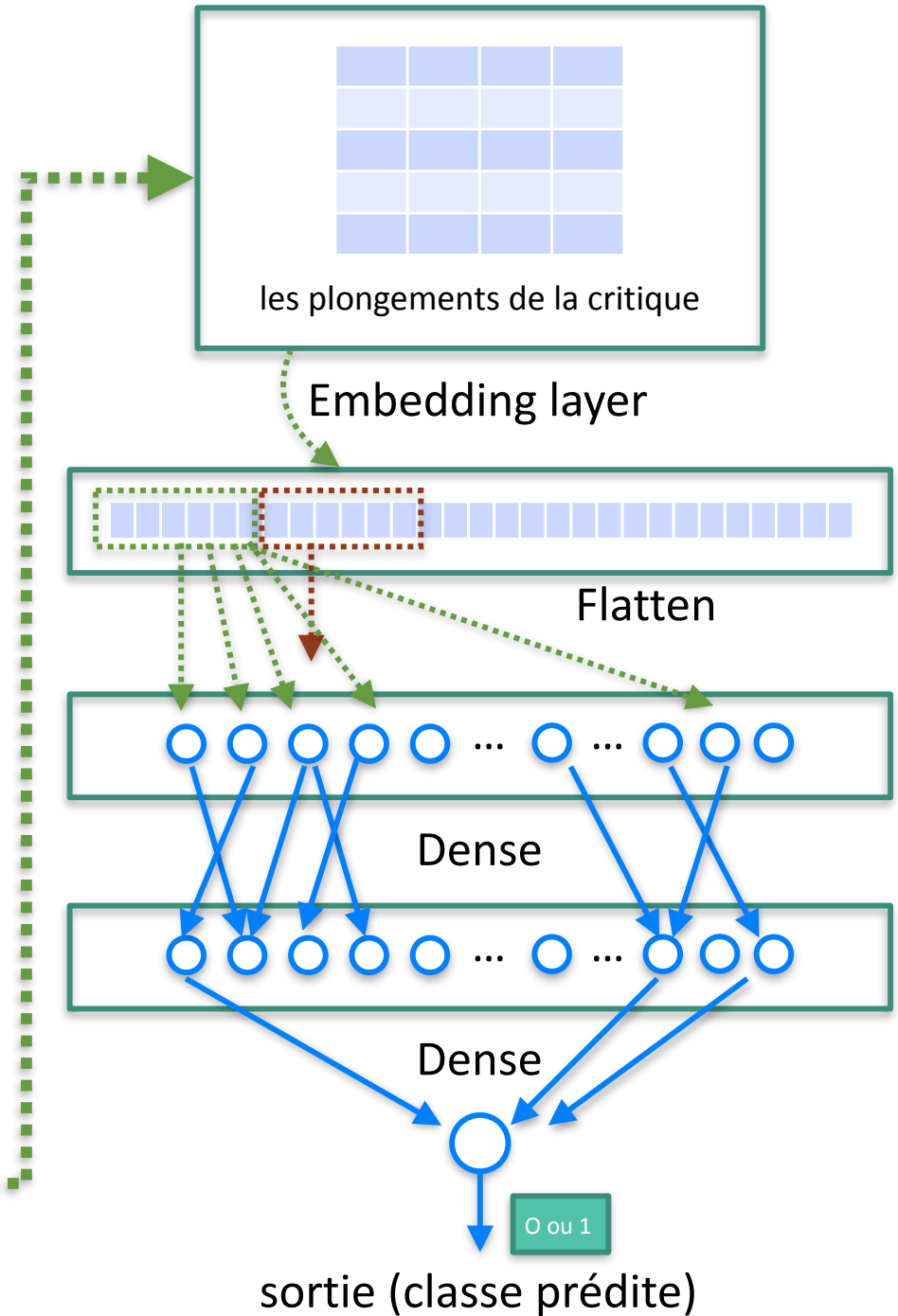
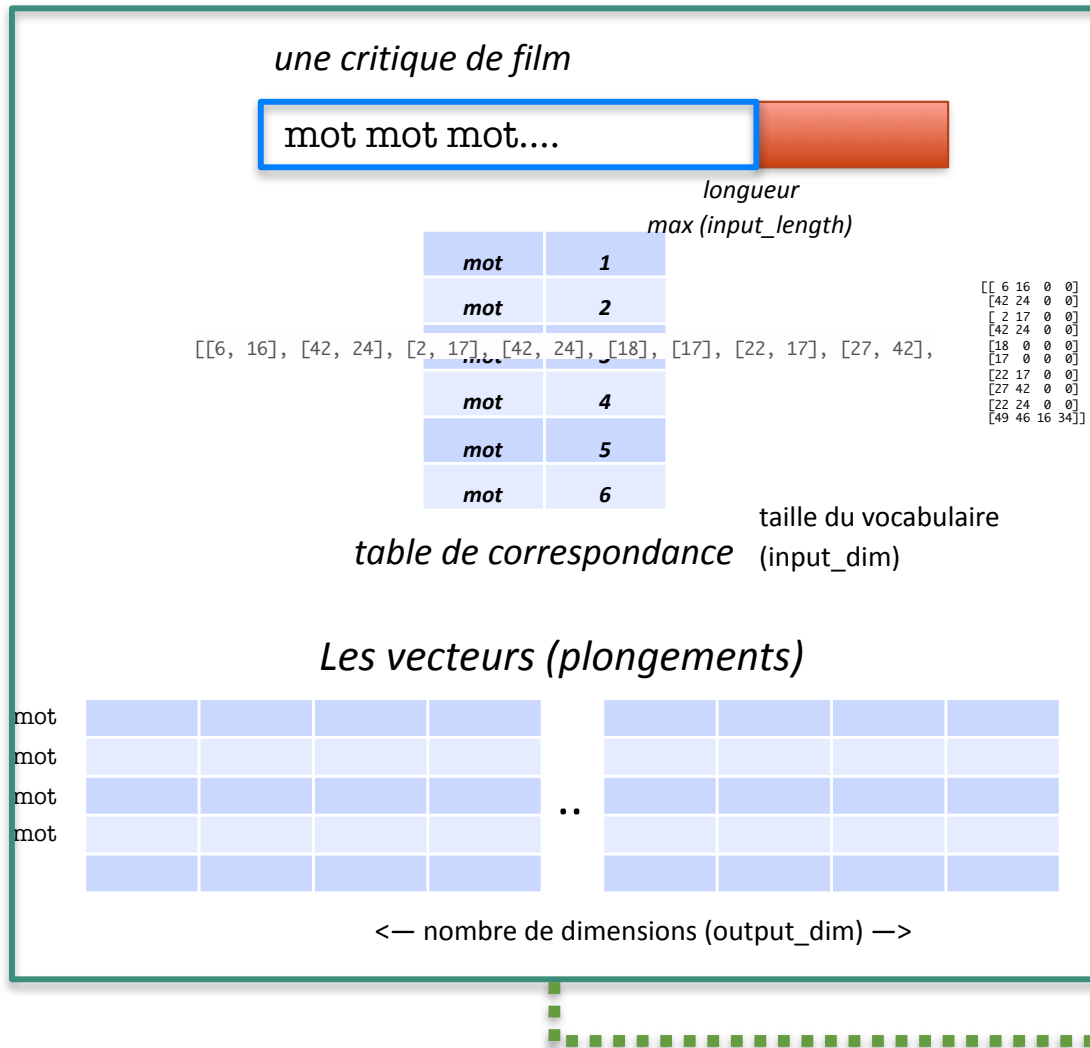
```
tsne_model = TSNE(perplexity=40, n_components=2, init='random', n_iter=1000, random_state=23, verbose=True, n_jobs=12)
```

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



```
tsne_model = TSNE(perplexity=40, n_components=2, init='random', n_iter=2500, random_state=23, verbose=True, n_jobs=12)
```

Architecture testée



```
DIMENSION_EMBEDDINGS = 200
```

```
modelEmbeddings = gensim.models.Word2Vec(sentences=review_lines, size=DIMENSION_EMBEDDINGS, window=5, workers=12, min_count=100)
```

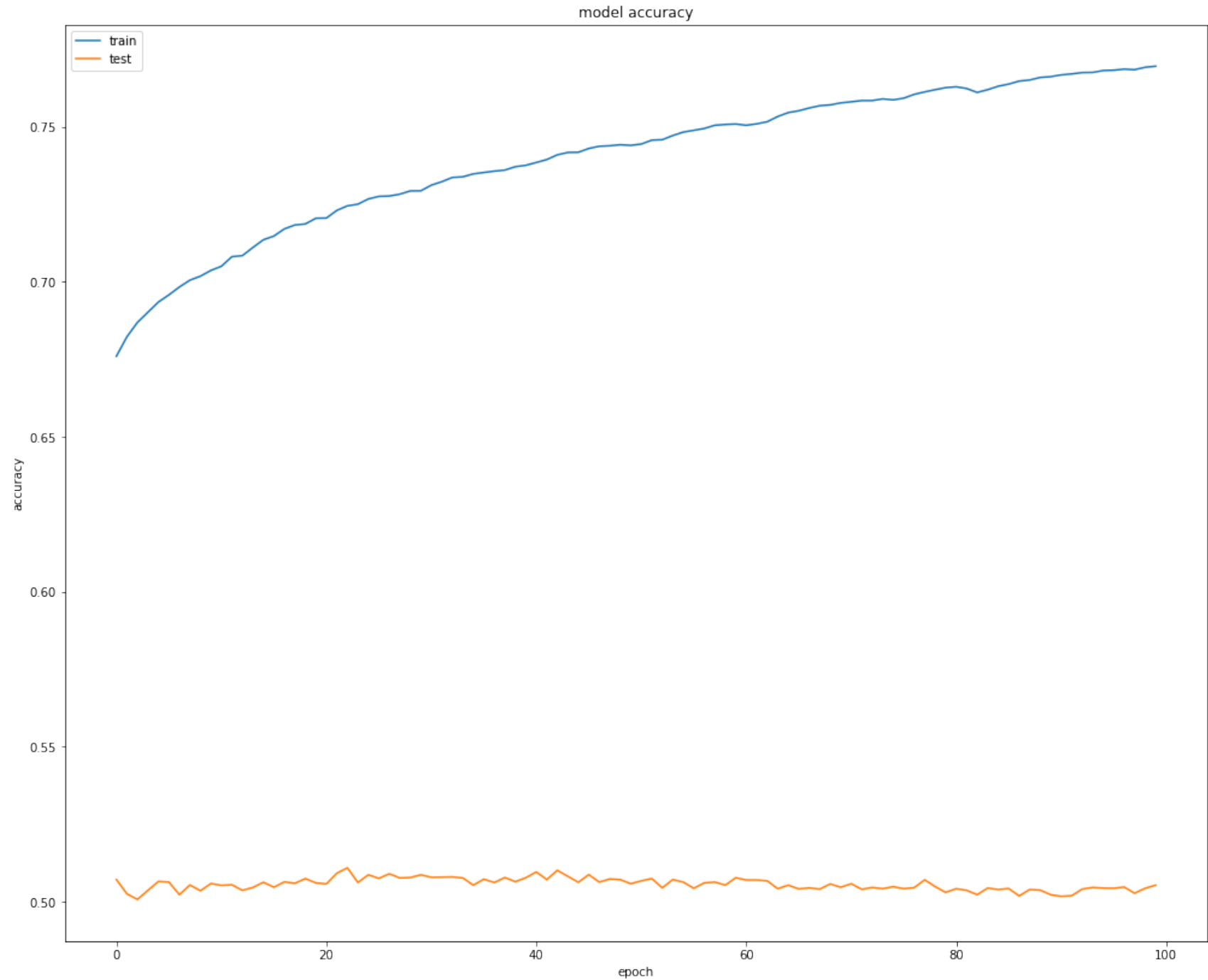
```
model.add(embedding_layer)
model.add(Flatten())
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

1er essai : on ne garde
que le mots qui apparaissent
au moins 100 fois
Réseau : une seule couche cachée

```
history = model.fit(X_train_pad, y_train, batch_size=128, epochs=100, validation_data=(X_test_pad, y_test), verbose=1)
```

Train on 35000 samples, validate on 15000 samples

Epoch	Train Samples	Train Loss	Train Accuracy	Val Loss	Val Accuracy
Epoch 1/100	35000/35000	0.4944	0.6760	1.0262	0.5071
Epoch 2/100	35000/35000	0.4831	0.6823	1.0537	0.5025
Epoch 3/100	35000/35000	0.4750	0.6869	1.1092	0.5007
Epoch 4/100	35000/35000	0.4708	0.6902	1.1248	0.5037
Epoch 5/100	35000/35000	0.4667	0.6935	1.2038	0.5065
Epoch 6/100	35000/35000	0.4619	0.6958	1.1686	0.5063
Epoch 7/100	35000/35000	0.4562	0.6983	1.2811	0.5023
Epoch 8/100	35000/35000	0.4543	0.7005	1.2396	0.5054
Epoch 9/100	35000/35000	0.4502	0.7018	1.3060	0.5035
Epoch 10/100	35000/35000	0.4459	0.7037	1.3753	0.5059
Epoch 11/100	35000/35000	0.4448	0.7050	1.3746	0.5053
Epoch 12/100	35000/35000	0.4424	0.7081	1.3833	0.5055
Epoch 13/100	35000/35000	0.4404	0.7085	1.4658	0.5037
Epoch 14/100	35000/35000	0.4375	0.7111	1.3758	0.5046
Epoch 15/100	35000/35000	0.4329	0.7136	1.5291	0.5063
Epoch 16/100	35000/35000	0.4289	0.7148	1.4570	0.5047



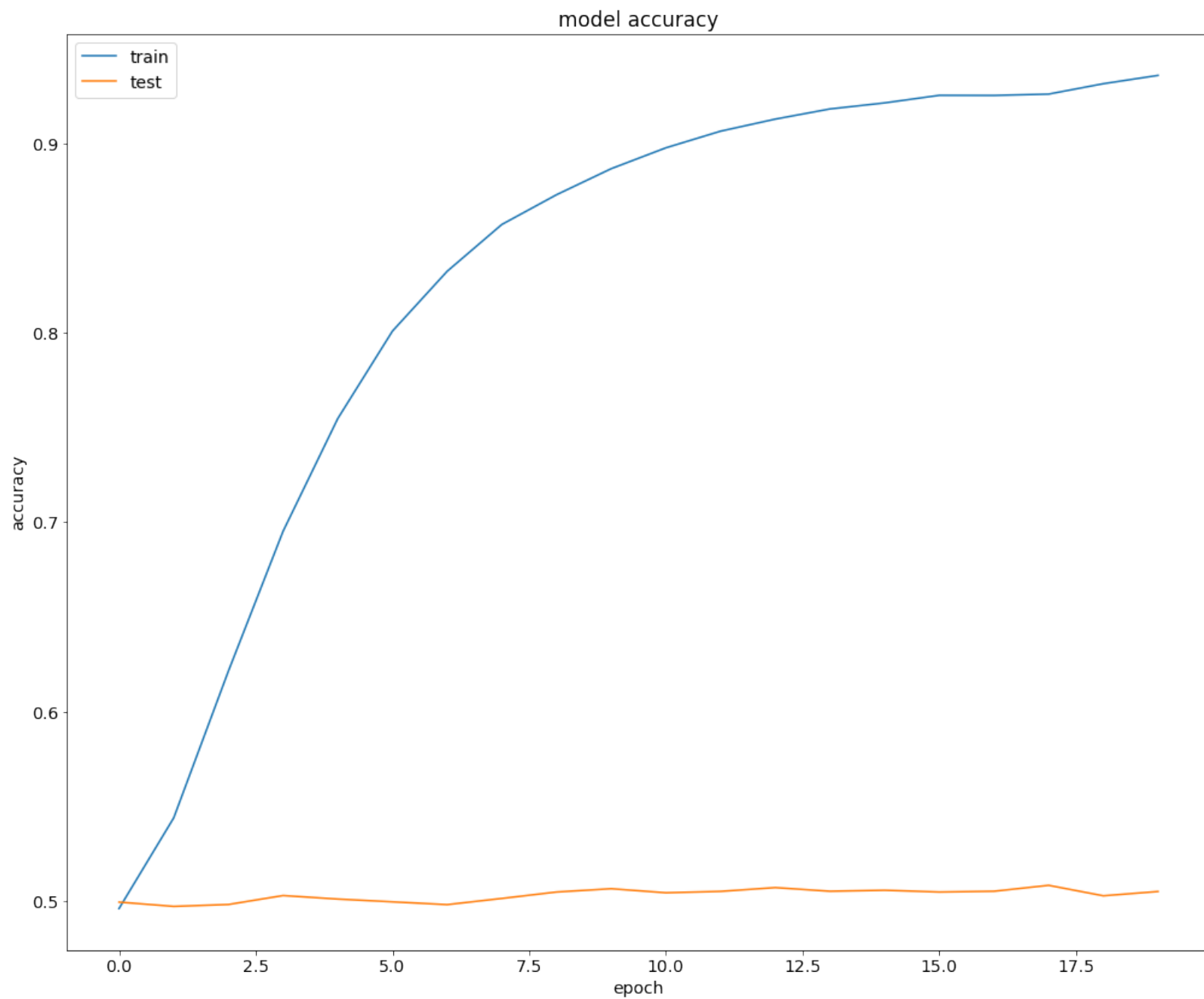
```
model.add(Flatten())
model.add(Dense(32, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

2^è essai :
on ajoute une 2^è couche

```
history = model.fit(X_train_pad, y_train, batch_size=128, epochs=20, validation_data=(X_test_pad, y_test), verbose=1)
```

Train on 35000 samples, validate on 15000 samples

Epoch	Train Samples	Time	Loss	Accuracy	Val Loss	Val Accuracy
Epoch 1/20	35000/35000	12s 343us/step	0.6959	0.4959	0.6937	0.4993
Epoch 2/20	35000/35000	11s 303us/step	0.6754	0.5438	0.7063	0.4970
Epoch 3/20	35000/35000	11s 302us/step	0.6101	0.6214	0.7265	0.4980
Epoch 4/20	35000/35000	11s 303us/step	0.5159	0.6953	0.7975	0.5027
Epoch 5/20	35000/35000	11s 302us/step	0.4270	0.7548	0.9331	0.5009
Epoch 6/20	35000/35000	11s 303us/step	0.3545	0.8011	1.0973	0.4994
Epoch 7/20	35000/35000	11s 303us/step	0.2949	0.8327	1.2164	0.4979
Epoch 8/20	35000/35000	11s 305us/step	0.2526	0.8574	1.4961	0.5012
Epoch 9/20	35000/35000	11s 304us/step	0.2285	0.8731	1.6137	0.5046
Epoch 10/20	35000/35000	11s 303us/step	0.2023	0.8869	1.7251	0.5063
Epoch 11/20	35000/35000	11s 305us/step	0.1832	0.8979	1.9321	0.5042
Epoch 12/20	35000/35000	11s 300us/step	0.1657	0.9067	1.9291	0.5049
Epoch 13/20	35000/35000	11s 302us/step	0.1539	0.9131	2.0730	0.5069



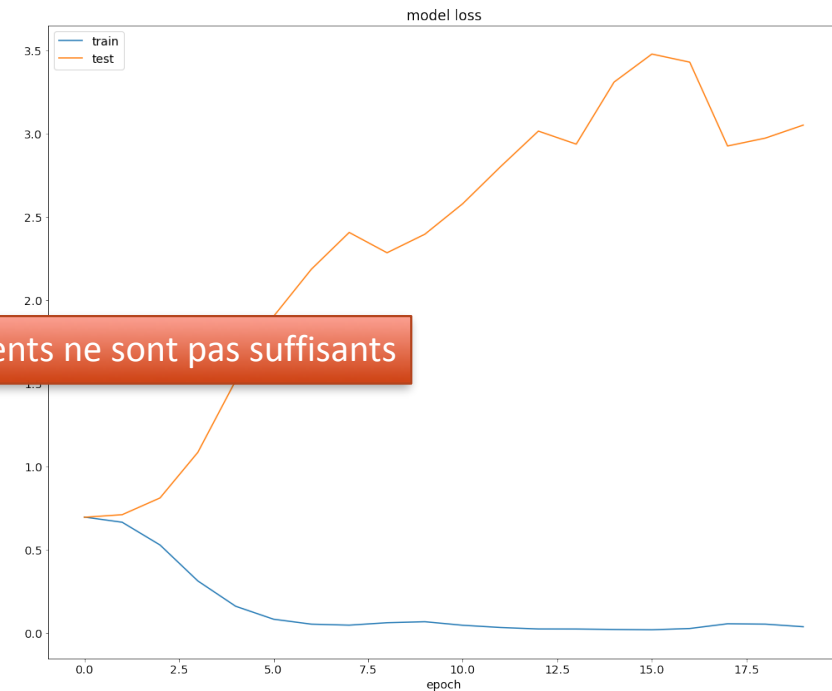
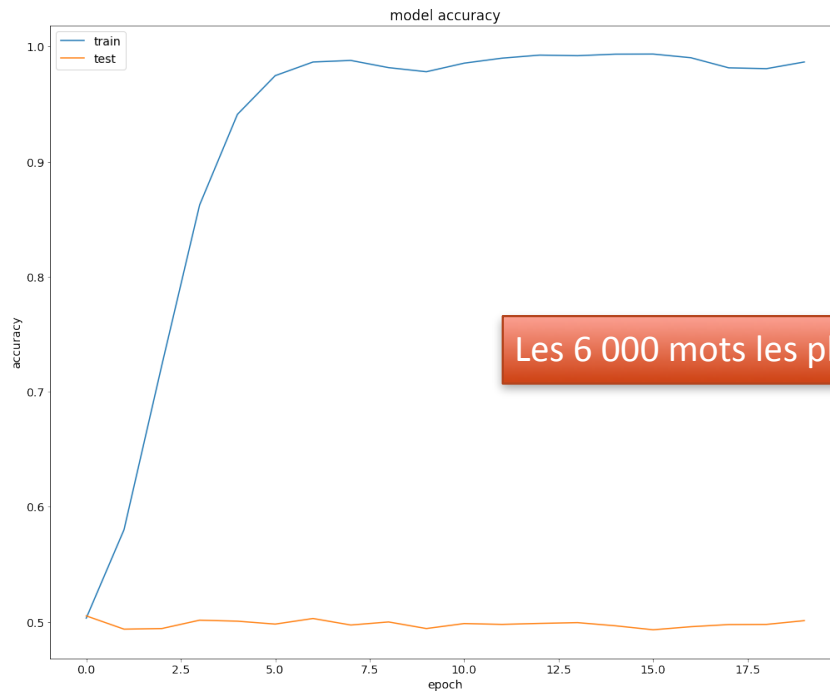
3è essai : on rajoute une 3è couche...

```
model.add(embedding_layer)
model.add(Flatten())
model.add(Dense(32, activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

Layer (type)	Output Shape	Param #
embedding_12 (Embedding)	(None, 256, 200)	19352000
flatten_7 (Flatten)	(None, 51200)	0
dense_18 (Dense)	(None, 32)	1638432
dense_19 (Dense)	(None, 32)	1056
dense_20 (Dense)	(None, 8)	264
dense_21 (Dense)	(None, 1)	9
Total params: 20,991,761		
Trainable params: 1,639,761		
Non-trainable params: 19,352,000		

Epoch 16/20

35000/35000 [=====] - 12s 330us/step - loss: 0.0199 - accuracy: 0.9935 - val_loss: 3.4786 - val_accuracy: 0.4931



Les 6 000 mots les plus fréquents ne sont pas suffisants

4è essai : on prend en compte plus de mots du vocabulaire

```
DIMENSION_EMBEDDINGS = 200
model = gensim.models.Word2Vec(sentences=review_lines, size=DIMENSION_EMBEDDINGS, window=5, workers=12, min_count=10)
```

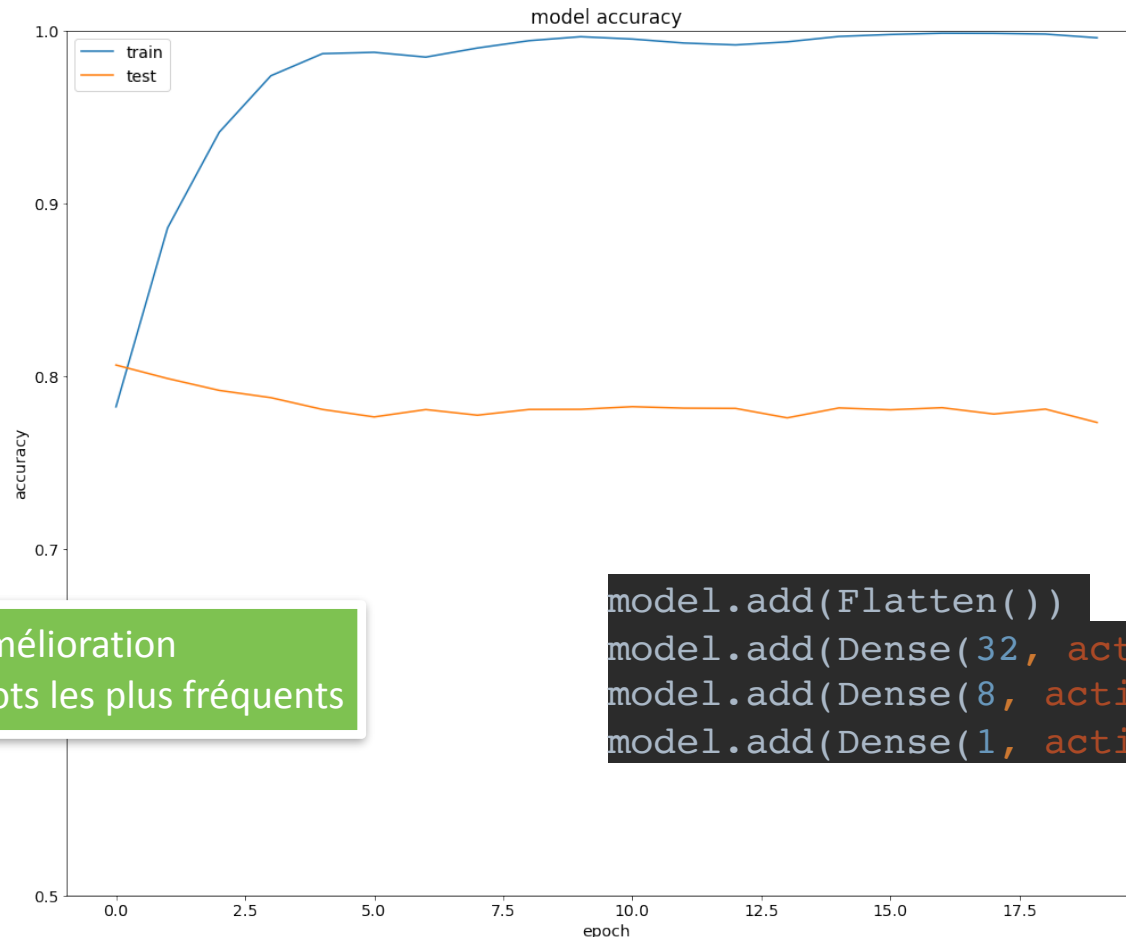
```
history = model.fit(X_train_pad, y_train, batch_size=128, epochs=20, validation_data=(X_test_pad, y_test), verbose=1)
```

Epoch 1/20

```
35000/35000 [=====] - 12s 340us/step - loss: 0.4560 - accuracy: 0.7825 - val_loss: 0.4157 - val_accuracy: 0.8066
```

Epoch 2/20

```
35000/35000 [=====] - 11s 304us/step - loss: 0.2631 - accuracy: 0.8859 - val_loss: 0.4545 - val_accuracy: 0.7988
```

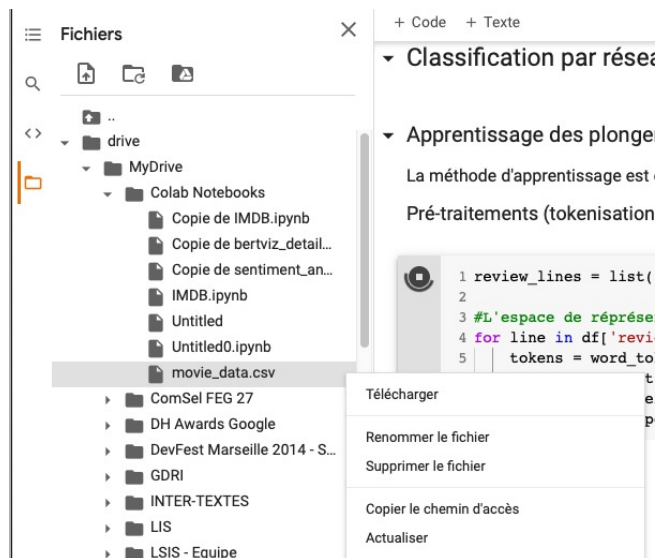
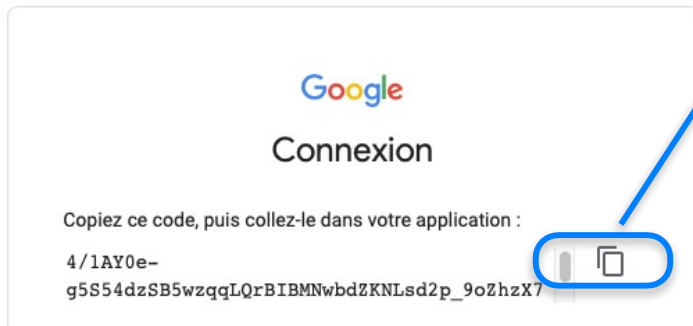
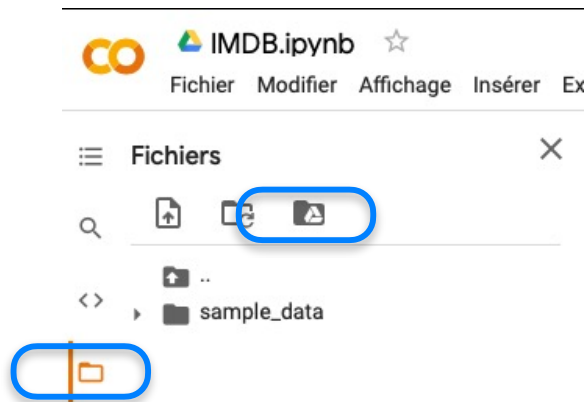


Nette amélioration
avec les 27 000 mots les plus fréquents

```
model.add(Flatten())
model.add(Dense(32, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
```

DANS GOOGLE COLAB

<https://colab.research.google.com>



▼ Analyse de sentiment sur les critiques d'IMDB

Le corpus peut être téléchargé ici : <http://ai.stanford.edu/~amaas/data/sentiment/> Placé dans le dossier de travail et décompresser.

Exemple inspiré de : <https://towardsdatascience.com/machine-learning-word-embeddings-part-1>

```
1 from google.colab import drive
2 drive.mount('/content/drive')
```

Mounted at /content/drive

!! Penser à changer le nom du répertoire/dossier de départ

```
[2] 1 repertoire_depart = '/content/drive/MyDrive/Colab Notebooks/'
    2 nomCSV = repertoire_depart + '/movie_data.csv'
```

Appel de la méthode entraînant le réseau et test au fur et à mesure des époques.

```
1 #selon la configuration de votre machine, des conflits entre bibliothèques peuvent survenir.
2 #Si Python quitte brutalement la ligne suivante peut permettre de contourner le problème
3 #sinon la mettre en commentaires
4 os.environ['KMP_DUPLICATE_LIB_OK']='True'
5
6 history = model.fit(X_train_pad, y_train, batch_size=32, epochs=3, validation_data=(X_test_pad, y_test), verbose=1)
```

```
Epoch 1/3
1094/1094 [=====] - 10s 9ms/step - loss: 0.5306 - accuracy: 0.7269 - val_loss: 0.4157 - val_accuracy: 0.8067
Epoch 2/3
1094/1094 [=====] - 8s 8ms/step - loss: 0.3056 - accuracy: 0.8690 - val_loss: 0.4371 - val_accuracy: 0.8022
Epoch 3/3
1094/1094 [=====] - 8s 8ms/step - loss: 0.2212 - accuracy: 0.9106 - val_loss: 0.4864 - val_accuracy: 0.7949
```

IMDB.ipynb

Fichier Modifier Affichage Insérer Exécution Outils Aide Toutes les modifications ont été enregistrées

Fichiers

- drive
 - MyDrive
 - Colab Notebooks
 - Copie de IMDB.ipynb
 - Copie de bertviz_detail...
 - Copie de sentiment_an...
 - IMDB.ipynb
 - Untitled
 - Untitled0.ipynb

Exécution

- Tout exécuter %/Ctrl+F9
- Exécuter avant %/Ctrl+F8
- Exécuter la cellule sélectionnée %/Ctrl+Enter
- Exécuter le code sélectionné %/Ctrl+Shift+Enter
- Exécuter à la suite %/Ctrl+F10
- Interrompre l'exécution %/Ctrl+M I
- Redémarrer l'environnement d'exécution %/Ctrl+M .
- Redémarrer et tout exécuter
- Réinitialiser l'environnement d'exécution
- Modifier le type d'exécution
- Gérer les sessions

Paramètres du notebook

Type d'exécution

py371

Accélérateur matériel

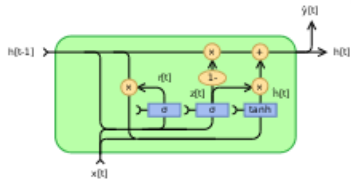
- ☒ None
- ☐ GPU
- ☐ TPU

ent de sortie des cellules de code lors de l'enregistrement de ce notebook

ANNULER ENREGISTRER

```
1 #selon la configuration de votre machine, des conflits entre bibliothèques peuvent survenir.
2 #Si Python quitte brutalement la ligne suivante peut permettre de contourner le problème
3 #sinon la mettre en commentaires
4 os.environ['KMP_DUPLICATE_LIB_OK']='True'
5
6 history = model.fit(X_train_pad, y_train, batch_size=32, epochs=3, validation_data=(X_test_pad, y_test), verbose=1)
```

```
Epoch 1/3
1094/1094 [=====] - 8s 6ms/step - loss: 0.4966 - accuracy: 0.7269 - val_loss: 0.4157 - val_accuracy: 0.8067
Epoch 2/3
1094/1094 [=====] - 6s 5ms/step - loss: 0.2925 - accuracy: 0.8690 - val_loss: 0.4371 - val_accuracy: 0.8022
Epoch 3/3
1094/1094 [=====] - 6s 5ms/step - loss: 0.2143 - accuracy: 0.9106 - val_loss: 0.4864 - val_accuracy: 0.7949
```

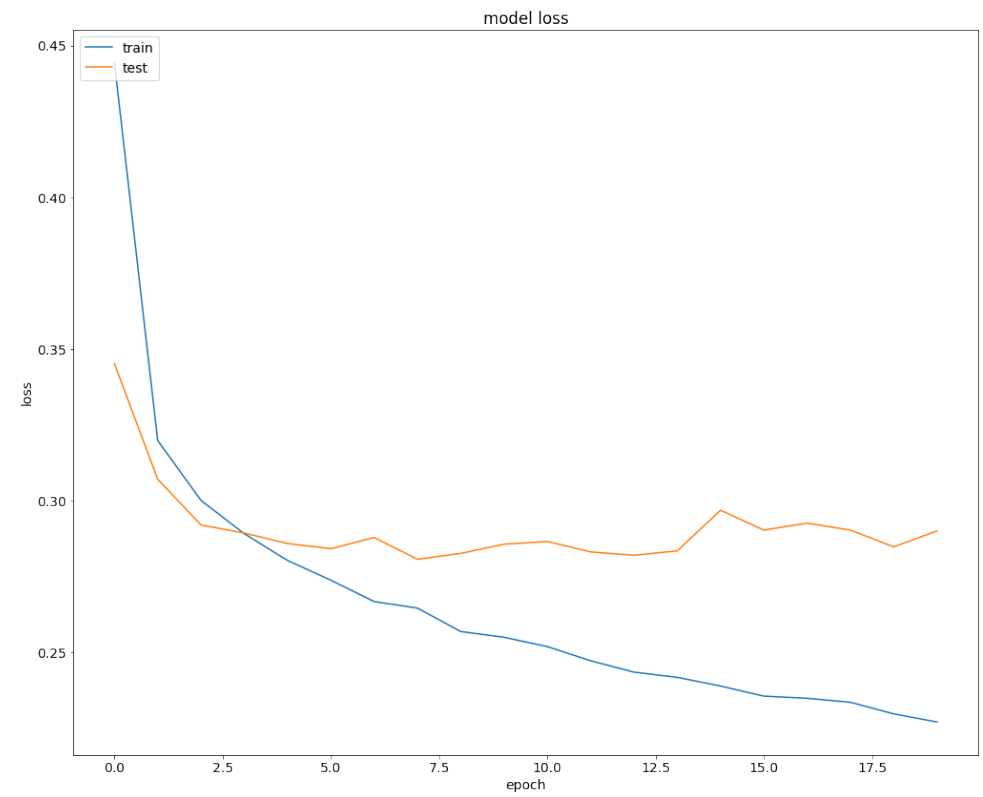
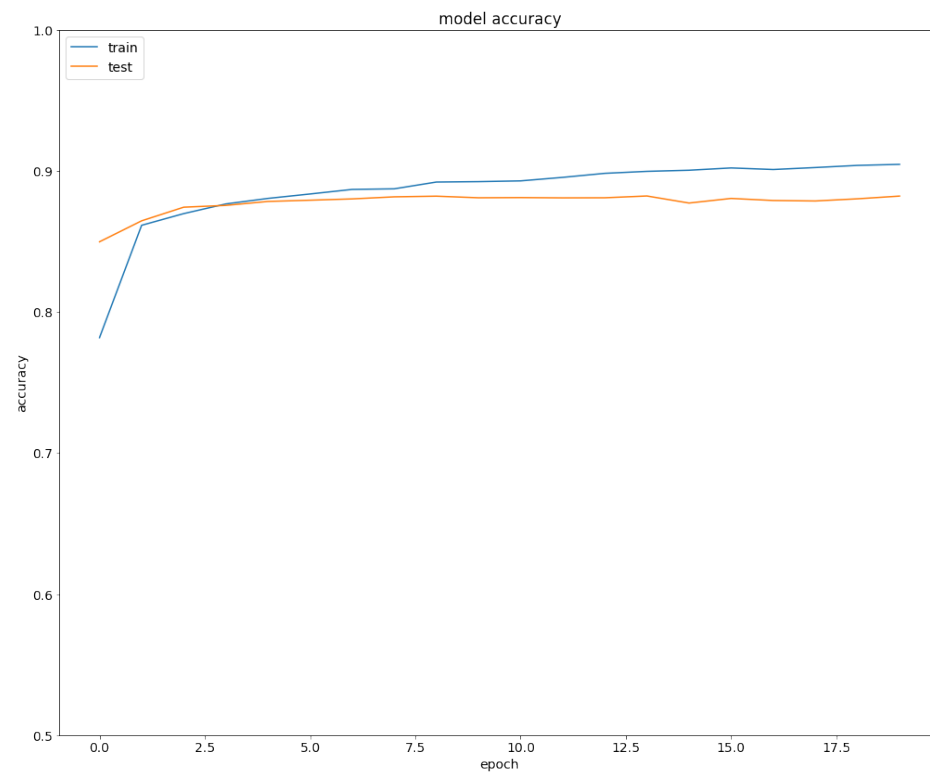
Model: "sequential_2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 128, 200)	19345800
gru (GRU)	(None, 32)	22464
dense_7 (Dense)	(None, 8)	264
dense_8 (Dense)	(None, 1)	9

Total params: 19,368,537
Trainable params: 22,737
Non-trainable params: 19,345,800

```
Epoch 1/20
1094/1094 [=====] - 187s 168ms/step - loss: 0.5532 - accuracy: 0.6923 - val_loss: 0.3451 - val_accuracy: 0.8497
Epoch 2/20
1094/1094 [=====] - 188s 172ms/step - loss: 0.3212 - accuracy: 0.8617 - val_loss: 0.3069 - val_accuracy: 0.8646
Epoch 3/20
1094/1094 [=====] - 188s 171ms/step - loss: 0.2948 - accuracy: 0.8736 - val_loss: 0.2919 - val_accuracy: 0.8743
Epoch 4/20
1094/1094 [=====] - 191s 175ms/step - loss: 0.2875 - accuracy: 0.8748 - val_loss: 0.2892 - val_accuracy: 0.8755
Epoch 5/20
1094/1094 [=====] - 185s 169ms/step - loss: 0.2820 - accuracy: 0.8804 - val_loss: 0.2858 - val_accuracy: 0.8783
Epoch 6/20
1094/1094 [=====] - 182s 166ms/step - loss: 0.2721 - accuracy: 0.8833 - val_loss: 0.2841 - val_accuracy: 0.8792
Epoch 7/20
1094/1094 [=====] - 185s 169ms/step - loss: 0.2629 - accuracy: 0.8881 - val_loss: 0.2877 - val_accuracy: 0.8801
Epoch 8/20
1094/1094 [=====] - 178s 162ms/step - loss: 0.2607 - accuracy: 0.8869 - val_loss: 0.2806 - val_accuracy: 0.8816
Epoch 9/20
1094/1094 [=====] - 182s 166ms/step - loss: 0.2553 - accuracy: 0.8929 - val_loss: 0.2825 - val_accuracy: 0.8821
```

```
1094/1094 [=====] - 24s 22ms/step - loss: 0.1825 - accuracy: 0.9275
469/469 [=====] - 10s 21ms/step - loss: 0.2899 - accuracy: 0.8821
```



En conclusion

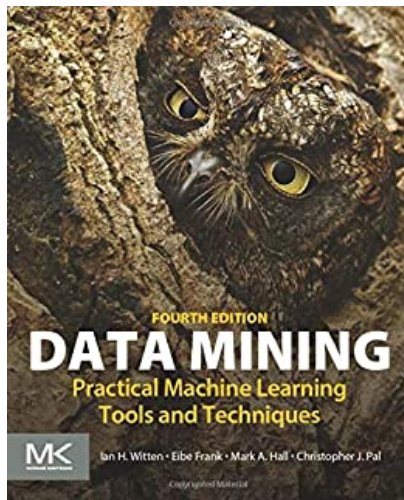
- **Approche bayésienne « naïve » : 0,85**
 - durée d'apprentissage : quelques secondes

- **Approche neuronale « plongements + couches denses » :**
 - mal configurée : 0,50 (soit l'équivalent d'un tirage aléatoire...)
 - **après quelques réglages et essais : 0,80**
 - durée d'apprentissage
 - avec CPU seul 12 cœurs : environ 10s / epoch, soit 3 mn
 - avec GPU (Google Colab) : environ 8s. / epoch

- **Approche neuronale « plongements + réseaux récurrents »**
 - **meilleur score : 0,88 (soit 3% de gain) — 0,9**
 - durée d'apprentissage :
 - avec CPU seul : environ 3000 s. / epoch, soit > 24 h.
 - avec TPU (Google Colab) : environ 200 s. / epoch

POUR ALLER PLUS LOIN

Références et liens pour WEKA



Data Mining: Practical Machine Learning Tools and Techniques
de Ian H. Witten , Eibe Frank, et al.



Data Mining with Weka

WekaMOOC
21 k abonnés

ACCUEIL VIDÉOS PLAYLISTS COMMUNAUTÉ CHAÎNES À PROPOS

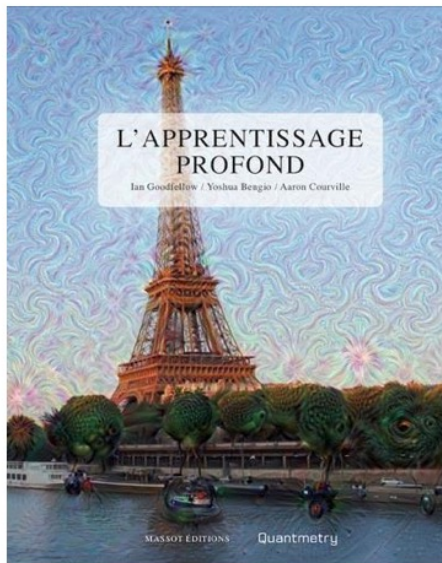
More Data Mining with Weka - FutureLearn
WekaMOOC • 11 k vues • il y a 3 ans
More Data Mining with Weka: online course with FutureLearn from the University of Waikato. First session starts 8 May 2017 [https://www.futurelearn.com/courses/more-data-mining-with-weka/...](https://www.futurelearn.com/courses/more-data-mining-with-weka/)

Data Mining with Weka ▶ TOUT REGARDER
Department of Computer Science, University of Waikato, New Zealand. <https://weka.waikato.ac.nz/>

Video Title	Duration	Views	Age
Data Mining with Weka: Trailer	1:40	171 k	7 ans
Data Mining with Weka (1.1: Introduction)	5:35	154 k	7 ans
Data Mining with Weka (1.2: Exploring the Explorer)	9:00	115 k	7 ans
Data Mining with Weka (1.3: Exploring datasets)	11:06	101 k	7 ans
Data Mining with Weka (1.4: Building a classifier)	10:38	106 k	7 ans

More Data Mining with Weka ▶ TOUT REGARDER
Department of Computer Science, University of Waikato, New Zealand. <http://weka.waikato.ac.nz/>

Video Title	Duration	Views	Age
More Data Mining with Weka: Trailer	3:39	31 k	6 ans
More Data Mining with Weka (1.1: Introduction)	6:39	19 k	6 ans
More Data Mining with Weka (1.2: Exploring the...)	11:17	23 k	6 ans
More Data Mining with Weka (1.3: Comparing classifiers)	7:53	23 k	6 ans
More Data Mining with Weka (1.4: The Knowledge Flow...)	7:43	20 k	6 ans



L'apprentissage profond

Yoshua Bengio, Ian Goodfellow, Aaron Courville

Massot Editions - 18 Octobre 2018

Sciences & Techniques

Voir les détails produits



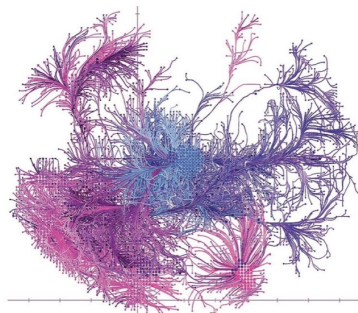
★★★★★ (Aucun avis)

À propos

Écrit par trois experts dans le domaine, Deep Learning est le seul livre complet sur le sujet. Il fournit une perspective générale et des préliminaires mathématiques indispensables aux ingénieurs en logiciel et aux étudiants qui entrent sur le terrain, et sert de référence aux autorités. Elon Musk, cofondateur et PDG de Tesla et SpaceX étudiants L'apprentissage profond (ou deep learning) est un apprentissage automatique qui permet à l'ordinateur d'apprendre par l'expérience et de comprendre le monde en termes de hiérarchie de concepts. Parce que l'ordinateur recueille des connaissances à partir de l'expérience, il n'est pas nécessaire qu'un opérateur humain spécifie formellement toutes les connaissances dont l'ordinateur a besoin. Cet ouvrage présente un large éventail de sujets d'apprentissage profond.

Le Lire la suite ▼

JEAN-CLAUDE HEUDIN
Comprendre le
DEEP LEARNING
Une introduction aux réseaux de neurones

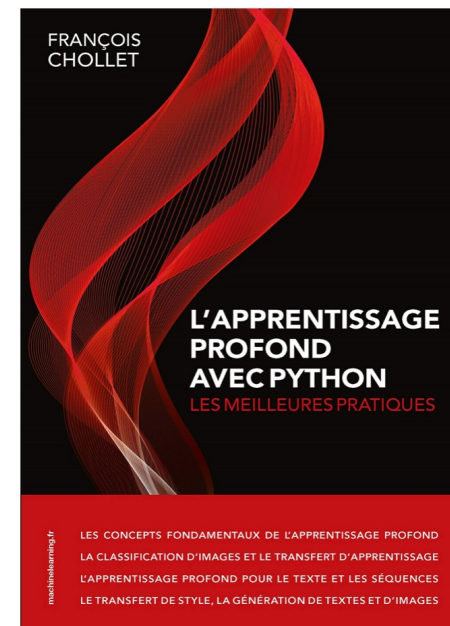


Yann Le Cun

Prix Turing

Quand la machine apprend

La révolution des neurones artificiels
et de l'apprentissage profond



<https://d2l.ai/index.html>



DIVE INTO DEEP LEARNING

Dive into Deep Learning

[Courses](#)
[PDF](#)
[All Notebooks](#)
[Discuss](#)
[GitHub](#)
[中文版](#)

Preface

Installation

Notation

1. Introduction

2. Preliminaries

3. Linear Neural Networks

4. Multilayer Perceptrons

5. Deep Learning Computation

6. Convolutional Neural Networks

7. Modern Convolutional Neural Networks

8. Recurrent Neural Networks

9. Modern Recurrent Neural Networks

10. Attention Mechanisms

11. Optimization Algorithms

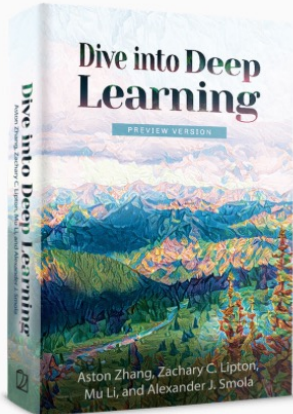
12. Computational Performance

13. Computer Vision

14. Natural Language Processing: Pretraining

15. Natural Language Processing: Applications

16. Recommender Systems



Dive into Deep Learning

Interactive deep learning book with code, math, and discussions

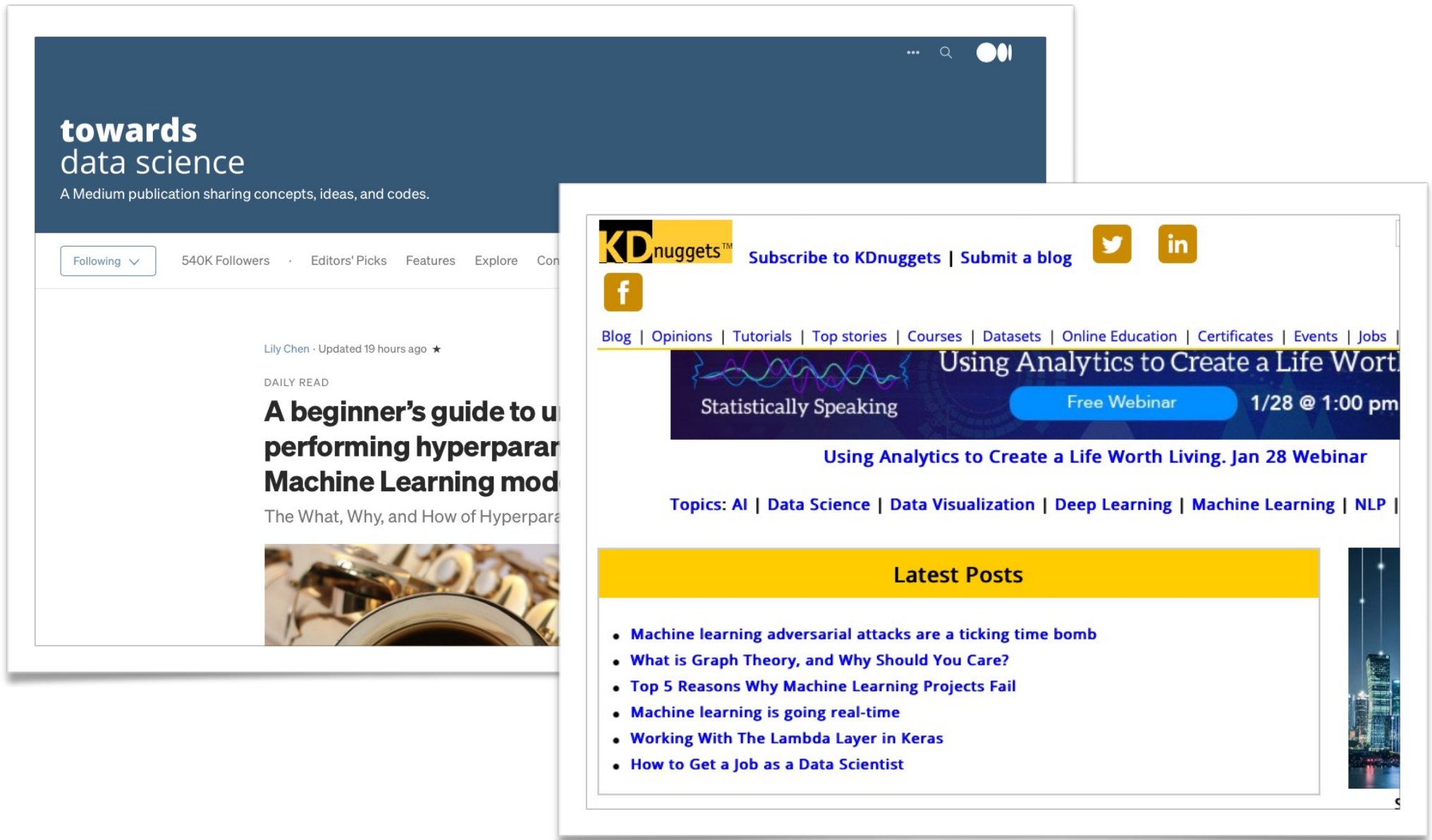
Implemented with **NumPy/MXNet, PyTorch, and TensorFlow**

Adopted at 175 universities from 40 countries

Announcements

- [Jan 2021] Check out the brand-new [Chapter: Attention Mechanisms](#). We have also completed PyTorch implementations. To keep track of the latest updates, please follow D2L's [open-source project](#).
- [Oct 2020] We have added TensorFlow implementations up to Chapter 7 (Modern CNNs).
- [Apr 2020] We have revamped [Chapter: NLP pretraining](#) and [Chapter: NLP applications](#), and added sections of BERT and natural language inference.
- [Jul 2019] The Chinese version is the [No. 1 best seller](#) of new books in "Computers and Internet" at the largest Chinese online bookstore.
- [May 2019] Slides, Jupyter notebooks, assignments, and videos of the Berkeley course can be found at the [syllabus page](#).

<https://towardsdatascience.com>



<https://www.kdnuggets.com>

<https://machinelearningmastery.com>

The Deck is Stacked Against Developers

Machine learning is taught by academics, for academics.

That's why most material is so *dry* and *math-heavy*.

Developers need to know what works and how to use it.

We need *less math* and *more tutorials with working code*.



Welcome to Machine Learning Mastery!

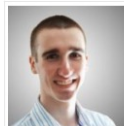
Hi, I'm Jason Brownlee PhD and I help developers like you skip years ahead.

Discover how to get **better results, faster**.

Click the button below to get my free EBook and accelerate your next project
(and access to my exclusive email course).

I'm Ready! Send it To Me!

Join over 150,000 practitioners who already have a head start.



I love your site by the way. It's one of the few ML sources I've come across that explains things clearly rather than writing everything as if it were an academic paper.

Kevin Beaulieu
Software Engineer



Your work has been VERY helpful for me as an aspiring Data Scientist!

David Dalisay
Junior Data Scientist

Quick-Start Guides

Discover the shortest path to a result.

Top guides include:

- [How to get started](#)
- [Understand algorithms](#)
- [Time series forecasting](#)

[>>Guides](#)

Latest Tutorials

Save weeks of searching and debugging.

Top tutorials include:

- [How to install everything](#)
- [Your first complete project](#)
- [Your first neural network](#)

[>>Blog](#)

EBook Catalog

Get years of experience in a PDF.

Top sellers include:

- [Master Machine Learning Algorithms](#)
- [Machine Learning With Python](#)
- [Deep Learning With Python](#)

[>>EBooks](#)