Clustering de textes courts pour relier des députés

Présentation de stage technique, MAIN 4

Achille Baucher
Encadré par Marie-Jeanne Lesot et Adrien Revault d'Allones

Octobre 2020

SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Présentation

Données

- 20 000 lettres
- $\sim$ 160 mots

Député : Bertrand Sorre
Ministre : Ministre de l'intérieur
Date : 12 Juin 2018
Rubrique : Sécurité routière
Titre : Précision sur les 80 km/h

Lettre :
M. Bertrand Sorre attire l'attention de M. le ministre d'État, ministre de l'intérieur, **sur la décision prise de réduire la vitesse maximale autorisée à 80 km/h**, à compter du 1er juillet 2018, sur les routes à double sens sans séparateur central (limitée actuellement à 90 km/h). Cette disposition permettra de sauver entre 300 et 400 vies par an selon le comité des experts du conseil national de la sécurité routière dans son rapport du 29 novembre 2013. Fréquemment questionné à ce sujet dans la circonscription de La Manche dont il est l'élu, *il aimerait savoir si cette décision implique également des modifications sur les vitesses actuellement autorisées pour les professionnels de la route (transport routier, autobus) ou pour les apprentis conducteurs d'un véhicule léger, titulaires d'un permis depuis de moins de 2 ans.*

SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Présentation

Objectif

- Clustering des lettres
- Même sujet précis
- Lien entre députés

SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Présentation

Normalisation

M. Bertrand Sorre attire l'attention de M. le ministre d'État, ministre de l'intérieur, sur la décision prise de réduire la vitesse maximale autorisée à 80 km/h.

- Ponctuation
- Stopwords
- Minuscule
- Infinitif singulier

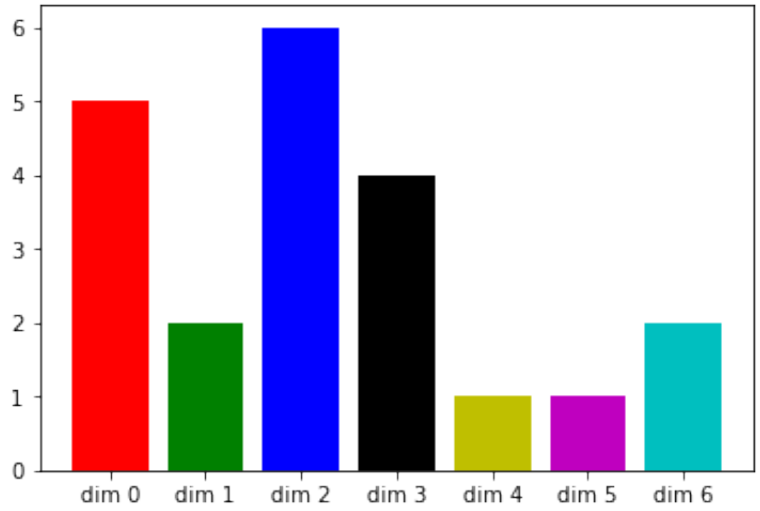
$\rightarrow$ *bertrand sorre attirer attention ministre état ministre intérieur décision prendre réduire vitesse maximale autorisation 80 km/h.*

SORBONNE UNIVERSITÉ

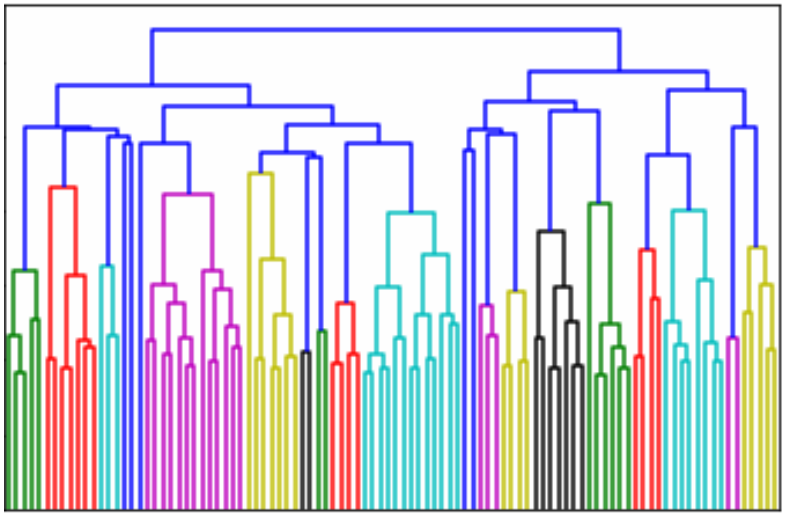
POLYTECH SORBONNE

Clustering de documents courts

État de l'art



- Représentation vectorielle



- Clustering classique

Représenter un document sous forme de vecteur

Principe et TF

- Un mot par dimension
- Term-Frequency

Mots / Documents	Covid	chloroquine	...	santé
Document 1 (TF)	4	2	...	3
Document 2 (TF)	2	0	...	3

SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Représenter un document sous forme de vecteur

TF-IDF

- Term-Frequency Inverse Document Frequency
- + Mots singuliers
- - Mots communs

Documents \ Mots	Covid	chloroquine	...	santé
Document 1 (TF)	4	2	...	3
Document 2 (TF)	2	0	...	3
Document 1 (TF-IDF)	3.	3.5	...	1.2
Document 2 (TF-IDF)	1.5	0.	...	1.2

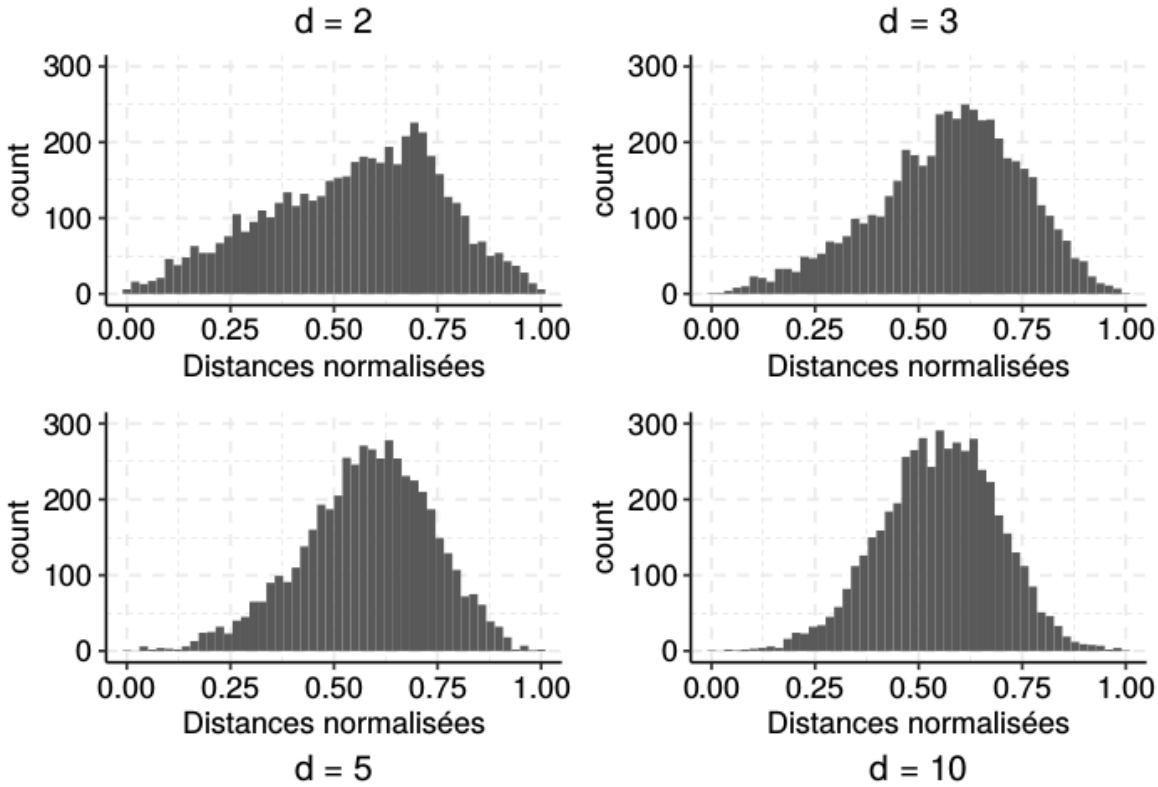
- $n \approx 9000$ dimensions !

SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Malédiction de la dimensionnalité

Exemple : Concentration des distances

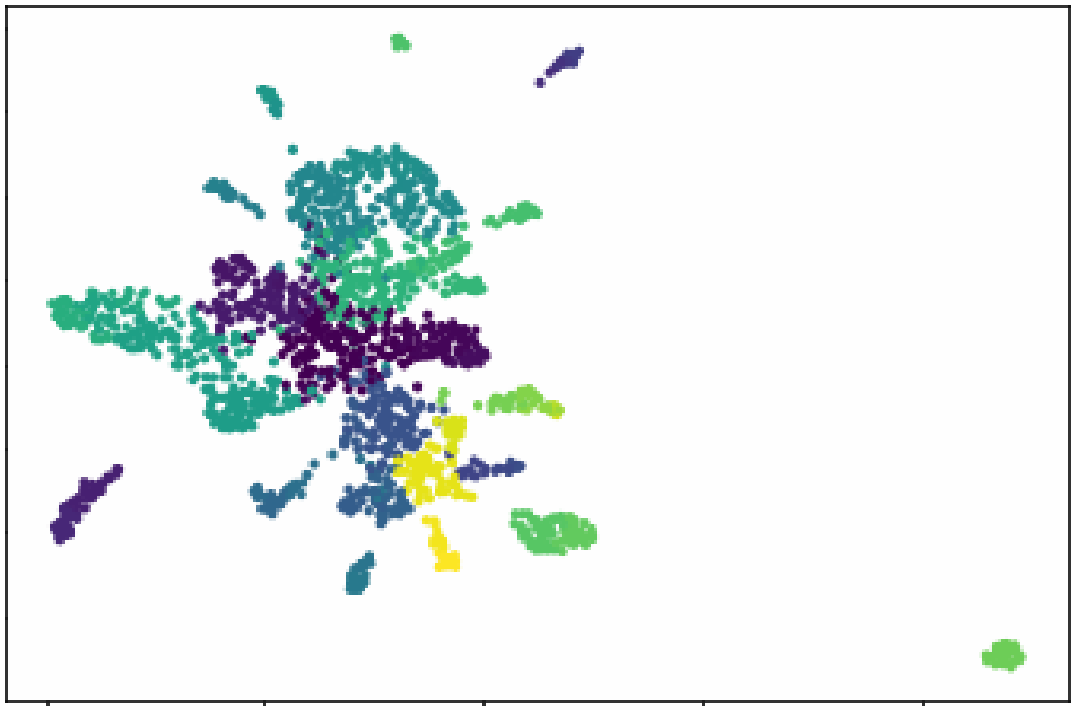


- Clustering difficile

Réduire la dimension

Enjeux

- Réduction drastique
- Préserver l'information
- Repérer la structure
- Diverses méthodes



SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Réduire la dimension

Principe

- Partir des TF ou TF-IDF
- Choisir P
- Nouvelles dimensions ($\sim$ *thèmes*)

Documents	dim 1 *(Ex : Santé)*	dim 2 *(Urgence)*	...	dim P *(BTP)*
...blabla Covid...	3.5	4.7	...	0.0
...blabla EDF ...	0.0	0.6	...	3.7

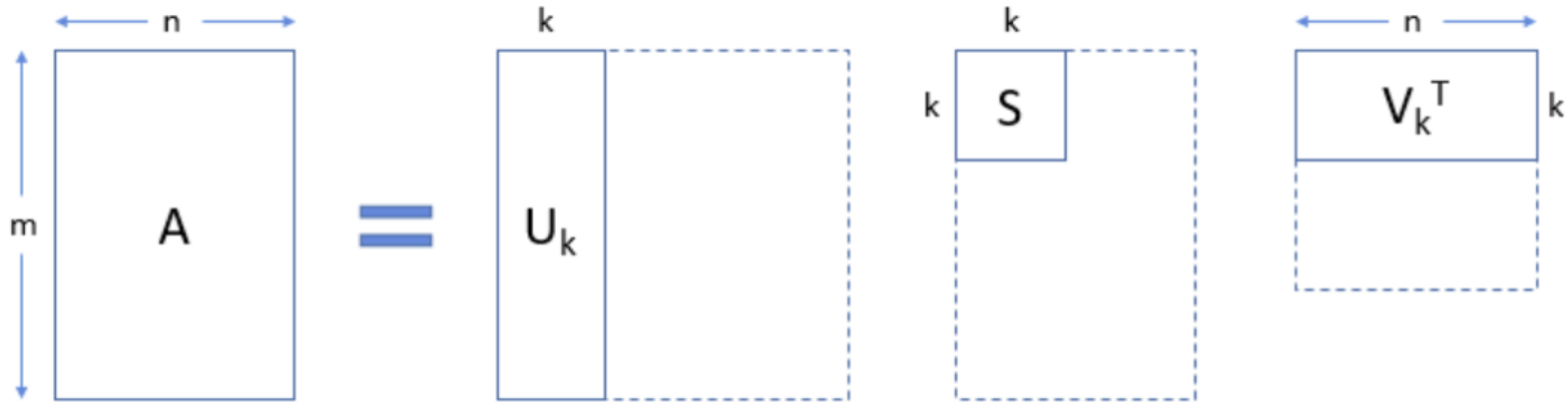
SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

Réduire la dimension

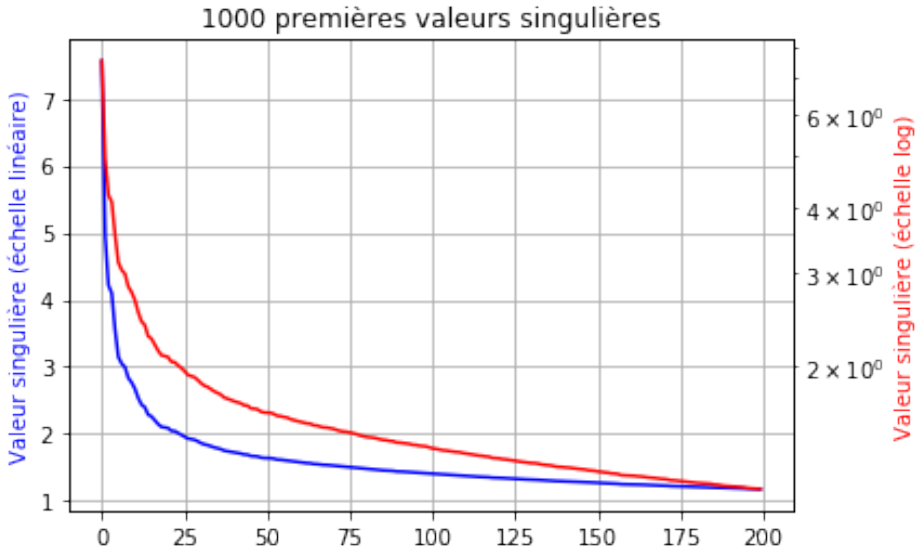
Algorithmes utilisés

- Factorisation du TFIDF (NMF, LSA)
- Distribution de probabilités des mots (LDA)

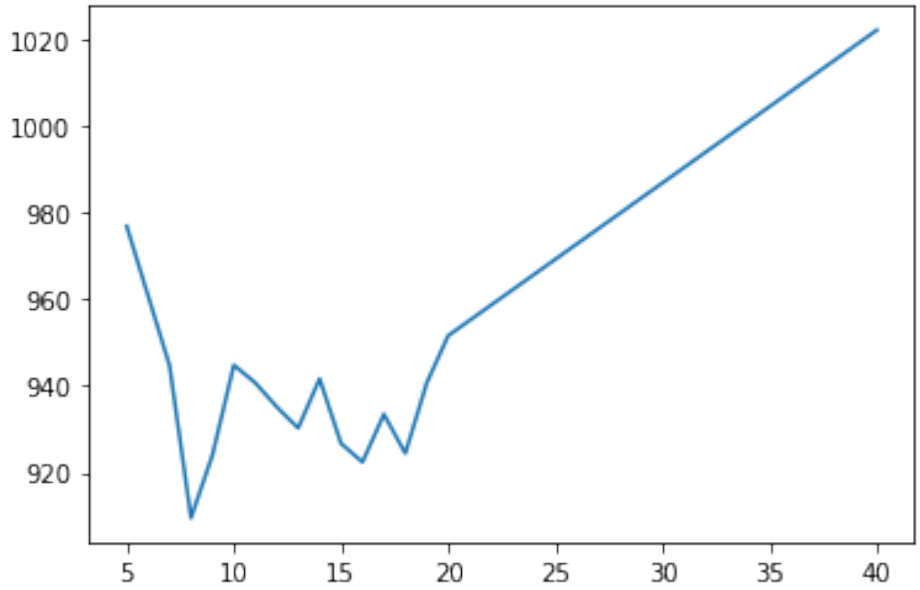


Réduire la dimension

Choix de la dimension



Critère de choix (LSA, NMF)



Mesure de pertinence (LDA)

Réduire la taille de dimension

Conclusion

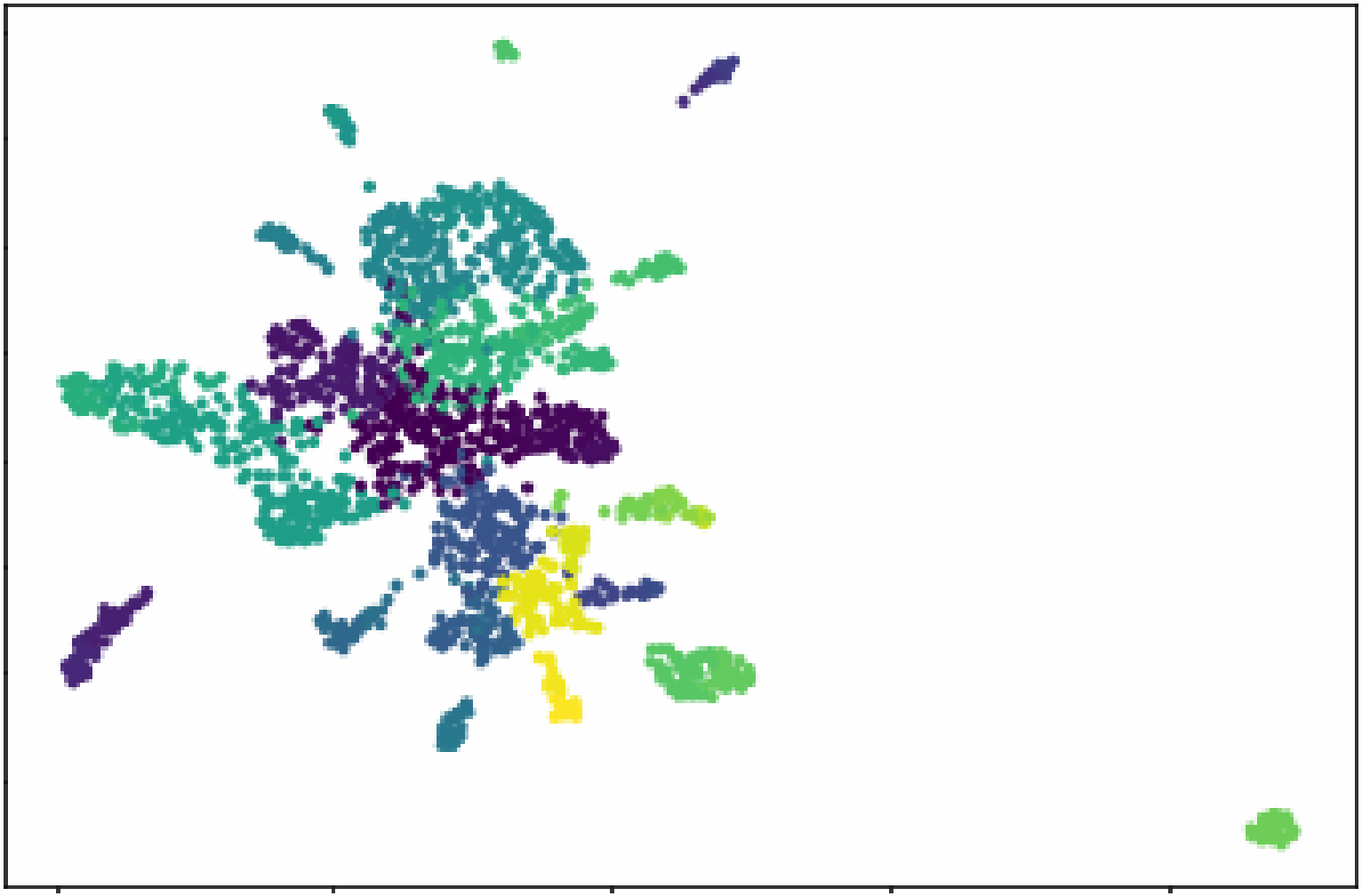
Vecteur	Topic 1	Topic 2	Topic 3
LSA	maladie, entreprise, santé, pouvoir, charge, mme, public, demande, prise, français	maladie, santé, lyme, prise, patient, charge, cancer, malade, atteindre, fibromyalgie	carte, gris, ant, usager, préfecture, service, titre, site, délai, sécuriser
LDA	tabac, cigarette, commune, logiciel, buraliste, tiers, prix, paquet, retrait, public	boulangerie, journée, administratif, paneterie, frontalier, tabac, effet, dispositif, gérant, hausse	art, métier, loi, roumanie, exercice, lme, fromager, oniam, examiner, lyme
NMF	000, pajemploi, pajot, pakistan, palais, palette, palier, palliatif, pallier, palmarès	charge, prise, patient, santé, atteindre, maladie, traitement, solidarité, autorité, soin	carte, gris, véhicule, problème, intérieur, demande, service, particulier, mois, agence

SORBONNE UNIVERSITÉ

POLYTECH SORBONNE

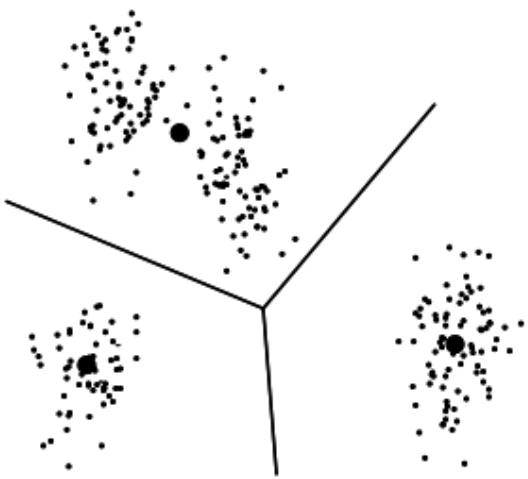
Clustering

Principe

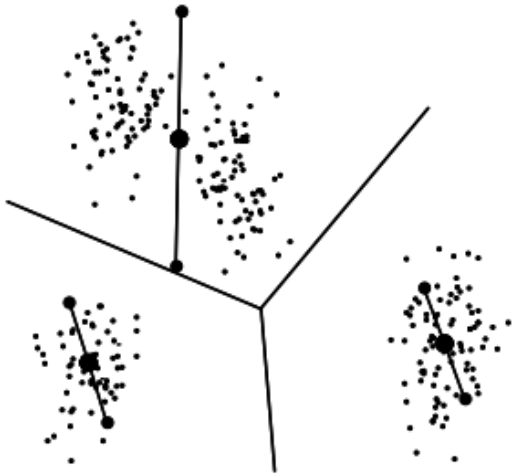


Clustering

K-means + X-means



K-means

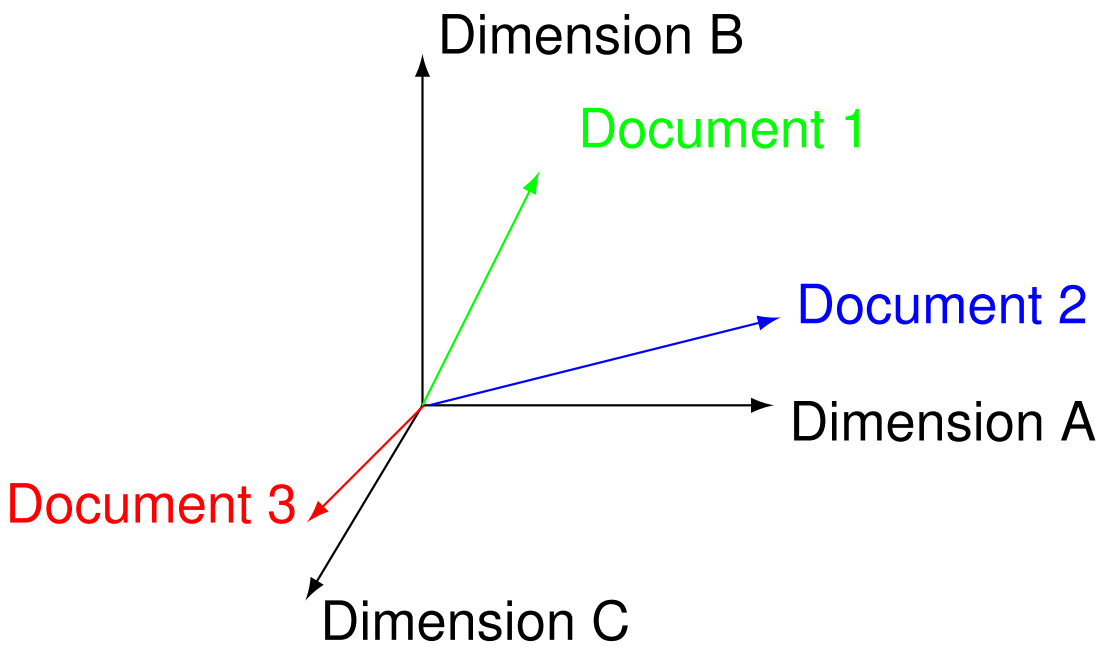


X-means step 1

X-means step 2

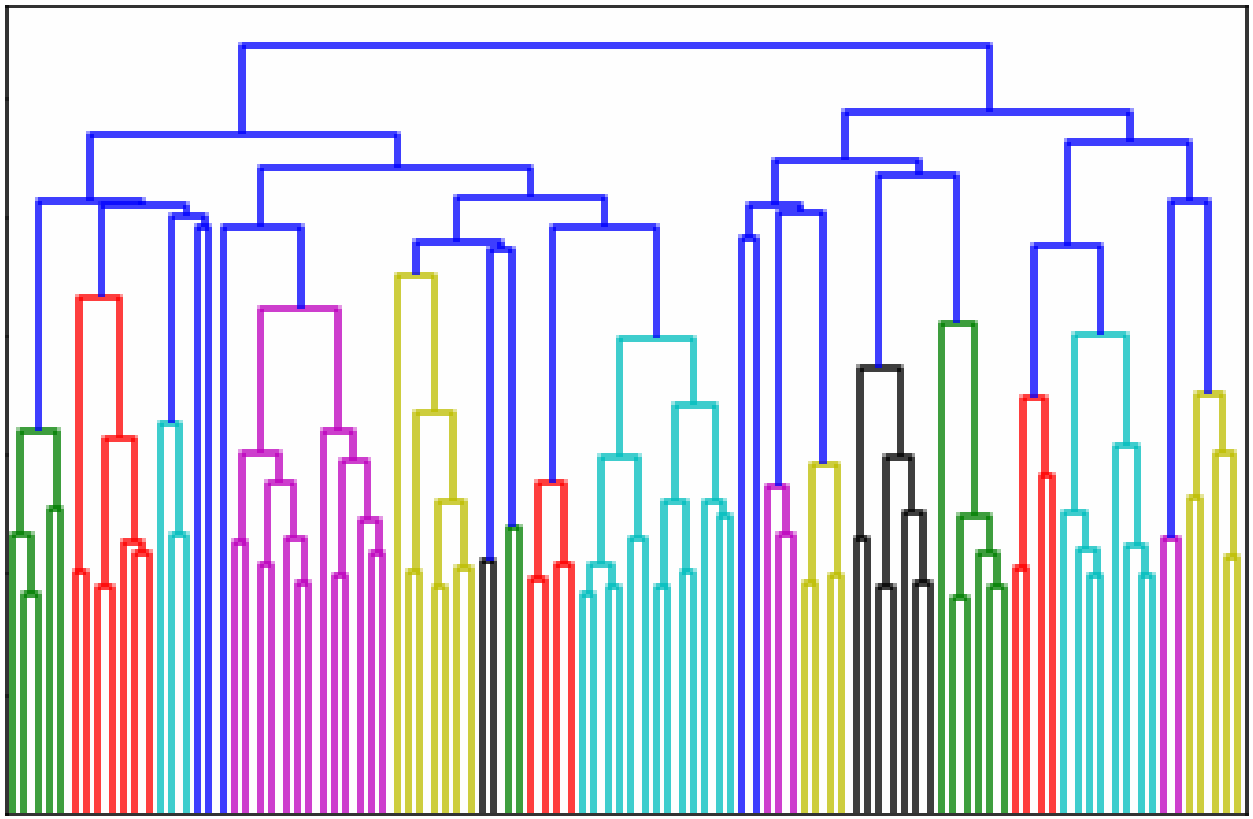
Distances

- Position ou direction ?



Clustering

Clustering Ascendant Hiérarchique (HAC)



Clustering

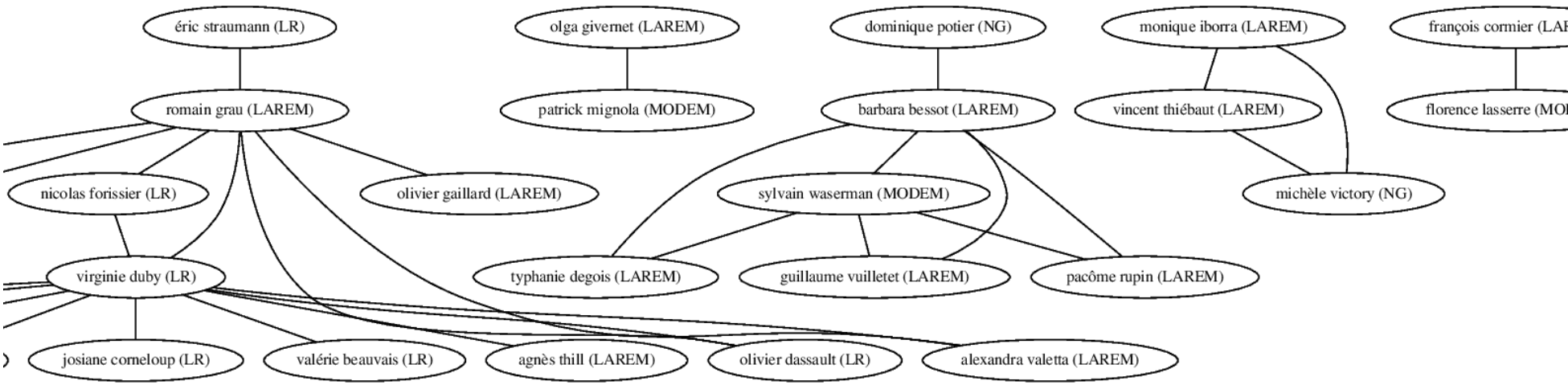
Conclusion

- LSA et HAC
- NMF et X-means

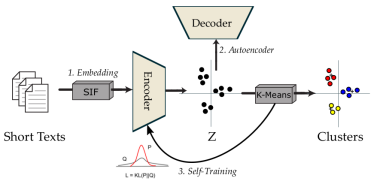
Cluster A
déficit école national administration
déficit compte ena
ant - dématérialisation - dysfonctionnement péritoine
ena - déficit - gestion
...

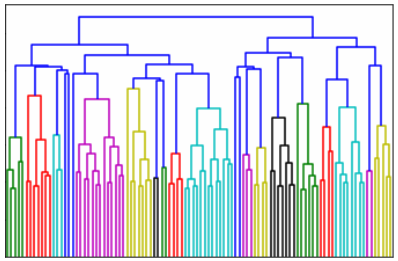
Lien entre déuptés

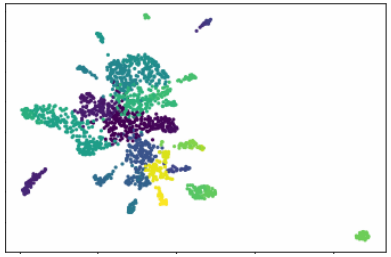
Premier graphe

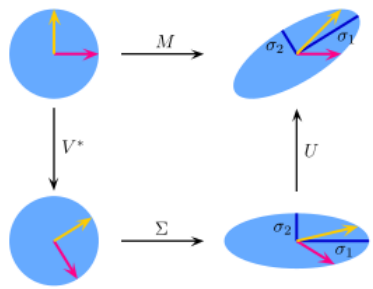


Conclusion

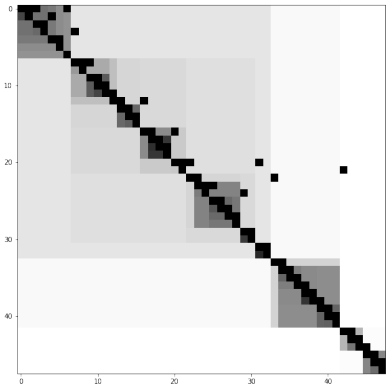


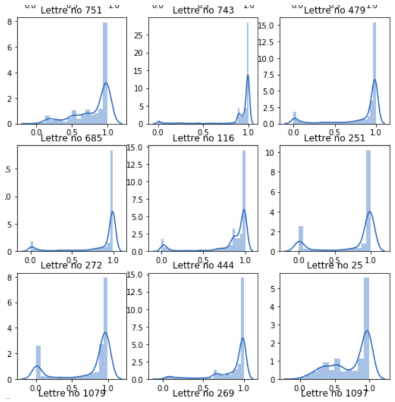


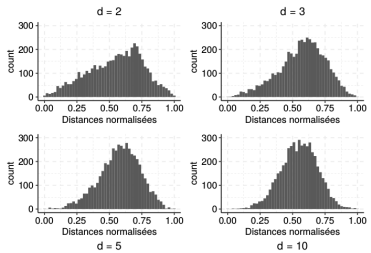


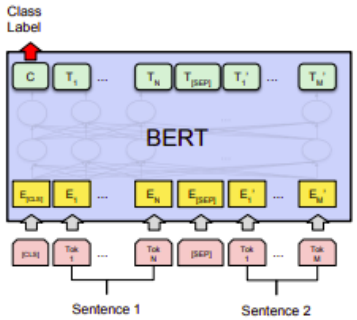


$$M = U \cdot \Sigma \cdot V^*$$









SORBONNE UNIVERSITÉ

POLYTECH SORBONNE