POSTER PROJET LU3IN026

Dans ce projet, nous avons étudié les performances de plusieurs algorithmes d'apprentissage automatique sur les données textuelles du forum 20 Newsgroups, qui contient des messages issus de 20 forums de discussion différents.
Nous avons appliqué à la fois des approches supervisées en approche supervisée et non-supervisées (clustering), en expérimentant différentes représentations vectorielles du texte : bag-of-words, fréquences, TF-IDF, avec ou sans suppression des stopwords.
L'objectif est d'appliquer et comparer différents algorithmes d'apprentissage supervisé et non-supervisé sur des données textuelles.
Étudier l'influence des représentations et des classifieurs sur la performance (accuracy, temps, robustesse).

ECHANTILLON

	messages	target
0	\n\nI am sure some bashers of Pens fans are pr...	10
1	My brother is in the market for a high-perform...	3
2	\n\n\n\n\tFinally you said what you dream abou...	17
3	\nThink!\n\nIt's the SCSI card doing the DMA t...	3
4	1) I have an old Jasmine drive which I cann...	4
...	...	...
18841	DN> From: nyeda@cnsvax.uwec.edu (David Nye)\nD...	13
18842	\nNot in isolated ground recepticles (usually ...	12
18843	I just installed a DX2-66 CPU in a clone mothe...	3
18844	\nWouldn't this require a hyper-sphere. In 3-...	1
18845	After a tip from Gary Crum (crum@fcom.cc.utah....	7

API

- Python
- NumPy
- Pandas
- Graphviz
- Matplotlib

SCIENCES SORBONNE UNIVERSITÉ

GUILLAUME DUPART
YANIS TOUTAIN
GROUPE 2
LU3IN026 - 2024 2025

PRÉTRAITEMENT

NETTOYAGE DU TEXTE

- Nettoyage → Met le texte en minuscule, remplace la ponctuation par un espace
- text2vect → Supprime les mots inutiles (conjonction, verbes trop communs ...).

REPRÉSENTATION VECTORIELLES

- Bag-of-words binaire (1 si au moins une fois dans le document, 0 sinon).
- Bag-of-words par comptage (compteur de mots par document)
- Fréquence relative (compteur mot / total)
- TF-IDF (l'importance d'un mot dans un document donné, tout en réduisant le poids des mots trop fréquents dans l'ensemble du corpus).

DONNÉES

- 18846 lignes
- Train (avec stop-words) : 173 lignes
- Test (avec stop-words) : 932 lignes
- Train (sans stop-words) : 173 lignes
- Test (sans stop-words) : 932 lignes

APPROCHE SUPERVISÉE

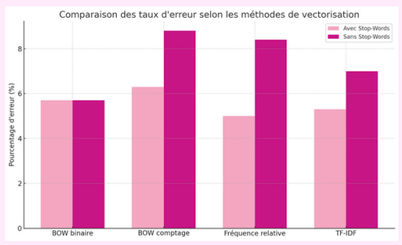
KNN DISTANCE EUCLIDIENNE

Avec BOW binaire
- Avec Stop-Words : 5.7% → 105 sec
- Sans Stop-Words : 5.7% → 84 sec

Avec BOW comptage
- Avec Stop-Words : 6.3% → 108 sec
- Sans Stop-Words : 8.8% → 79 sec

Avec Fréquence relative
- Avec Stop-Words : 5% → 74 sec
- Sans Stop-Words : 8.4% → 21.7 sec

Avec TF-IDF :
- Avec Stop-Words : 5.3% → 73 sec
- Sans Stop-Words : 7% → 22 sec



Nous remarquons que la distance euclidienne est inéfficace pour le pourcentage de bonne prédiction et le temps d'exécution.

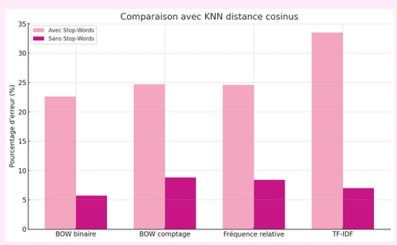
KNN DISTANCE COSINUS

Avec BOW binaire
- Avec Stop-Words : 22.6% → 5.14 sec
- Sans Stop-Words : 5.7% → 41.9 sec

Avec BOW comptage
- Avec Stop-Words : 24.7% → 5.63 sec
- Sans Stop-Words : 8.8% → 44.3 sec

Avec Fréquence relative
- Avec Stop-Words : 24.6% → 5.34 sec
- Sans Stop-Words : 8.4% → 23.31 sec

Avec TF-IDF :
- Avec Stop-Words : 33.5% → 5.11 sec
- Sans Stop-Words : 7% → 24.3 sec



Mieux que distance euclidienne, même si le pourcentage de bonne prédiction laisse à désirer. Amélioration nette du temps d'exécution et du pourcentage lorsque les mots inutiles ont été filtrés. Nous pouvons voir les mots inutiles comme une sorte de bruit qui peut altérer les résultats de prédiction.

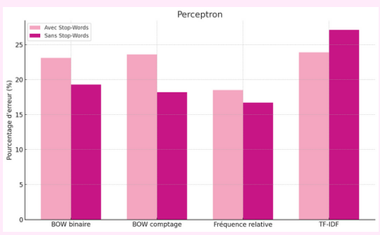
PERCEPTRON

Avec BOW binaire
- Avec Stop-Words : 23.1% → 0.05 sec
- Sans Stop-Words : 19.3% → 0.05 sec

Avec BOW comptage
- Avec Stop-Words : 23.6% → 0.05 sec
- Sans Stop-Words : 18.2% → 0.05 sec

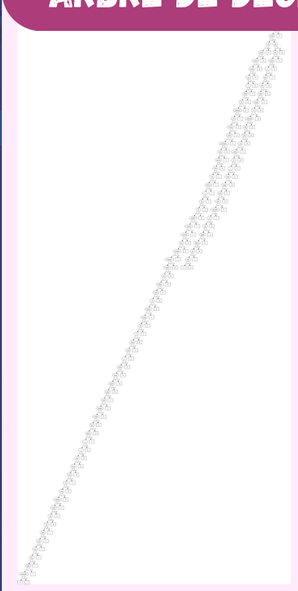
Avec Fréquence relative
- Avec Stop-Words : 18.5% → 0.02 sec
- Sans Stop-Words : 16.7% → 0.02 sec

Avec TF-IDF :
- Avec Stop-Words : 23.9% → 0.02 sec
- Sans Stop-Words : 27.1% → 0.02 sec

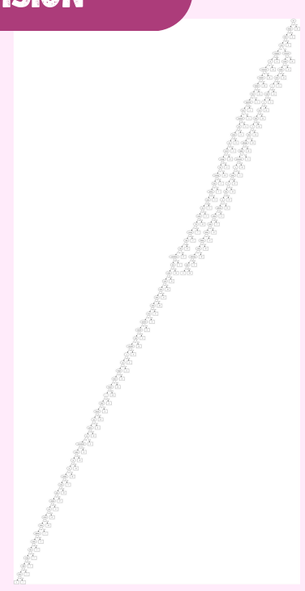


Avec ou sans stop words, le pourcentage de précision de prédiction stagne entre 16 et 24 %, le temps d'exécution est très efficace.

ARBRE DE DÉCISION

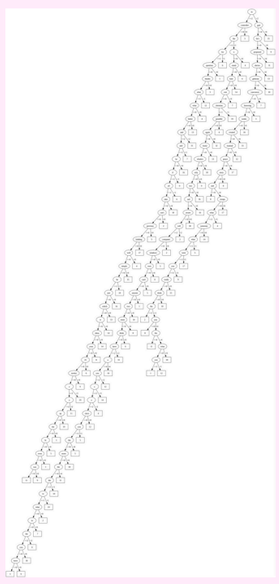


Arbre binaire avec sw

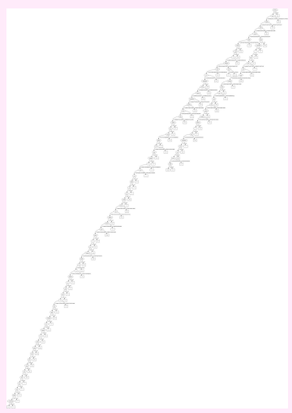


Arbre binaire sans sw

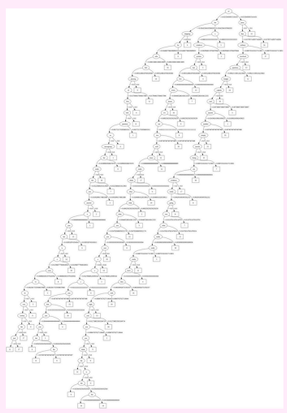
Arbre comptage avec sw



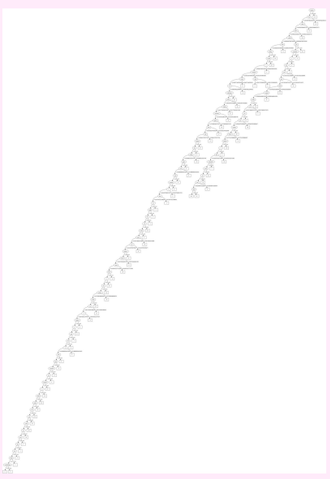
Arbre comptage sans sw



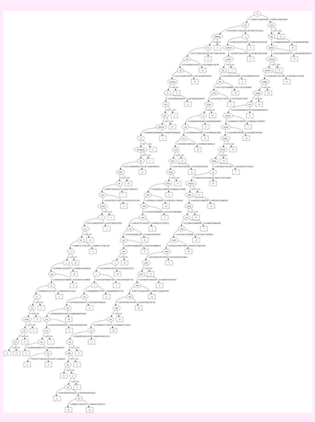
Arbre frequence avec sw



Arbre frequence sans sw



Arbre tf idf avec sw



Arbre tf idf sans sw

APPROCHE NON SUPERVISÉE

REMARQUE

Nous n'avons pas eu le temps de faire cette partie.

CONCLUSION

Nos résultats ne sont pas assez représentatif, car notre jeux de données est limité par la capacité de nos ordinateurs (PPTI compris).
Néanmoins, pour ce qui est de l'apprentissage supervisé, dans la plupart des cas, la méthode TF-IDF est la plus approprié pour le problème de classification de mots.
Nous remarquons que sans l'application des tris sur les stop-words, les performances sont biaisés par tous les mots inutiles, ce qui baissent les performances de bonne prédiction.
Concernant les classifieurs, le Perceptron de Rosenblatt est le classifieur ayant obtenu les meilleurs résultats.
Pour les arbres de décision, la représentation en fréquence et TF-IDF se découpent en branches plus homogènes, comparé aux autres qui s'enchainent.