Clustering et réduction de dimension avec les embeddings Word2Vec et GloVe

Florent GIAUNA

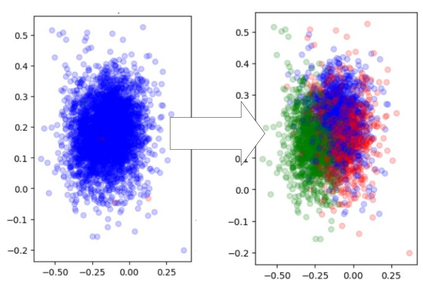
M1 AMSD/MLSD, Université Paris Cité - UFR Sciences Fondamentales et Biomédicale

Ce projet a pour objectif d'étudier les représentations textuelles Word2Vec et GloVe et de les utiliser pour comparer différentes méthodes de réduction de dimension et de clustering.

Clustering

Le clustering est une technique utilisée en apprentissage automatique (*machine learning*) pour **regrouper des données similaires** en fonction de leurs caractéristiques communes.

Image 1 : Exemple de clustering sur Classic3



- Objectif : trouver des similarités entre les données et les organiser en groupes distincts, c'est-à-dire des clusters.
- Application : segmentation de clients, classification d'images, détection d'anomalies, recommandation de produits, etc.
- Utilité : découvrir des structures cachées dans les données pour mieux les comprendre.

C'est un **processus non supervisé**, ce qui signifie qu'il n'y a pas de réponse préalablement connue. Les clusters sont déterminés uniquement en fonction des similarités entre les données.

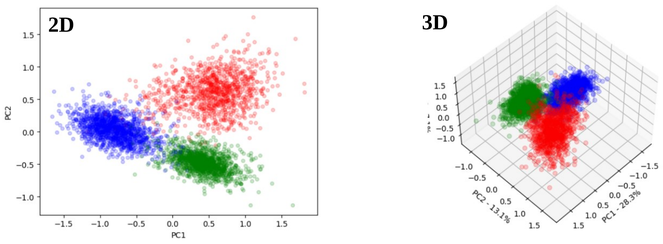
Réduction de dimension

La réduction de dimension est une technique utilisée pour **réduire le nombre de caractéristiques ou de variables** d'un ensemble de données.

- Objectif : simplifier les données tout en conservant autant d'informations pertinentes que possible.
- Utilité : obtenir une représentation plus concise des données, ce qui facilite leur compréhension et leur visualisation.

Réduire la dimension d'un ensemble de données à deux ou trois dimensions permet de le représenter graphiquement, ce qui **facilite la visualisation et la détection de structures ou de clusters**. Cela peut être particulièrement utile pour comprendre les relations entre les données.

Image 2 : Exemples de réduction de dimension (ACP sur Classic3 avec GloVe)



Objectif

Chaque méthode de clustering et chaque technique de réduction de dimension a ses propres avantages et inconvénients, et il est important de choisir celui qui convient le mieux pour chaque problème spécifique.

Le projet a pour objectif de réaliser une étude comparative de différentes méthodes. Pour réaliser cette comparaison le projet nécessite plusieurs parties :

1) **clustering dans l'espace d'origine** : le clustering est effectué sans réduction de dimension.

2) **approche tandem ou additive** : une opération de réduction de la dimension est effectuée avant de réaliser un clustering ;

2) **approche simultanée** : exploration de méthode permettant d'effectuer en même temps une réduction de dimension et un clustering.

- **Méthodes de réduction de dimension** : PCA, t-SNE, UMAP, Autoencodeurs
- **Méthodes de clustering** : Kmeans++, Kmedoids, spherical Kmeans, CAH avec différents critères d'agrégation.
- **Méthodes tandem ou simultanées**: Reduced k-Means[1], Deep k-Means[2].

Méthodologie

L'étude porte sur des **données textuelles**. Il s'agit de retrouver les sujets dont les documents traitent. Les vrais groupes sont connus pour pouvoir évaluer les différentes approches.

- Données : **Classic 3**, **Classic4** et **BBC**.
- Modèles d'apprentissage non supervisé : ils produisent des représentations vectorielles (*word embedding*) pour que des mots similaires soient plus proches les uns des autres.

  - **Word2Vec** : les vecteurs captent les relations entre les mots en fonction de leur proximité contextuelle.

  - **GloVe** : les vecteurs capturent les relations sémantiques et syntaxiques entre les mots en se basant sur la distribution de leur co-occurrence dans le texte.
- Représentations visuelles en 2D et en 3D des espaces pour comparer les clustering.
- Métriques pour évaluer le clustering.

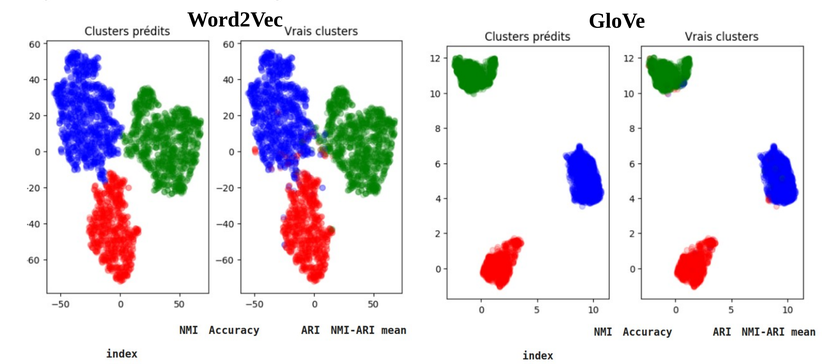
  - **NMI** (Normalized Mutual Information) : tient compte de la similarité entre les regroupements et de leur pureté.

  - **ARI** (Adjusted Rand Index) : mesure la similarité entre les paires d'échantillons et en ajustant les résultats en fonction du hasard.

  - **Accuracy** (Exactitude) : calcul le pourcentage de prédictions correctes.

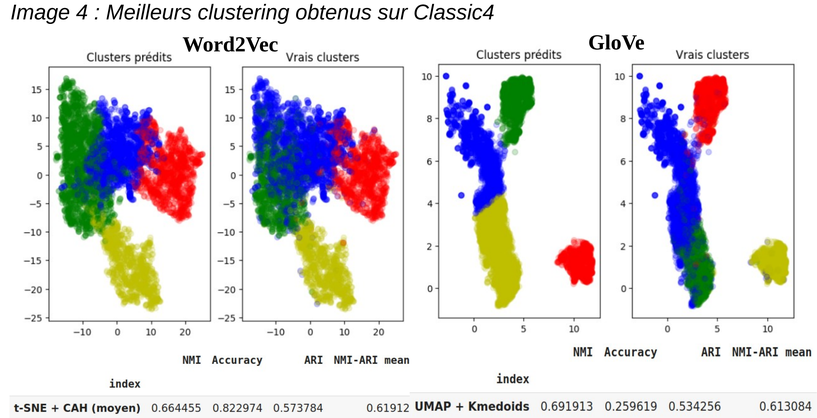
Résultats

Image 3 : Meilleurs clustering obtenus sur Classic3



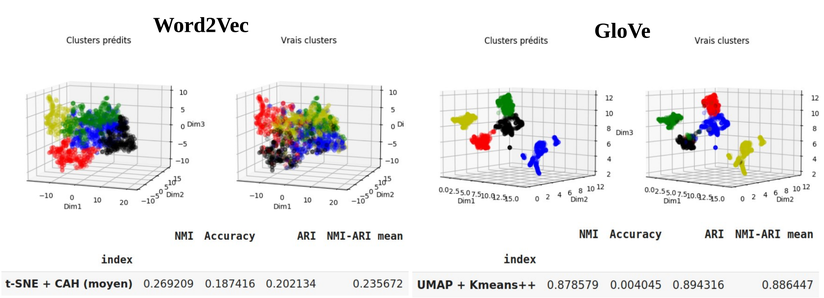
t-SNE + CAH (moyen)	0.898662	0.979183	0.940088	0.919375
UMAP + Kmeans++	0.932574	0.987407	0.962335	0.947455

Image 4 : Meilleurs clustering obtenus sur Classic4



t-SNE + CAH (moyen)	0.664455	0.822974	0.573784	0.61912
UMAP + Kmedoids	0.691913	0.259619	0.534256	0.613084

Image 5 : Meilleurs clustering obtenus sur BBC



t-SNE + CAH (moyen)	0.269209	0.187416	0.202134	0.235672
UMAP + Kmeans++	0.878579	0.004045	0.894316	0.886447

Conclusion

Sur les jeux de données :

- Classic3 : Word2Vec et GloVe permettent d'obtenir de très bons clusters.
- Classic4 : Word2Vec et GloVe trouvent des clusters médiocres.
- BBC : GloVe obtient de biens meilleurs résultats que Word2Vec.

Sur les *word embeddings* :

- Word2Vec : t-SNE performe mieux.
- GloVe : UMAP performe mieux.

Afin de mieux juger des méthodes simultanées Reduced k-Means et Deep k-Means plus d'hyperparamètres devraient être d'expérimentés.

Bibliographie

1. M. Yamamoto and H. Hwang. A general formulation of cluster analysis with dimension-reduction and subspace separation. Behaviormetrika, 41(1):115–129, 2014.
2. M. M. Fard, T. Thonet, and E. Gaussier. Deep k-means: Jointly clustering with k-means and learning representations. Pattern Recognition Letters, 138:185–192, 2020.