1. Jeux de données

Quatre jeux de données sont fournis, dont deux sont labellisés et deux sont dépourvus de labels. Les deux jeux de données labellisés (classic4 et BBC) doivent être analysés, afin que les constats émis par rapport à l'analyse comparative soient consolidés. Concernant les jeux de données non-labellisés (Articles_1 et Articles_2), le choix est laissé à l'étudiant quant au jeu de donnée à analyser. Il est bien entendu possible de fournir une étude pour les deux jeux de données non-labellisés.

- Classic4: Il contient 7095 documents catégorisés en 4 classes.
- BBC: Il contient 2225 articles de presse catégorisés en 5 classes.
- Articles_1: Il est composé du contenu textuel de 3985 articles de presse non-labellisés.
- Articles_2: Il contient un échantillon de 5089 documents non-labellisés.

2. Travaux à réaliser

Les jeux de données proposés sont utilisés pour étudier les représentations textuelles vues en cours, à savoir Word2vec, GloVe[1], BERT, RoBERTa, FastText (facultatif), ALBERT (facultatif). Ces données serviront de domaines d'application de méthodes vues en cours. Le projet comportera deux parties différentes : la première concerne principalement la réduction de la dimension à laquelle est ajoutée ensuite une tâche de clustering (approche tandem), la seconde se focalise sur l'obtention du clustering via une approche asimultanée combinanat les deux tâches simultanément. Ces deux parties conduiront à deux évaluations pour les deux UEs.

Partie 1: UE Réduction de la dimension

Une fois les représentations obtenues, il vous sera demandé de réaliser une étude comparatives de différentes méthodes (vues ou non en cours) de réduction de dimension (PCA, t-SNE, UMAP, Autoencodeurs, etc.) et de clustering (K-means, clustering spectral, HDBSCAN, CAH avec différents critères d'agragation, etc.) dans l'espace réduit et l'espace d'origine. Votre analyse doit être bien construite et doit notamment inclure :

- Une comparaison des différentes méthodes en termes de visualisation en se servant de différentes métriques existantes qui sont disponibles dans le package QVisVis et décrite dans [2].
- A l'aide des métriques accuracy, NMI et ARI, évaluer le clustering à partir de l'espace d'origine et l'espace réduit. Le nombre de classes doit être pris en compte dans les analyses.
- Pour les données labelisées ou pas, une interprétation des classes doit être réalisée.

Dans cette partie, on doit disposer de tableaux synthétiques, de visualisations en 2d ou 3d et des commentaires pertinents de chaque table et figure. A noter que le code de ces méthodes est disponible.

Partie 2: UE Apprentissage et factorisation matricielle

Dans cette partie et contrairement à la partie 1 (approche Tandem), il s'agit d'appliquer et d'évaluer des méthodes combinant simultanément les méthodes de la réduction de dimension et le clustering.

- Reduced k-means et Factorial k-means [3, 4]
- Deep Clustering Network (DCN) [5]
- Deep k-means (DKM) [1]

Comme dans Partie 1, on doit disposer de tableaux synthétiques, de visualisations en 2d ou 3d et des commentaires pertinents de chaque table et figure. En plus, des commentaires comparatifs de Partie 1 et Partie 2 seront nécessaires. A noter que le code de ces méthodes est disponible.

3. Rendus du projet en deux étapes

- @AMSD et @MLSD Le retour sur les données labelisées (Word2vec, Glove) du projet (Partie 1 + Partie 2) est programmé pour samedi 04 Décembre à minuit au plus tard 2021. Attention, une première notre sera attribuée à cette partie. Le reste sera également évalué, ci-après les dates de retours.
- @AMSD Le retour du reste est programmé pour Samedi 11 décembre à minuit.
- @MLSD Le retour du reste est programmé pour Samedi 18 décembre.

4. Envois des projets

Les envois sont à adresser (en spécifiant dans le sujet Partie1+Partie2) à
Mr Mohamed Nadif mohamed.nadif@u-paris.fr
Mme Mira Ait Saada mira.ait-saada@etu.u-paris.fr

References

[1] 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.

[2] Stephen L France and Ulas Akkucuk. A review, framework, and r toolkit for exploring, evaluating, and comparing visualization methods. Vis. Comput., 37(3):457–475, 2021.

[3] M. Vichi and H. Kiers. Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis, 37(1):49–64, 2001.

[4] M. Yamamoto and H. Hwang. A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika, 41(1):115–129, 2014.

[5] B. Yang, X. Fu, N. D. Sidiropoulos, and M. Hong. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In ICML, volume 70, pages 3861–3870. PMLR, 2017.

[1]Utiliser la version https://nlp.stanford.edu/data/glove.840B.300d.zip de GloVe en convertissant le modèle GloVe en un format word2vec avec la fonction : https://radimrehurek.com/gensim/scripts/glove2word2vec.html