Évaluation des plongements textuels des LLMs pour la classification non supervisée de documents

Imed Keraghel[1], Stanislas Morbieu[2], Mohamed Nadif[3]

[1]Centre Borelli UMR 9010 – imed.keraghel@u-paris.fr

[2]Kernix Software – smorbieu@kernix.com

[3]Centre Borelli UMR 9010 – mohamed.nadif@u-paris.fr

Abstract

Document clustering involves grouping a collection of unlabeled texts in a way that texts within the same cluster are more similar to each other than to those in different clusters. The effectiveness of this task relies on the representation of the documents. Although the Bag-of-Words (BoW) model was initially introduced, there have been significant advances in contextual representations, particularly with transformer architectures. This study aims to evaluate various textual representation techniques in an unsupervised setting, including BoW, TF-IDF, Word2Vec, GloVe, BERT, E5, JoSE, INSTRUCTOR, and GPT. The primary focus is to compare the textual embeddings of GPT with those generated by other methods. Additionally, due to the high dimensionality of these embeddings, an investigation will be conducted on how dimensionality reduction can impact clustering quality.

Keywords: LLM, GPT, clustering, dimensionality reduction.

Résumé

La classification non supervisée (*clustering*) joue un rôle clé dans le traitement de textes. Cette tâche repose essentiellement sur les plongements textuels extraits à partir de différents modèles. Alors que des techniques traditionnelles comme le sac de mots (*Bag of Words* - BoW) ont longtemps été privilégiées, les innovations récentes en Traitement Automatique du Langage Naturel (TALN), notamment les modèles basés sur l'architecture *Transformer* tels que GPT, offrent des plongements qui s'avèrent pertinents pour la classification. Cette étude vise à évaluer l'efficacité de tels plongements dans un cadre non supervisé, et à étudier l'impact des techniques de réduction de dimension sur la qualité de la classification. Nous comparons plusieurs représentations : BoW, TF-IDF, Word2Vec, GloVe, BERT, E5, JoSE, INSTRUCTOR et GPT.

Mots clés : LLM, GPT, classification non supervisée, réduction de la dimension.