**Plan de travail prévisionnel**

**Introduction :**

Dans le domaine du traitement du langage naturel, les méthodes d'incorporation de texte, ou text embeddings, sont essentielles pour convertir le texte en représentations numériques que les ordinateurs peuvent traiter efficacement. Ce projet vise à mettre en œuvre et à évaluer les performances du modèle text-embedding-ada-002 récemment publié par OpenAI, dans un contexte de classification de texte, et le comparer à une méthode de base (baseline) utilisant Tf-IDF couplé avec OneVsRest (OVR) pour la classification de questions sur Stack Overflow en fonction de leurs tags.

**Thématique choisie :**

Classification multi-label de texte (prédiction de tags pour les questions sur Stack Overflow)

**Méthode envisagée :**

**Méthode principale :** Text-embedding-ada-002, un modèle récent pour l’incorporation de texte présenté par OpenAI le 15 décembre 2022.

* Il est présenté comme surpassant les anciens modèles d’incorporation de texte en termes de recherche de texte, de recherche de code et, ce qui nous intéresse, de classification de texte.
* La longueur de contexte accrue permet de travailler plus facilement avec de longs documents, ce qui pourrait être utile lors de la manipulation de questions complexes de notre jeu de données.
* Ce modèle a une taille d'embedding plus petite, ce qui le rend plus rentable et peut accélérer l'entraînement et l'inférence. De plus, il est indiqué que le coût a été réduit de 90%.

**Méthode baseline :** Utilisation d’une approche TF-IDF combinée à un classificateur OneVsRest

* L'utilisation de Tf-IDF couplé avec OVR est une approche simple et couramment utilisée pour la classification de texte. Elle a fait ses preuves dans diverses applications et constitue une base solide pour la comparaison.

**Dataset :**

Le dataset utilisé contient des questions de Stack Overflow et nous devons prédire les tags associés à ces questions.

Ce dataset provient d’un précédent projet de prédiction de tags

* Comme mentionné, un projet précédent portait sur un dataset similaire, ce qui permet d'exploiter l'expérience existante et de se concentrer davantage sur l'évaluation du nouveau modèle.

**Métrique de performance :**

Score Jaccard

* Le score Jaccard est une métrique appropriée pour les tâches de classification multi-labels car il mesure l'intersection sur l'union des ensembles de labels prédits et réels. Cela est particulièrement utile pour évaluer comment le modèle performe dans la prédiction des tags de questions, où chaque question peut avoir plusieurs tags.

**Sources bibliographiques :**

OpenAI (15 Décembre 2022), "Présentation du modèle text-embedding-ada-002 qui remplace cinq modèles séparés pour la recherche de texte, la similarité de texte, et la recherche de code". [Lien vers l'article](https://openai.com/blog/text-embedding-ada-002/)

Documentation OpenAI, "Embeddings" - Ce document explique ce que sont les embeddings, comment les obtenir en utilisant l'API OpenAI et comment ils peuvent être utilisés pour différentes tâches, y compris la classification de texte. [Lien vers la documentation](https://platform.openai.com/docs/guides/embeddings)