OPENCLASSROOMS

Projet n°7

Plan prévisionnel

Étudiant :
Fayz El Razaz

Enseignant :
Amine Hadj-Youcef

30 septembre 2022

1 Thématique choisie

Nous nous sommes donnés pour ce projet la tâche d'étudier et d'approfondir le traitement du langage naturel (NLP). Nous souhaitons nous concentrer sur l'algorithme BERT et ses déclinaisons afin de l'utiliser au sein du projet n°5 du parcours Ingénieur machine learning.

2 Sources bibliographiques

Pour mener à bien notre projet, nous avons identifiés les sources bibliographiques suivantes :

- **Arxiv** : on y trouve notamment les articles scientifiques originaux.
- **Papers With Code** : site spécialisé en machine learning où l'on trouve les articles scientifiques accompagné du code utilisé pour illustrer l'article.
- **HuggingFace** : site qui recense un nombre important de modèles ainsi que de la documentation sur les techniques de NLP.

3 Prototype à implémenter

Nous avons choisis d'implémenter l'algorithme BERT et potentiellement certaines de ses déclinaisons (ALBERT, DistiBERT ou encore RoBERTa) pour effectuer de la classification multi-label de texte.

3.1 Dataset

Nous allons implémenter notre prototype sur le jeu de données du projet n°5 du parcours, à savoir le jeu de données de StackOverflow où l'on va chercher à tagger les questions des utilisateurs.

3.2 Méthode baseline

Nous avions retenu un modèle composé de deux modèles :

- Un SVM (qui prédit en général, un tagg correctement)
- Word2Vec (qui permet de donner les mots proches à celui prédit, et qui permet donc d'augmenter la quantité de taggs prédit, tout en gardant de la cohérence).

3.3 Méthode mise en oeuvre

Nous mettons en oeuvre dans le cadre de ce projet un modèle BERT et une de ses déclinaisons (à voir selon le temps disponible).