RAPPORT COMPETITION IFT3395

Titre du Projet :

Nom de l’équipe Kaggle

Nom des membres :

Introduction

Pour ce projet, vous participerez à une 2ème compétition Kaggle basée sur la classification de textes.  
Vous disposez de plus d'un million de textes annotés comme négatifs, positifs et neutres. La tâche consiste à créer des caractéristiques basées sur les données fournies et à utiliser des algorithmes d'apprentissage automatique pour les classer.

Vous devez implémenter et entraîner plusieurs algorithmes de classification basés sur l'ensemble des données fournies. L'évaluation sera basée sur les performances sur le jeu de test fourni et sur un rapport écrit.

Voyons comment nous pouvons mettre en œuvre la classification de textes avec Tensorflow https://www.tensorflow.org/ TensorFlow est une plateforme open source de bout en bout pour l'apprentissage automatique. Le jeu de données provient de la competition de Kaggle (https://www.kaggle.com/competitions/kaggle-competition-2/overview) Nous implémenterions la classification de texte en utilisant un simple réseau convolutif développé à l'aide de Tensorflow sur les données de tweet pour classer les tweets comme "positifs", "négatifs" ou "neutres".

Pré-requis :

* Python 3.5 ou supérieur
* scikit-learn (<https://github.com/scikit-learn/scikit-learn>)
* Pandas (<https://pandas.pydata.org/>)
* Tensorflow GPU 2.1
* Pydot

Features Design :

Nous commençons par définir des méthodes utiles au pre-process des textes :

Nous lisons les données d'entraînement et de test donné à partir de chaque fichier csv. Au début, nous lisons le fichier train.csv et test.csv ainsi que les données d’un label (train-result.csv) nécessaire à l’entrainement.