CentraleSupélec

OPENCLASSROOMS

# Recommandation de Tags

## Synthèse


### *Contexte :*

Dans le but d'aider les utilisateurs a bien tagger leurs questions sur le Site Stack Overflow, on nous demande de mettre en place un modèle supervisé et non supervisé permettant de fournir à l'utilisateur des tags censé pour sa question

### *Problème :*

Ce problème est un problème de classification multi-labels. A l'aide de différents modèles, divers tags vont être proposés. Les modèles seront évalués sur leur capacité à prédire le plus de tag en commun avec le post en question.

### *Données :*

Les données sont extraites d'une API publique de Stack Exchange. A partir de celle-ci, 50 000 questions avec un score supérieur à 3 ont été pris au hasard afin d'avoir un dataset considéré propre.

### *Approche :*

Après un nettoyage des questions, les matrices Term Frequency et Term Frequency-Inverse Document Frequency ont été mis en place. Pour le modèle non supervisé, le LDA sur la matrice TF a été mis en place. Par la suite, une approche de type KNN en utilisant la Similarité de Jensen Shannon a été mis en place pour prédire les tags les plus proches. Pour le modèle supervisé, plusieurs modèles ont été testés sur la matrice TF-IDF et un Fine-Tuning a été fait sur le modèle le plus performant.

### *Performances des modèles :*

Le modèle non supervisé avec le KNN customisé a été évalué manuellement sur 20 posts. La prédiction est assez mauvaise dans l'absolue mais environ 40% des tags sont en rapport avec la question. Quant au modèle supervisé, il y a quelques mots qui ont une tendance à fausser le modèle. Un fort overfitting existe sans régularisation. De ce fait un Grid Search est mis en place pour trouver de bons paramètres de régularisation. Au final, la précision sur le test set diminue mais reste très proche de la précision du train set.

### *Résultats :*

Sur ce projet, les résultats sont correctes avec la méthode supervisée mais moins avec la méthode non-supervisée. L'avantage du modèle non-supervisé est de proposer des tags "moins fréquents" et ainsi proposer des tags parfois difficile à trouver par soi-même. Diverses solutions sont proposées pour améliorer les modèles en passant d'un nettoyage plus poussé du dataset (au niveau des StopWords) à une architecture différente pour la classification avec 2 modèles.


MINÉ Nicolas