ALLAIN Lucas
REVALIER Quentin
2A31

Rapport – Challenge R

Pour mener à bien ce projet nous avons principalement été chercher sur le web (forum, site de cours de R …). Par la suite nous avons regroupé toutes les informations trouvées et nous les avons combinées afin de mieux classifier les documents.

Nous avons principalement utilisé la méthode KPPV (k plus proches voisins) durant ce challenge.

Méthodes utilisées:
Classifieur 1-plus-proche-voisin
Descripteurs = mots apparaissant au moins 400 fois dans le corpus
Pondération = fréquence du mot dans le document
Distance euclidienne

On a augmenté le nombre de mots dans le dictionnaire de mots supprimés afin d'améliorer le taux de réussite du classifieur.
Nous avons aussi utilisé des fichiers de sauvegardes (fichiers .rds) afin de stocker les mots qui apparaissent plus de 400 fois dans les documents ainsi que la matrice de classification.
Nous avons également remarqué que certains mots étaient inutiles, nous les avons regroupés dans motSuppr et nous les supprimons.

Librairies utilisées:
NLP
tm

Taux d'erreur calculée sur les données d'entraînement:
Taux d'erreur d'apprentissage pour le classifieur des 1-PPV sur les données d'entraînement = 0.2304762 soit 23.04762 %
Taux de réussite pour le classifieur des 1-PPV sur les données d'entraînement = 0.7695238 soit 76.95238 %