Master 1 MEDAS (CNAM, Nantes)

Fouille de Données 1

S. Quiniou et C. Marinica

Projet : analyse de tweets de la campagne présidentielle 2017

L'objectif de ce projet est d'analyser les tweets des candidats à l'élection présidentielle de 2017, en effectuant différents types d'analyses, à l'aide des différentes méthodes vues pendant le cours.

Le travail est à réaliser en binôme ; vous avez jusqu'à **27 mai 2020** pour nous faire parvenir les groupes par retour de mail. Un rapport d'une dizaine de pages, en anglais, au format pdf, sera à rendre pour le **3 juillet 2020 18h**. Il contiendra vos réponses aux exercices ci-dessous, en précisant le langage (R ou Python) et les modules utilisés pour réaliser chaque exercice ainsi que les résultats obtenus pour chaque exercice.

Il y aura également une soutenance de 20-25 minutes (10 minutes de soutenance + 10-15 minutes de questions) le **3 juillet 2020**, en binôme (planning à venir). Pour limiter les problèmes de connexion et de partage d'écran, vous devrez également envoyer votre support de présentation, en français, au format pdf, pour le **2 juillet 2020 12h**.

Exercice 1 : Récupération des données.

Les données sont à télécharger sur le site http://ideo2017.ensea.fr/outil-twitter/index.php . Vous devez choisir le corpus 'data_presidentielle2017', et puis récupérer tous les tweets avec toutes les caractéristiques. Attention, si vous faites un téléchargement, les caractéristiques ne sont pas gardées, donc privilégier le copier-coller du tableau affiché.

Exercice 2 : Analyse stylistique des candidats

L'analyse stylistique consiste à étudier le style d'un auteur, à l'aide de différents indicateurs. Dans le cadre des tweets, les indicateurs intéressants peuvent être le nombre de mots différents utilisés, les mots ou les lemmes les plus utilisés (par exemple, les 10 les plus fréquents), les catégories grammaticales les plus utilisées ou les pourcentages d'utilisation de chaque catégorie grammaticale… La fouille de motifs séquentielle peut également être utilisée pour calculer les motifs les plus fréquents de chaque candidat, au niveau des mots, des lemmes et/ou des catégories grammaticales.

Vous pourrez comparer les valeurs de ces indicateurs entre les candidats ainsi que par rapport aux valeurs globalement calculées sur l'ensemble du corpus de tweets.

Exercice 3 : Recherche des thématiques des tweets de chaque candidat

Vous utiliserez des méthodes de clustering pour identifier les principales thématiques abordées par chaque candidat ainsi que les thématiques globalement abordées dans l'ensemble du corpus de tweets.

Il vous faudra choisir le nombre de clusters le plus approprié. Vous pourrez également étiqueter chaque cluster à l'aide des mots les plus fréquents du cluster (pour décrire la thématique abordée dans le cluster).

Exercice 4 : Identification d'un candidat à partir de ses tweets

Vous utiliserez des méthodes supervisées pour construire un classifieur permettant d'attribuer un nouveau tweet au candidat qui est le plus susceptible de l'avoir écrit.

Pour cela, vous découperez le corpus global des tweets par candidat en un corpus de développement, un corpus de validation et un corpus de test (vous pourrez également utiliser de la validation croisée). Vous appliquerez des méthodes de classification supervisée et vous donnerez les résultats obtenus sur le corpus de test. Vous pourrez réutiliser les indicateurs identifiés à l'exercice 2, pur choisir les caractéristiques les plus pertinentes à utiliser dans vos classifieurs.