

Cours : Traitement Automatique du Langage Naturel

Promo : Master Polytech

Auteur : Amir HAZEM

Date : Novembre 2018

TP 3

Vers l'évaluation des approches de classification

L'objectif de ce troisième et dernier TP est de préparer vos systèmes à la phase d'évaluation. Ainsi vous aurez à conduire une étude comparative des différents classifieurs utilisés selon les traits choisis. Vous devez préparer trois systèmes pour chaque tâche. Chaque système correspondra à un classifieur avec un ensemble de traits.

À rendre :

1- un script contenant toutes les fonctions d'extraction de traits. Nommez ce fichier

extract_features.py

Ce script devra prendre en entrée un corpus (train.csv par exemple) et renvoyer en sortie un ensemble de traits. Au minimum *extract_features.py* devra produire les traits suivant :

- la longueur du tweet en nombre de caractères
- la longueur du tweet en nombre de mots
- la présence ou non d'urls
- la présence ou non d'émoticônes
- la présence ou non de hashtags (#)
- la présence ou non d'arobases (@)
- la présence ou non de ponctuations consécutives (!!!,??, etc.)
- la présence ou non de répétitions (Nooooon, Ahhhh, etc.)
- la présence ou non d'entités nommées
- le nombre de noms dans un tweet
- le nombre de verbes dans un tweet
- le nombre d'adjectifs dans un tweet

Tout autre trait que vous jugerez pertinent pourra être rajouté.

2 - Trois scripts *systemes_tacheX.py* (X correspond au numéro de la tâche)

Chaque script contiendra une fonction par système.

Un des 3 systèmes devra obligatoirement utiliser le classifieur Naive Bays.

Votre script *systemes.py* prendra en entrée le fichier de développement et renverra un score pour chaque système.

3 - Un rapport (maximum 3 pages en pdf) présentant vos 3 systèmes (système + traits choisis) ainsi que les résultats obtenus sur le corpus de développement.

Vous devez fournir pour chaque tâche et chaque système un tableau ou graphique illustrant le score d'exactitude en fonction des traits choisis.

Vous devez justifier le choix des traits ou à défaut donner l'intuition motivant vos choix.

Libre à vous d'adopter différentes stratégies selon le type de prétraitement, de filtrage et/ou d'utilisation de lexiques de sentiments.