Méthodes de la science de données
Projet :
Classification de documents d'opinions

Encadrement :
Fati Chen, Christophe Menichetti
Konstantin Todorov, Pascal Poncelet

Mars 2019

Le but de ce projet consiste à mettre en oeuvre et évaluer des méthodes de classification de documents d'opinions

Le corpus

Un jeu de données textuelles est mis à disposition sur Moodle. Il s'agit d'un corpus d'à peu près 8000 documents contenant des avis d'internautes sur des films. A chaque document est associé sa polarité selon l'avis (+1 : positif, -1 : négatif). Le fichier des documents est formaté dans un tableau cvs (un avis par ligne), un autre fichier csv contient les polarités d'avis par document (- 1/+1). Une correspondance directe existe entre les numéros des lignes des documents et des polarités.

Objectifs :
L'objectif du projet est de classer le mieux possible les avis. Pour cela vous pouvez utiliser différentes versions du corpus : textes bruts avec ou sans suppression de stop-words, textes lémmatisés, stématisés, utilisation d'un outil de part-of-speech-tagging (librairie NLTK, TreeTagger – http://www.cis.unimuenchen.de/~schmid/tools/TreeTagger/), utilisation de n-grams, etc.
Vous devrez rechercher des ressources externes pour traiter de l'ironie et aider à la classification. Il y a beaucoup d'ironie dans les avis.
Vous devrez utiliser différents classifieurs pour voir ceux qui donnent les meilleurs résultats, tester les éventuels hyperparamètres pour au final fournir un modèle qui soit apte à être utilisé avec des données de même type.

Etape 1 : Rechercher les meilleurs pré-traitements et les meilleurs classifieurs

Dans cette étape vous pourrez effectuer différents prétraitements et utiliser différents classifieurs afin de rechercher ceux qui sont les plus efficaces. Il ne faut pas hésiter à utiliser différents classifieurs car un classifieur qui s'avère efficace sur un jeu de données est peut être inefficace sur un autre jeu de données. Attention également à l'évaluation de vos modèles.

Etape 2 : Analyse

Une analyse complète de la qualité de la classification selon les différents types d'entrées et types de prétraitements par modèle de classification et paramétrage doit être proposée. Autrement dit, les combinaisons différentes de modèle + paramètres + pondération + type de données d'entrée donneront des performances différentes. A vous de les comparer et configurer votre fonction de classification pour qu'elle soit le plus performante possible sur les données de test en proposant une analyse approfondie de vos résultats.

Remarque 1 : Le thème de la classification des textes laisse penser que certains types de mots peuvent se révéler particulièrement discriminants (par exemple, les adjectifs pour la classification d'opinion). Une discussion sur l'influence de tels marqueurs morphosyntaxiques sera bienvenue.

Remarque 2 : Différents traitements (par exemple, pondérations, algorithmes de fouille de données comme l'extraction des règles d'association) ont été proposés par les encadrants du projet. Vous pourrez vous en inspirer pour présenter des résultats complémentaires aux résultats de classification.

Remarque 3 : Attention à la négation et à l'ironie/sarcasme.

Etape 3 : Challenge

Un challenge sera organisé qui aura pour but de tester vos meilleurs classifiers sur des données de tests non-vues. Pour cela vous devrez sauvegarder vos modèles en utilisant pickle. Attention au fait que vos pré-traitements doivent être mis dans le pipeline pour les nouvelles données.
Les résultats seront évalués par l'équipe d'encadrants et un classement des groupes en résultera. Vous aurez les résultats du challenge 2 semaines avant la soutenance : vos rapports devraient intégrer une analyse de vos résultats sur le challenge et les éventuelles améliorations que vous avez pu effectuer sur cette base

Organisation

- Le travail s'effectuera en groupes de 2 à 5 étudiants (limite ferme).
- Une soutenance orale de 15 minutes suivie de 10 minutes de questions est prévue à la fin du semestre. La soutenance a pour objectif de présenter vos approches, vos choix et de mettre en avant également l'analyse des résultats que vous avez obtenu. Lors de la présentation vous présenterez également les résultats obtenus dans le challenge et discuterez des résultats (meilleurs, moins bons, pourquoi, etc). Il est inutile de perdre du temps lors de la présentation sur les données initiales (qui sont communes) ni sur la problématique du projet ou bien la théorie des méthodes utilisées.
- Le rendu final sera soumis sous la forme d'un fichier compressé (gzip) identifié par le numéro du groupe **à déposer sur Moodle au plus tard 3 jours avant la soutenance** consiste en :
  1. Un rapport de **max 10 pages**
  2. Le notebook au pdf et ipynb de vos codes. format Les codes de l'ensemble des traitements automatiques
- Attention à bien mettre le prénom, nom et numéro d'étudiant de chaque personne du groupe dans les documents rendus.