Extraction de connaissances à partir de données

HMIN208

2017

Projet: **Classification de documents par opinion**

Encadrement : Dino Ienco, Konstantin Todorov, Pascal Poncelet

Le but de ce projet consiste à mettre en oeuvre et évaluer des méthodes de classification de documents par opinion.

Le corpus

Un jeu de données textuelles vous est mis à disposition sur Moodle. Il s'agit d'un corpus de 5000 documents par classe contenant des avis d'internautes sur des films. A chaque document est associé sa polarité selon l'avis (+1 : positif et -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.

Etape 1 : Transformation des données et vectorisation

WEKA prend en entrée des fichiers du type .arff. Le but ici est de transformer vos données de départ en fichiers .arff, compatibles avec Weka. Les programmes pourront être développés en Perl, Python, PHP, Java ou autres. Par la suite, les valeurs textuelles doivent être rendues numériques en utilisant une pondération fréquentielle (tf-idf, tf, ou autres). Cela peut se faire à l'aide de la fonction stringToWordVector, un filtre de WEKA. Pensez à la normalisation de vos vecteurs.

Etape 2 : Prétraitements des documents

Vous utiliserez les différents types de données d'entrée selon les prétraitements. Le but est d"évaluer vos modéles avec différentes répresentations pour choisir la représentation qui vous semble plus adapté à la classificaiton des documents, par example:

(1) Textes bruts (avec ou sans suppression de stop-words),
(2) Textes lemmatisés,
(3) Textes lemmatisés avec analyse morphosyntaxique (à l'aide de l'outil Tree-tagger vu en cours).
(4) ...

Pensez aussi à la possibilité d'appliquer des prétraitements personnalisés selon vos besoins et votre corpus (e.g., liste de stop-words personnalisée). Avec l'outil Tree-tagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) vous pouvez ajouter à chaque mot sa catégorie grammaticale et enrichir l'espace des descripteurs et ainsi comprendre si cette information peut aider (ou non) à classer votre corpus. Attention au format d'entrée utilisé par Tree-tagger. Vous pouvez également vous intéresser à d'autres types de connaissances linguistiques (par exemple, la terminologie, la sémantique, etc.).

Etape 3 : Mise en oeuvre d'algorithmes de classification

La suite du travail consistera à utiliser Weka et à évaluer rigoureusement les résultats de classification obtenus en prenant en entrée les différents corpus préparés dans l'étape précédent.

Rappelons que de nombreuses approches d'apprentissage peuvent alors être utilisées pour la classification de textes :

• K plus proches voisins,
• Arbres de décisions,
• Naïve Bayes,
• Machines à support de vecteurs

Paramétrage : Pour chaque méthode de classification, ils existent plusieurs paramètres à choisir, tels que le paramètre K de l'algorithme des KPPV, le noyau pour les SVM, le support pour les règles, *etc*.

Etape 4 : 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 teste 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 : est-ce qu'une opinion contenant le mot “génial” est forcément positive...? Comment traiter ce problème ?

Etape 5 : Challenge

Un seconde jeu de données (jeu de données de test) contenant des avis uniquement (sans polarités) vous sera fourni peu avant l'examen. Il sera formaté de la même façon que les données que vous avez utilisé précédement. Le but de ce challenge est d'évaluer le meilleur classifieur (avec ses paramètres, descripteurs et prétraitements associés) que vous avez construit auparavant sur ces nouvelles données. Une comparaison des résultats obtenus par les différents groupes sera effectué. Les trois meilleurs groupes pourront bénéficier de points complémentaires dans l'évaluation du TP. Attention, l'objectif de ce challenge n'est pas de modifier vos classifieurs dans la mesure où les résultats seront ceux qui vous seront rendus. Par contre il est important d'analyser les résultats obtenus. Pour cette étape, chaque groupe chargera sur le site MOODLE un fichier .csv avec autant de lignes que la taille du jeu de données de test et chaque ligne doit contenir la prédiction faite par le classifieur (+1 : positif et -1 : négatif). Le fichier doit être nommé avec l’identifiant du groupe (exemple GROUPEA.csv)

Organisation:

– Le travail s'effectuera en groupes de 3 à 4 étudiants (limite ferme).
– Une soutenance orale de 10 minutes suivie de 10 minutes de question 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.

Devoir à rendre:
Un fichier .ZIP identifié par le groupe (exemple GROUPEA.ZIP). Tous les autres formats sont refusés. Ce fichier doit contenir :

- 1 document pdf identifié par l'identifiant du groupe. Ce document doit être écrit en utilisant le format Latex fourni. Il ne doit pas dépasser 10 pages. Attention il n'est pas nécessaire de parler des données dans la mesure où tout le monde possède les mêmes données. Les numéro de carte d'étudiants, prénom et nom des membres du groupe doivent explicitement apparaître en entête.

1 répertoire contenant :
- Les sources des programmes (en entête de chaque programme doivent apparaître les numéro de carte d'étudiants, prénom et nom des membres du groupe)
- Les ressources supplémentaires que vous avez utilisées si c'est le cas

Echeance du projet:

- 02/03 : Mis à disposition sur Moodle des données d'entraînement
- 27/04 : Mis à disposition sur Moodle des données de TEST
- 01/05 à 18h : Les étudiants chargent sur Moodle leur classification
- 03/05 : Mis à disposition sur Moodle des labels de TEST
- 15/05 à 12h : Remise des rapports sur Moodle :
- 18/05 : Soutenance