Projet « RSS-Intelligence »

Sous projet « Classifier »

October 2020

Etape 1 : Constitution du corpus d'apprentissage et de test

A l'aide du sous-projet RSS-Collector, collecter sur des flux RSS appropriés des documents relevant des 5 classes suivantes et dans les langues anglaise et française :

- ART_CULTURE
- ECONOMIE
- POLITIQUE
- SANTE_MEDECINE
- SCIENCE
- SPORT

Taille du corpus : a minima une 100 aine de documents (donnée RSS + donnée url source) par classe et par langue.

Etape 2 : Prétraitement des données

Détecter la langue, extraire le texte des données, supprimer les balises, et autres « scories », puis segmenter le texte en mots que l'on « racinisera » en utilisant l'implémentation d'un « stemmer » pour chacune des langues considérées. On utilisera pour cela la bibliothèque Snowball https://pypi.python.org/pypi/snowballstemmer
Par l'utilisation de listes de mots-vides (stoplist), on éliminera des dictionnaires les mots outils (stop words) a priori peu discriminants. Exploiter les listes de stop words proposées sur https://pypi.python.org/pypi/stop-words.

Etape 3 : constitution des dictionnaires et vectorisation des documents

On pourra pour calculer les vecteurs *creux* le module *countVectorizer,* en utilisant ou non la pondération tf-idf, de la bibliothèque de *machine learning* scikitLearn :

http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Les vecteurs creux devront être sauvegardés sur disque et rechargeables en mémoire RAM pour éviter de les recalculer à chaque lancement du programme.

Etape 4 : Implémentation de quelques classifieurs de type k-NN, Régression Logistique, Bayésien Naif, SVM, Neural Network, Random Forest, etc.

Implémenter un classifieur de chaque type pour chacune des langues.

Effectuer une validation croisée pour sélectionner les méta paramètres et effectuer l'évaluation des classifieurs. On présentera les métriques usuelles (taux d'erreur/accuracy, précision/rappel, courbes ROC et AUC) en utilisant les micro et macro moyennes.

Etape 5 : Exploitation

Pour les nouvelles données collectées, on ajoutera dans la structure item-RSS un champ de classe_prédite (sport, international, économie, etc.) proposé par le meilleur classifieur en ajoutant la valeur de similarité/probabilité fournie par le classifieur.

Etape 6 : Documentation et code

- Fournir spécification, conception, code, installation, tests et le cas échéant les conditions d'interopérabilité avec le sous-projet « RSS-Classifier».