Classification de documents

Encadrement : H. Alatrista Salas, E. Kergosien

Le but de ce projet consiste à mettre en oeuvre et évaluer une méthode de **classification de documents par thème ou opinion**. Les programmes développés (vectorisation des données textuelles) pourront être développés en Perl, Python, PHP, Java ou autres. Les documents sont a priori au format texte mais la proposition d'autres types de documents peut être proposée et discutée avec les encadrants du projet pour validation.

Première étape : constitution du corpus

Dans un premier temps, un corpus devra être constitué. Nous proposons d'acquérir un corpus véhiculant un thème ou une opinion. Deux à cinq catégories seront alors proposées. Par exemple, pour la classification d'opinion (à partir de corpus de critiques de films, restaurants ou autres), trois catégories pourraient être identifiées : positif, négatif et neutre. Ces catégories seront attribuées au regard des notes attribuées par les utilisateurs.
Pour ce faire, vous devrez rechercher au moins 15 à 20 textes écrits en français ou en anglais relatifs à chaque catégorie.
Ce corpus devra être normalisé (suppression des balises HTML, etc).

Deuxième étape : mise en oeuvre d'un algorithme de classification

La seconde étape consistera à représenter les données textuelles sous forme vectorielle (approche dite de Salton) afin d'appliquer les algorithmes de fouille de données. La suite du travail consistera à utiliser Weka et évaluer **rigoureusement** les résultats de classification. 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,
- Réseaux de neurones,
- Machines à support de vecteurs.

Troisième étape : prise en compte d'informations linguistiques

Le but ici est d'utiliser vos textes avec différentes informations :

- Textes bruts,
- Textes lemmatisés,
- Textes lemmatisés avec analyse syntaxique.

Pour obtenir de telles connaissances, vous pouvez utiliser l'analyseur syntaxique **Sygmart** (pour les textes en français) également vu en cours. Une analyse

complète de la qualité de la classification selon les différents cas pourra être proposée.

Les étudiants pourront également s'intéresser à d'autres types de connaissances linguistiques (par exemple, la terminologie), sémantiques, etc. Dans ce projet, différents critères peuvent aussi être étudiés (paramètre K de l'algorithme des KPPV), élagage, normalisation du type tf*idf, etc. Bien entendu, tous ces critères ne pourront être étudiés dans le cadre de ce projet. **Il est donc préférable que chaque groupe étudie des aspects précis en y apportant une évaluation rigoureuse et une analyse approfondie.**

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 morpho-syntaxiques 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 seuls résultats de classification.*