Classification de documents

Le but de ce projet consiste à mettre en oeuvre et évaluer une méthode de **classification de documents par thème et opinion**. Les programmes développés pourront être développés en Perl, Python, PHP, Java ou autres. Les documents sont au format json.

```
{
        "id":"1",
        "commentaires":"Genesis",
        "catégorie":["cat1","cat2","catn"],
        "polarité":"0.8"
}
```

Commentaires :

Du texte libre, des commentaires, des tweets, des articles

Catégorie :

Entre 1 et n catégorie

Polarité :

0 négatif
0.5 neutre
1 positif

Première étape : constitution du corpus

Dans un premier temps, un corpus devra être constitué. Nous proposons d'acquérir un corpus véhiculant un thème et une opinion. Trois à cinq catégories seront alors proposées pour les thèmes et l'opinion sera évaluée au travers d'une note (0 très négative, 1 très positif).
Pour ce faire, vous devrez rechercher au moins 15 à 20 textes écrits en français ou en anglais relatifs à chaque catégorie et y évaluer l'opinion véhiculée.

Seconde étape : Préparation des données et création du fichier Arrf

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.

- Choisir et justifier un type de descripteurs: (terme, ngramme de mot ou de caractères …).
- Proposer et appliquer un prétraitement si nécessaire sur les textes (lemmatisation, suppression des mots creux …).
- Proposer une pondération (Booléen, TF, Occurrence, TF*IDF…).

Générer un fichier Arrf compatible avec Weka.

Troisième étape : Mise en oeuvre d'un algorithme de classification pour la ou les catégories

La suite du travail consistera à utiliser Weka et évaluer **rigoureusement** les résultats de classification. Rappelons que de nombreuses approches d'apprentissage pour la classification de textes ont été vues en Tp.

Quatrième étape : Étude comparative

Choisir au choix un autre type de descripteurs, une autre pondération ou un autre type de prétraitement et analyser l'impact du changement sur les résultats de classification.

Cinquième étape : Mise en oeuvre d'un algorithme de régression pour détecter l'opinion