PROJET Fouille de données Master IC (2009-2010)

Le but de ce projet est de classifier des corpus politiques issus du Web.

Le projet s'organisera autour des différentes étapes suivantes :
1. acquisition du corpus,
2. normalisation/nettoyage du corpus,
3. "vectorisation" des données textuelles,
4. application des algorithmes de classification de Weka,
5. restitution et analyse des résultats.

La première étape du projet consiste donc à construire un corpus de 10 textes par catégorie. Nous proposons de travailler à partir de **cinq (ou éventuellement trois) catégories politiques** : Extrême Gauche, Gauche, Centre, Droite, Extrême Droite. Les textes à acquérir seront sur la base de pages web de partis politiques ou de candidats aux dernières élections présidentielles.

Le but du projet est d'appliquer un protocole expérimental rigoureux afin d'évaluer la qualité des méthodes de classification en utilisant le logiciel Weka.

L'étude devra notamment répondre aux questions suivantes :
- *Les classes politiques sont-elles automatiquement séparables ?*
- *Quelles classes sont plus faciles/difficiles à discriminer ?*
- *Quels sont les descripteurs (mots) discriminants ?*
- *Quels filtres linguistiques peuvent être ajoutés pour améliorer les résultats ?*
- *etc.*