Acquisition de connaissances à partir de données

Catégorisation de documents textuels

L'objectif de ce projet est
- d'une part, de mettre en œuvre et d'évaluer les performances d'un algorithme de catégorisation de textes à l'aide de représentations « standards » des documents,
- d'autre part,
  - soit de comparer ses performances à celles d'un second algorithme de catégorisation de nature différente,
  - soit d'enrichir les représentations grâce à des connaissances linguistiques pré-existantes ou acquises sur les données textuelles et d'évaluer l'impact de ces informations linguistiques sur les performances de l'algorithme implanté.

Pour implémenter et tester les algorithmes de catégorisation, vous trouverez une collection de documents classables en différentes catégories (135 sujets) et la vérité-terrain associée sous http ://www.daviddlewis.com/resources/testcollections/reuters21578/ (également accessibles sous Moodle dans le répertoire Data du projet). Par ailleurs, voici quelques ressources ou outils TAL (liste totalement non exhaustive) qui pourraient vous servir au cours du projet, en particulier – mais pas uniquement – si vous choisissez la seconde extension :
- existence de segmenteurs et autres *tokenizers* ; listes de mots vides dans diverses langues également disponibles (*e.g.*, sous http ://torvald.aksis.uib.no/corpora/1999-1/0042.html, liste de Jean Véronis mais il en existe d'autres) ;
- *stemmers* ou analyseurs morphologiques : racineurs de Lovins, Porter ou Paice-Huster ; Flemm (analyseur morphologique du Français ; site de F. Namer) ;
- étiqueteurs morpho-syntaxiques, avec ou sans lemmatisation : TreeTagger, Brill... ;
- synonymes ou unités lexicales en relations paradigmatiques : WordNet (univ. Princeton ou version Java sur source.net/projects/jwordnet) ; the Roget's thesaurus, dictionnaire de synonymes du GREYC de Caen... ;
- acquisition de relations sémantiques paradigmatiques sur corpus par des méthodes d'apprentissage non supervisé ; calcul de relations sémantiques sur corpus par PLI... ;
- extraction de termes complexes : sur le Français ou l'Anglais, Acabit de B. Daille LINA Nantes, Ana de C. Enguehard LINA Nantes, Lexter de D. Bourigault ERSS Toulouse et une version à rôle beaucoup plus étendu Syntex.

Ce que vous avez à réaliser consiste donc en 3 parties.

1- Mise en forme des données : mettre en forme la collection de documents pour les tests (lire le fichier README très explicite).

2- Implémentation et évaluation d'un algorithme de catégorisation : dans cette première phase, on vous demande d'implémenter, sur le jeu de données mentionné ci-dessus, un algorithme de catégorisation de textes de votre choix. Les documents sont représentés à l'aide de termes simples. Vous devrez également mettre en place une procédure soignée d'évaluation des performances de cet algorithme.

3- Choix entre tests sur un second algorithme de catégorisation et comparaison des performances avec le premier sur le même jeu de données, ou **intégration de connaissances linguistiques dans votre représentation des documents et comparaison des performances** de votre algorithme initial quand les représentations intègrent ou pas ces connaissances. On peut par exemple s'intéresser à l'utilisation de *stemmers* (si pas utilisés lors de la première phase), de termes complexes, de proximités sémantiques entre mots..., connaissances linguistiques que vous pouvez soit obtenir à l'aide d'outils existants, à l'aide de bases lexicales disponibles, ou en les apprenant sur vos données.