**Projet de recherche documentaire**

Le but du projet est de vous faire comprendre les techniques fondamentales mises en œuvre dans les systèmes de recherche documentaire en réalisant la maquette d’un tel système et en l’expérimentant sur le corpus de documents fourni.

Le système que vous proposerez devra

* Pouvoir effectuer les recherches correspondant au corpus de questions fourni à partir du corpus de documents donné. Vous pouvez bien entendu élargir le champ de fonctionnement de votre système à partir d’autres corpus existant sur le Web.
* Concevoir ou utiliser une procédure de stemming.
* Définir ou utiliser une stop list. (Y figurent les mots dont l’inverse document frequency est très bas)
* Autoriser des formes plus ou moins complexes pour les requêtes allant des formes les plus simples (c.a.d.requêtes composées d’une union ou d’une conjonction de termes simples) aux plus sophistiquées (c.a.d. les *phrasal queries* requêtes composées d’une suite de mots devant apparaître de façon consécutive dans les documents, prise en compte de formules logiques plus ou moins complexes, présence du caractères jocker \* dans les termes de la requête, recherche tolérante permettant d’accepter des fautes d’orthographe)
* Disposer d’une interface conviviale permettant à l’utilisateur de fournir sa requête et d’accéder facilement aux résultats (par exemple en surlignant dans les documents les termes de la requête)
* Vous pourrez de plus envisager des extensions à votre système :
* En élargissant le type de requêtes que votre système accepte
* Fournir les réponses du système en ordonnant les documents résultats par ordre décroissant de pertinence
* Permettre l’utilisation de différentes mesures de similarité entre documents et requêtes
* …

**Ce que vous devrez fournir**

**Un rapport** :

Il ne doit pas excéder 10 pages qui décrivent le fonctionnement du système : l’indexation, le traitement des requêtes, et le cas échéant, la façon dont le système évalue la pertinence des documents.

**Présentation orale :**

Vous devrez prévoir une présentation orale d’une durée de 10 minutes au cours de laquelle chacun des participants au projet devra intervenir.

Vous y présenterez rapidement le modèle choisi (booléen, vectoriel), la façon dont il a été implémenté,le type des requêtes acceptées par votre système et la façon dont le système traite chacun des types, etc.

Une démonstration mettant en avant les points forts de votre réalisation devra accompagner la présentation.

**Quelques liens susceptibles de vous être utiles :**

<http://tartarus.org/~martin/PorterStemmer/> (Stemmer)

<http://www.web-mining.fr/methodes/stop-words> (Stopwords lists)

[http://www.algolist.net/Data\_structures/Dictionary\_%28ADT%29](http://www.algolist.net/Data_structures/Dictionary_(ADT)) (Dictionary data structures)

<http://www.gutenberg.org/catalog/> (corpora)

<http://www.umiacs.umd.edu/~jimmylin/downloads/brill-javadoc/edu/mit/csail/brill/BrillTagger.html> (Brill’s tagger)