Description du projet

MC AP8 - RI et Web

Antoine.Doucet@iutc3.unicaen.fr

Semestre 4, printemps 2014

Déroulement :

- Le travail se fait en groupe de 3-4 étudiants.
- Chaque groupe présentera son travail lors de la dernière séance (semaine du 24 au 28 mars), lors d'une présentation de 15-20 minutes.
- Le rapport devra être envoyé par email, en un seul fichier PDF, ou une page html (ce qui vous permet d'inclure les documents de votre collection), **avant le 28 mars à 12h.**

Tâches :

- Chaque groupe choisit un thème qui lui est propre.
- Chaque membre récupère 15 documents portant sur ce sujet, par exemple sur Internet.
- Chaque membre définit deux tâches (2 besoins d'information).
- Les documents sont indexés à l'aide du moteur de recherche Open Source « Lucene ». Il est alors possible de soumettre des requêtes à cette collection de documents.
- Chaque groupe liste toutes ses tâches et en sélectionne 2 avec l'avis de l'enseignant. Pour chacune des ces 2 tâches, vous préparerez 2 requêtes:
    1. Une expression Booléenne (le résultat sera l'ensemble des documents satisfaisant l'expression, sans classement).
    2. Une liste de termes (le résultat sera une liste de documents ordonnés suivant leur pertinence supposée).
- Pour chaque groupe, pour chaque besoin d'information au moins 2 juges évaluent la pertinence de chaque document.
- Calculez la mesures d'accord entre juges (mesure de Kappa).
- Exécutez les requêtes avec Lucene.
- Calculez le rappel et la précision pour tous les résultats
- Quand le résultat est une liste ordonnée, dessinez les courbes de rappel-précision. Utilisez la moyenne sur toutes les tâches.

Rapport :

Le rapport doit décrire votre travail. Il doit notamment inclure les éléments suivants :

- Description de la collection de documents; nombre de documents, sujet, nombre total de mots et nombre par document (en moyenne).
- Les tâches et les requêtes.
- Retour sur expérience des jugements de pertinence. Les juges étaient-il d'accord ? Pourquoi ? Donnez des exemples.
- Présentation des résultats de Lucene (rappel, précision, graphes, …).
- Différence d'intérêt entres les résultats des requêtes Booléennes et des requêtes pondérées (modèle d'espace vectoriel).
- Avec le recul, quelles requêtes auriez-vous tapées afin d'optimiser rappel et précision ?