Trabajo Práctico No. 1

September 19, 2008

1 Introducción

El trabajo es individual. Se permite trabajar en grupos de 2 personas. Los grupos de 2 personas deben considerar al menos 2 sistemas de RI, y justificar en detalle la decisión por la que eligen el sistema con el que trabajaran. También deben realizar más de un opcional a elección.
El sistema de trabajo a utilizar es de elección libre. Hay una lista de sistemas en la página de la cátedra. Uno de los más conocidos es el sistema Lucene (Java). Otro digno de mencionar es WIRE (http://www.cwr.cl/projects/WIRE/) del grupo de Baeza-Yates en Chile, escrito en C/C++.

La tarea es explorar las distintas posibilidades del sistema elegido y tratar de mejorar su performance. El objetivo principal del trabajo es familiarizarse con un sistema de RI y las técnicas que afectan la precisión del sistema.

Se recomienda utilizar un sistema que permita indexar diferentes campos en un documento (título, autor, etc.). Se recomiendo utilizar un sistema en un lenguaje que se domine. Hay sistemas en Java en C, y sistemas con interfaces para otros lenguajes.

2 Corpus

Corpus: Usamos una parte del corpus OHSUMED, una collección de abstracts de articulos médicos. Información sobre el corpus se encuentra en http://trec.nist.gov/data/t9 filtering/README. Usaremos el conjunto de datos de entrenamiento (training-set) que se encuentra en http://trec.nist.gov/data/t9_filtering.html, en el archivo ohsu-trec.tar.gz
El archivo del corpus propio **ohsumed.87**
(disponible en (http://www-2.dc.uba.ar/materias/riwm/ohsumed.87.gz)
Archivo de *queries*: **query.ohsu.1-63**
(disponible en (http://www-2.dc.uba.ar/materias/riwm/query.ohsu.1-63)
consiste de un conjunto de 63 *queries* para este corpus. Cada query contiene un número (OHSU1-OHSU63), un título y una descripción. Utilizar el título y la descripción como términos de la *query*.
El archivo de evaluación **qrels.ohsu.batch.87**, contiene juicios de relevancia y sus valores (uno por línea).

(disponible en (http://www-2.dc.uba.ar/materias/riwm/qrels.ohsu.batch.87)

3 Pasos preliminares

Usar la función de indexar con el corpus. Asegurarse que la indexación se ejecute sobre los distintos campos (titulo,keywords, contenido, MESH keywords), por cada documento, si es necesario, modificarla para que lo haga. Usar la función que ejecuta las queries. Verificar que forma tiene el output. Verificar si el sistema contiene *stopwords*. Probar usarlos

4 Requisitos

Qué es lo que tiene que hacer:

- a) Calcular Precisión, Recall, Precisión-R del sistema en el corpus
- b) En este ejercicio se espera que explore las distintas posibilidades del sistema que usa y que verifique como afectan la precisión-R del sistema.
- c) Verificar cual es la "scoring function" y si el sistema permite usar funciones alternativas, usarlas. Crear también una función nueva que no se encuentre entre las alternativas de acuerdo a las que se encuentran en el manual. (parametros a modificar son tf, idf,normalización de longitud).
- d) Verificar el analizador de texto (parser) y si es posible mejorarlo. Verificar estrategias para Mayúsculas/minúsculas (case folding), stemming. Si no está disponible agregar el Porter stemmer. Si tiene lista de stop-words verificar cual es la lista. Ver si encuentra otra lista de stop-words más adecuada.

  Importante: el analizador de texto debe funcionar igual para la query y los documentos.

Opcionales:

- i) Calcular precision-k. Elija 2 valores de k y justifiquelos.
- ii) Calcular *F-measure*.
- iii) Modifique los pesos asignados a distintos campos (título, keywords, etc.)
- iv) Incorporar un stemmer para español, probar con un corpus de español.

5 Entrega:

Debe presentar un escrito. Longitud máxima 2 paginas. Sin contar tablas. Debe fundamentar la elección del sistema. En los resultados informe cómo se afectó precisión y analice los distintos resultados que obtuvo.

Debe presentar todo código que ha creado, y todo código que ha modificado. El mismo debe estar debidamente comentado.

Fecha de entrega máxima: 22 de Octubre. Tanto el informe como el código deben ser enviados a jcastano@dc.uba.ar