Organización de Datos (75.06)
Segundo Cuatrimestre de 2006
Curso 1
Trabajo Práctico
Search Engine

El objetivo de este TP es la construcción de un Search Engine para páginas web, basado en el método del coseno para la resolución de consultas ranqueadas.
El TP deberá contar con 2 funcionalidades mínimas: el módulo indexador y el módulo de consultas. El módulo indexador deberá indexar un conjunto de páginas HTML almacenadas en un directorio y construir las estructuras necesarias para resolver las consultas.
Una vez indexados los datos solo se realizarán consultas sobre los mismos, sin modificar los datos.
El módulo de consultas resolverá consultas ranqueadas respetando la siguiente sintaxis

q: t1 t2 t3
q: +t1 t2 (t1 debe aparecer en los documentos relevantes)
q: -t1 t2 t3 (t1 no debe aparecer)
q: t1^2 t2 t3 (el peso de t1 es el doble de t2 y t3 para la consulta)

La interfase del módulo indexador es por linea de comandos de la forma:
index directorio
Ej: index /tps/grupoxx/html

La interfase del módulo de consultas será de tipo web, con la salvedad de que el programa debe esta hecho en C o C++ puede usarse cualquier tecnología para la interfase comunicandose de alguna forma con el módulo de consultas en C.

Evaluacion: Este TP será evaluado en función de los siguientes parametros:
- Tiempo que se demora en indexar el conjunto de datos
- Cantidad de espacio en disco que se usan para estructuras auxiliares en función del tamaño de los datos
- Calidad de respuesta del módulo de consultas: relevancia de los resultados, falsos positivos, falsos negativos etc
- Tiempo de respuesta del módulo de consultas

Bibliografía a consultar:
- Apunte de Indices de la cátedra
- Managing Gigabytes (Witte, Bell, etc)
- Understanding Search Engines (Berry, Browne)
- Material sobre search engines en general