Práctica 2: Motor de recuperación y Evaluación

Creación de un Motor de Recuperación Web

Prerrequisitos de la práctica

- Programación (la práctica se podrá realizar en el lenguaje que el alumno desee)
- Conocimiento de Bases de Datos

Objetivo de la Práctica

- Desarrollar un Motor de Recuperación de Información a partir del desarrollo de la práctica Implementación básica del Modelo Vectorial. El motor de recuperación deberá ser capaz de indizar y recuperar documentos Web a partir de una colección dada. Ante consultas realizadas en lenguaje natural el sistema deberá dar como salida una lista ordenada de documentos relevantes de entre los previamente indizados, así como su similitud con la consulta.
- Partiendo de un código con un modelo vectorial implementado para procesar documentos, incorporar las clases que proporcionen la funcionalidad relativa al Motor de recuperación no incluidas en la implementación inicial

Recursos

- Se suministra la colección **EIREX 2010**, que se deberá utilizar para **realizar pruebas y ajustar** el motor a desarrollar. Para ello se suministran los documentos, las necesidades de información (topics) y un conjunto de juicios de relevancia para evaluar.
- El fichero union.trel incluye los juicios de relevancia que relacionan los documentos con las necesidades de información.
- Los documentos se encuentran en el directorio documents.biased, organizados en directorios.
- Las necesidades de información y los niveles de relevancia se incluyen en el fichero topics.xml.

Módulo de evaluación

Objetivos

- Una vez desarrollado el motor de recuperación, deberán añadirse las clases que soporten la funcionalidad relativa a su evaluación mediante el cálculo de métricas.
- La salida de la aplicación deberá mostrar los valores de las métricas de evaluación descritas a continuación, haciendo uso de los juicios de relevancia incluidos en la colección de documentos descrita anteriormente.

Medidas a Incluir en las Presentaciones y Memoria

La aplicación deberá proporcionar como salida los valores de las siguientes métricas de efectividad:

- Precisión en cortes 5, 10 (relevancia mínima 1).
- Exhaustividad en cortes 5, 10 (relevancia mínima 1).
- Valor-F basado en los puntos de corte anteriores.
- Reciprocal Rank (relevancia mínima 1).
- Reciprocal Rank (relevancia mínima 2).
- Average Precision@100 (relevancia mínima 1).
- nDCG en cortes 10, 100 en base 2.
- Se debe reportar, por cada medida y corte, la media para los topics de la colección, usando siempre 4 decimales.

Sobre el Ground Truth

Como se ha visto en el curso un groundtruth es un fichero realizado manualmente que nos muestra cuales son los documentos relevantes para cada consulta, realizado a mano. Lo puedes encontrar en la colección de datos, el fichero para esta colección se denomina 2010.union.trel, esto es una muestra

2010.union 2010-001 2010-56-062 1

2010.union 2010-001 2010-67-004 1

2010.union 2010-001 2010-00-072 1

2010.union 2010-001 2010-13-080 1

2010.union 2010-001 2010-26-075 0

2010.union 2010-001 2010-38-057 1

2010.union 2010-001 2010-00-094 1

La segunda columna representa el topic, la tercera el documento y la cuarta el grado de relevancia. En esta lista puede haber documentos no relevantes como el 2010-26-075, debeis ignorar los documentos no juzgados. No hay orden en los resultados, es decir el 2010-00-094 no es menos relevante que el 2010-38-057, ya que ambos tienen un uno.

Al ejecutar vuestro programa obtendréis unos resultados similares

1 2010-001 2010-56-062 10.00025

2 2010-001 2010-67-082 0.00019

3 2010-001 2010-26-075 0.000017

4 2010-001 2010-38-057 0.000016

Donde la última columna podría ser el peso del coseno con tf-idf, indicando asi el grado de relevancia, en este caso si hay orden el segundo es menos relevante que el primero.

Seguramente vuestra lista será mucho más larga, podéis poner un umbral de corte si lo veis interesante, por ejemplo que por debajo de 0.000014 no se considera relevante. Eso mejorará vuestras estadísticas si bien reduce algo el reuso con otras colecciones.

Confrontando ambas listas podéis obtener los valores de precision, recall, etc. aunque se pueda hacer a mano no resulta muy lógico, por lo que se recomienda que el cálculo de las métricas esté programado.