uc3m | Universidad Carlos III de Madrid

Práctica 2: Motor de recuperación con Reconocimiento de Entidades y Expansión de Consultas

Ampliación del Motor de Recuperación Web

Objetivo de la Práctica

- Reutilizando la implementación del Motor de Recuperación de Información desarrollado en la entrega anterior, deberán incluirse las clases que amplíen la funcionalidad de motor mediante la implementación de uno de los dos módulos descritos a continuación.
- Deberá desarrollarse un módulo base con la funcionalidad de un Motor de Recuperación. El módulo base será **ampliado con uno de los siguientes módulos**: el Módulo de Expansión de Consultas (módulo I) y el Módulo de Reconocimiento de Entidades (módulo II). El grupo deberá seleccionar uno de los dos módulos de ampliación para su implementación. El desarrollo de ambos módulos por iniciativa del grupo será valorado positivamente.

Módulo Base: Recuperación Básica

Objetivo

Desarrollar un Motor de Recuperación de Información. El motor de recuperación deberá ser capaz de indizar documentos Web incluidos en la colección EIREX de la entrega anterior. 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. En la implementación se utilizará el modelo vectorial y la función coseno para el cálculo de la similitud.

Módulo I: Expansión de Consultas

Objetivo

- Realizar expansión de consultas utilizando la base de datos léxica Wordnet:

  http://wordnet.princeton.edu

Requisitos

- Deberán expandirse los términos de las consultas a otros términos relacionados semánticamente. Deberá tenerse en cuenta:
  - Qué *synset* de WordNet es el apropiado para cada término.

  - Qué relaciones semánticas de las recogidas en WordNet se tendrán en cuenta.
  - Cómo se incorporará esta información en el modelo de recuperación utilizado.
- La aplicación deberá mostrar como salida los resultados con y sin expansión de la consulta de modo que puedan compararse.

Recursos

- Se entregan Diccionarios Wordnet y un ejemplo que muestra cómo se utiliza Wordnet en un proyecto Java.

Módulo II: Reconocimiento de Entidades

Objetivo

Incorporar técnicas de reconocimiento de entidades al motor, con la finalidad de aproximarlo a uno de tipo *Question Answering*.

Requisitos

- Indización de documentos
  - Deberán añadirse al proceso de indización del Módulo base técnicas de reconocimiento de entidades de tipo persona.
  - Se deberán utilizar listados de nombres de persona proporcionados en aula global, que no **podrán modificarse**.
- Introducción de consultas
  - Deberá implementarse algún mecanismo para detectar que la respuesta a la consulta es una entidad de tipo persona.
- Recuperación de documentos
  - Deberán tenerse en cuenta en la recuperación las páginas que contienen una entidad que pueda responder a la consulta y promocionarlas respecto a las restantes.

Recursos

- Se entregan dos listados con nombres de personas, masculinos y femeninos.

Estructura del proyecto propuesta