Recuperación y acceso a la información

GRADO EN INGENIERÍA INFORMÁTICA

Práctica 2
Modelo vectorial

Realizado por:

Rodrigo Borges (Grupo 80-100317579)

(100317579@alumnos.uc3m.es)

David del Rey García (Grupo 80-100315243)

(100315243@alumnos.uc3m.es)

Contenido

Introducción 2
Descripción de la arquitectura 2
Diagrama de clases 2
CreadorDiccionario 2
Calculador 3
Modelos 3
Resultados obtenidos 4
Consulta 1 What video game won Spike's best driving game award in 2006? 4
Consulta 2 What is the default combination of Kensington cables? 4
Consulta 3 Who won the first ACM Gerard Salton prize? 5
Mejoras implementadas 6

Índice de ilustraciones

Ilustración 1. Diagrama de clases. 2
Ilustración 2. Primera consulta Q1. 4
Ilustración 3. Segunda consulta Q2. 4
Ilustración 4. Tercera consulta Q3 5
Ilustración 5. Conjunto de consultas estáticas 5

Introducción

El proyecto presentado está compuesto por algoritmos que permiten conocer la relevancia que tienen ciertos documentos frente a una consulta, en función de las palabras que la componen. Concretamente se implementarán las siguientes funciones de similitud:

- Producto escalar TF.
- Producto escalar TF IDF.
- Coseno TF.
- Coseno TF IDF.

Descripción de la arquitectura

Diagrama de clases

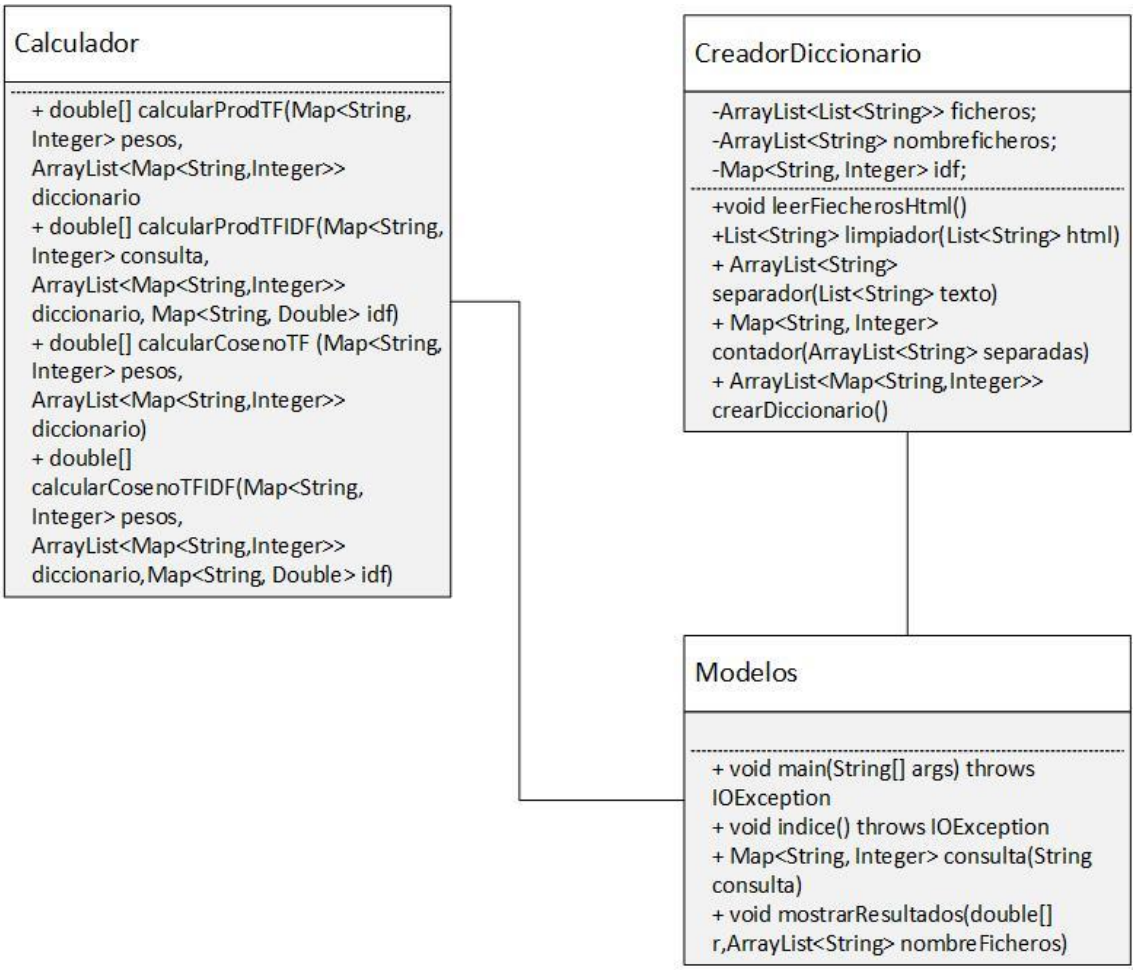

Ilustración 1. Diagrama de clases.

CreadorDiccionario

Esta clase es la encargada de crear el diccionario con las palabras de los documentos.

Para poder crear el diccionario se parte de unos documentos html. Estos documentos han de estar almacenados en la ruta "*./htmls/*", que se encuentra en la carpeta raíz del proyecto.

Una vez leídos los documentos se limpian para obtener únicamente el texto relevante de cada uno, para ello se ha utilizado la librería "*Jsoup*"; y se separan las palabras, almacenando cada palabra con su frecuencia en el documento, de manera que el diccionario se compone de páginas, y cada página representa a un documento, almacenando sus palabras con la frecuencia.

Calculador

Esta clase es la encargada de realizar los cálculos de los índices de relevancia en función de la consulta realizada y del diccionario proporcionado.

Los índices de similitud calculados son:

- Producto escalar TF: $\sum_{i=1}^{n}(TF_{iJ} * w_{iq})$

- Producto escalar TF-IDF: $\sum_{i=1}^{n}(TF_{iJ} * IDF_i)$

- Producto vectorial TF: $\frac{\sum_{i=1}^{n}(TF_{iJ} * w_{iq})}{\sqrt[2]{\sum TF_j * \sum IDF_Q}}$

- Producto vectorial TF-IDF: $\frac{\sum_{i=1}^{n}(TF_{iJ} * IDF_i)}{\sqrt[2]{\sum TF_j * \sum IDF_Q}}$

Modelos

Es la clase encargada de mostrar al usuario las opciones disponibles, y en función de la elección del usuario el programa realizará la acción solicitada.

Las opciones disponibles son:

- Crear/Actualizar diccionario: con esta opción el programa llama a la clase CrearDiccionario para que esta lea la ruta "./htmls/" y cree un diccionario con todos los documentos ".html" que se encuentren en ella.
- Realizar consulta: una vez creado el diccionario el usuario podrá realizar una consulta, para ello ha de seleccionar la segunda opción del índice y escribir la consulta que desee realizar. Se mostrarán los resultados por pantalla y se le volverá a dar al usuario la opción de elegir acción en el índice.
- Mostrar consultas estáticas: en la práctica se proponen 3 consultas. Esta opción realizará las consultas y las mostrará por pantalla en un formato diferente al general para poder compararlas.
- Salir: esta opción terminará la ejecución del programa.

Resultados obtenidos

Consulta 1 What video game won Spike's best driving game award in 2006?

```
Introduzca la consulta que desea realizar
What video game won Spike's best driving game award in 2006?
Los resultados se mostrarán por categoría y relevancia

-- Producto escalar TF --
El documento 2010-42-103.html tiene una relevancia de 265.0 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 95.0 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 32.0 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 8.0 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 7.0 sonbre la consulta realizada

-- Producto escalar TF-IDF --
El documento 2010-42-103.html tiene una relevancia de 37.705888720785524 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 6.535252654476011 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 0.5261252694463808 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 0.2648674397741473 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 0.009391550621221665 sonbre la consulta realizada

-- Coseno TF --
El documento 2010-42-103.html tiene una relevancia de 0.2538894262507483 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.12717662165819243 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 0.08073654849666627 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 0.039777864208786505 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 0.03553134984368876 sonbre la consulta realizada

-- Coseno TF-IDF --
El documento 2010-42-103.html tiene una relevancia de 0.3225880801334257 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.05485820503178932 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 0.025675556637310963 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 0.005698052408579516 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 2.1820475994280514E-4 sonbre la consulta realizada
```

Ilustración 2. Primera consulta Q1.

Consulta 2 What is the default combination of Kensington cables?

```
Introduzca la consulta que desea realizar
What is the default combination of Kensington cables?
Los resultados se mostrarán por categoría y relevancia

-- Producto escalar TF --
El documento 2010-42-103.html tiene una relevancia de 317.0 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 160.0 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 129.0 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 60.0 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 11.0 sonbre la consulta realizada

-- Producto escalar TF-IDF --
El documento 2010-58-044.html tiene una relevancia de 3.129664896659923 sonbre la consulta realizada
El documento 2010-42-103.html tiene una relevancia de 0.7231493978340682 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 0.5715844778356546 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.3662704742276449 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 0.0 sonbre la consulta realizada

-- Coseno TF --
El documento 2010-58-044.html tiene una relevancia de 0.4510385844931978 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 0.43055529602545506 sonbre la consulta realizada
El documento 2010-42-103.html tiene una relevancia de 0.4017695575674871 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.2833496534643605 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 0.0646298918600735 sonbre la consulta realizada

-- Coseno TF-IDF --
El documento 2010-58-044.html tiene una relevancia de 0.27470938671811285 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 0.022116848663674558 sonbre la consulta realizada
El documento 2010-42-103.html tiene una relevancia de 0.0111278718941821 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.005530012028939013 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 0.0 sonbre la consulta realizada
```

Ilustración 3. Segunda consulta Q2.

Consulta 3 Who won the first ACM Gerard Salton prize?

```
Introduzca la consulta que desea realizar
Who won the first ACM Gerard Salton prize?
Los resultados se mostrarán por categoría y relevancia

-- Producto escalar TF --
El documento 2010-42-103.html tiene una relevancia de 196.0 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 115.0 sonbre la consulta realizada
El documento 2010-99-086.html tiene una relevancia de 99.0 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 31.0 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 5.0 sonbre la consulta realizada

-- Producto escalar TF-IDF --
El documento 2010-99-086.html tiene una relevancia de 13.46602226583512 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.8486561097095154 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 0.4885590669614942 sonbre la consulta realizada
El documento 2010-42-103.html tiene una relevancia de 0.4523433599782888 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 0.009391550621221665 sonbre la consulta realizada

-- Coseno TF --
El documento 2010-99-086.html tiene una relevancia de 0.3304261574148841 sonbre la consulta realizada
El documento 2010-42-103.html tiene una relevancia de 0.24841272329093841 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 0.23303660198815218 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.2036575634275091 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 0.02937722357276068 sonbre la consulta realizada

-- Coseno TF-IDF --
El documento 2010-99-086.html tiene una relevancia de 0.2885166639494375 sonbre la consulta realizada
El documento 2010-58-044.html tiene una relevancia de 0.02374550947310593 sonbre la consulta realizada
El documento 2010-76-088.html tiene una relevancia de 0.00709487357753341 sonbre la consulta realizada
El documento 2010-42-103.html tiene una relevancia de 0.0038542604794564684 sonbre la consulta realizada
El documento 2010-22-100.html tiene una relevancia de 2.173191151240664E-4 sonbre la consulta realizada
```

Ilustración 4. Tercera consulta Q3

En la siguiente imagen se muestra la relevancia de las 3 consultas propuestas en los 5 documentos "*.html*". No están ordenados por relevancia como en las consultas más generales.

-- Producto Escalar TF --	Q1	Q2	Q3
2010-22-100.html	8.0	11.0	5.0
2010-42-103.html	265.0	317.0	196.0
2010-58-044.html	7.0	60.0	31.0
2010-76-088.html	95.0	160.0	115.0
2010-99-086.html	32.0	129.0	99.0
-- Producto Escalar TF-IDF --	Q1	Q2	Q3
2010-22-100.html	0.009391550621221665	0.0	0.009391550621221665
2010-42-103.html	37.705888720785524	0.7231493978340682	0.4523433599782888
2010-58-044.html	0.5261252694463808	3.129664896659923	0.4885590669614942
2010-76-088.html	6.535252654476011	0.3662704742276449	0.8486561097095154
2010-99-086.html	0.2648674397741473	0.5715844778356546	13.46602226583512
-- Producto Escalar TF--	Q1	Q2	Q3
2010-22-100.html	0.03553134984368876	0.0646298918600735	0.02937722357276068
2010-42-103.html	0.2538894262507483	0.4017695575674871	0.24841272329093841
2010-58-044.html	0.039777864208786505	0.4510385844931978	0.23303660198815218
2010-76-088.html	0.12717662165819243	0.2833496534643605	0.2036575634275091
2010-99-086.html	0.08073654849666627	0.43055529602545506	0.3304261574148841
-- Producto Escalar TF--	Q1	Q2	Q3
2010-22-100.html	2.1820475994280514E-4	0.0	2.173191151240664E-4
2010-42-103.html	0.3225880801334257	0.0111278718941821	0.0038542604794564684
2010-58-044.html	0.025675556637310963	0.27470938671811285	0.02374550947310593
2010-76-088.html	0.05485820503178932	0.005530012028939013	0.00709487357753341
2010-99-086.html	0.005698052408579516	0.022116848663674558	0.2885166639494375

Ilustración 5. Conjunto de consultas estáticas

demás, en algunos índices de similitud. Esto se debe a que poseen mayor cantidad de texto, por lo que se dan más coincidencias de palabras de las consultas en los textos, la mayoría palabras carentes de valor. Por ello una de las posibles mejoras sería la eliminación de palabras carentes de valor para que las consultas fuesen más precisas. En cambio, al no haber realizado

dicha mejora aún, hay ciertos ficheros que no son relevantes para las consultas realizadas pero aparecen como relevantes al tener un alto número de coincidencias.

Mejoras implementadas

Normalización de términos: tanto para los documentos como para las consultas hemos eliminado los caracteres especiales (solo mantenemos caracteres alfanuméricos) y transformado a minúsculas todos los términos. Esta es la primera aproximación a la normalización de términos, en la siguiente versión se reducirá cada termino a su raíz.

Limpieza de etiquetas: para llevar a cabo la eliminación de las etiquetas hemos utilizado la librería Jsoup de java. Durante el desarrollo tuvimos problemas con la eliminación del símbolo '>' al final de las etiquetas, si no existía un espacio entre texto etiqueta y texto Jsoup nos juntaba las palabras. Para solucionar este problema añadimos un espacio cada vez que encontramos el símbolo '>' y de esta manera evitamos que nos concatenara palabras que debían estar separadas.

Separación de cálculo de pesos de la función de similitud: los valores de TF e IDF se calculan durante la creación del diccionario y las funciones de similitud se calculan utilizando la clase calculador, cada función de similitud tiene su propio método. En la siguiente versión se re-factorizara para restar responsabilidad la clase calculador.

Filtrado de palabras vacías: no se ha realizado ningún filtrado sobre estas palabras. En la siguiente versión se incluirá esta funcionalidad.