



UNIVERSIDAD DE GRANADA

Recuperación de Información

Práctica 3

Implementación de un Sistema de Recuperación de Información utilizando Lucene

Diseño, indexación, búsqueda e indexación de
facetas

David Carrasco Chicharro
Daniel Terol Guerrero

1. Indexación

Para poder realizar el proceso de indexación se ha creado una clase, *ExtraerPelícula*, que almacena la información relacionada a una película. A partir del archivo *wiki_movie_plots_deduped.csv*, se ha leído el fichero entrada a entrada, extrayendo los diferentes campos a través de una biblioteca capaz de manejar archivos CSV de forma cómoda y eficiente.

Cada entrada del archivo CSV, se ha almacenando en un vector de objetos de la clase *ExtraerPelícula* para, posteriormente, ser procesado de forma completa todo el *dataset* y realizar así la indexación:

Se recorre el vector de películas y, por cada película, se establecen los diferentes campos de indexación.

- a. Título
- b. Director
- c. Reparto
- d. Género
- e. Origen

Además de ser campos de indexación, *Género* y *Origen* se han establecido como *Facetas*, pudiendo así, posteriormente, filtrar los resultados de la búsqueda a partir de estos campos.

El índice creado genera varios ficheros que se almacenan en el directorio */indice*.

2. Búsqueda

Para realizar la búsqueda, se hace uso de una interfaz donde se puede establecer el campo de búsqueda deseado y la consulta. Una vez se introduce la consulta, se realiza una *Query* a través de un *parser*.

```
1 Query q = parser.parse(campo+":\""+consulta+"\");
```

Es entonces, a partir de la consulta y del número de documentos que se quieren mostrar como resultado, cuando se realiza la búsqueda en el índice.

```
1 TopDocs tdc = FacetsCollector.search(searcher, q, num_docs, fc);
```

A partir del resultado se recorren todos los documentos incluidos en él y se muestra el título de los diferentes documentos.

```
1 String str = "";
2 for (ScoreDoc sd : tdc.scoreDocs) {
3     Document d = searcher.doc(sd.doc);
4     str+=sd.score + "\t\r-░░░"+d.get("titulo")+"\n";
5 }
6
7 JTextAreaPelisEncontradas.setText(str);
8 JTextAreaPelisEncontradas.update(jTextAreaPelisEncontradas
9 .getGraphics());
```

Una vez se tienen los resultados, aparecen las categorías asociadas a la búsqueda.

The screenshot shows a web application interface for searching books. It has two main sections: 'Campo de búsqueda' (Search Field) and 'Consulta' (Query). In the 'Campo de búsqueda' section, there is a dropdown menu set to 'Titulo' and a text input field containing 'Harry Potter'. In the 'Consulta' section, there is a 'Buscar' button. Below these, there is a 'Categoría' section with a list box containing 'genero' and 'origen', and a 'Buscar categoría' button. To the right, it says 'Número de documentos encontrados: 11 hits'. Below this, a list of results is displayed, each with a score and a title: '4.4960327 - Harry Potter and the Sorcerer's Stone', '4.4960327 - Harry Potter and the Philosopher's Stone', '4.0850296 - Harry Potter and the Chamber of Secrets', '4.0850296 - Harry Potter and the Prisoner of Azkaban', '4.0850296 - Harry Potter and the Goblet of Fire', '4.0850296 - Harry Potter and the Half-Blood Prince', '3.7428758 - Harry Potter and the Order of the Phoenix', '3.7428758 - Harry Potter and the Deathly Hallows: Part 1', '3.7428758 - Harry Potter and the Deathly Hallows: Part 2', and '3.7428758 - Harry Potter and the Deathly Hallows: Part I'.

Figura 1: Resultado de búsqueda

Al seleccionar una de las categorías se muestran las diferentes facetas asociadas a la consulta, pudiendo seleccionar una faceta y, por tanto, filtrando la búsqueda a través de una categoría y faceta determinada.

Para poder realizar el filtrado, se hace uso de *DrillDownQuery* y *DrillSiblings*.

```

1 DrillDownQuery dq = new DrillDownQuery(fconfig, q));
2 [...]
3 DrillSideways ds = new DrillSideways(searcher, fconfig,
4   taxoReader);
5 DrillSideways.DrillSidewaysResult dsresult = ds.search(dq,10);

```

Al filtrar por categorías y faceta, el resultado queda como se muestra a continuación:

The screenshot shows a web-based search interface. At the top, there is a search bar labeled 'Campo de búsqueda' with a dropdown menu set to 'Titulo'. The search term 'Harry Potter' is entered in the 'Consulta' field, and a 'Buscar' button is next to it. Below the search bar, the 'Categoría' section shows a list of categories: 'genero' and 'origen'. A 'Buscar categoría' button is below this list. To the right of the categories, it says 'Número de documentos encontrados: 2 hits'. Below this, a list of results is shown: '3.7428758 - Harry Potter and the Deathly Hallows: Part 1' and '3.7428758 - Harry Potter and the Deathly Hallows: Part 2'. At the bottom, there is a 'Faceta' section with a list of facets: 'fantasy (4)', 'family, fantasy (4)', 'action-adventure, fantasy (2)', and 'unknown (1)'. A 'Buscar Faceta' button is at the bottom right of the facets list.

Figura 2: Resultado de búsqueda por facetas

3. Ejecución

Para ejecutar el programa hay que añadir el archivo CSV *wiki_movie_plots_deduped.csv* al directorio del proyecto y ejecutar el archivo *ejecucion.sh*.