Proyecto Final: Sistema de Recuperación de Información

Universidad Carlos III de Madrid, Multimedia, Curso 2020-2021

1. Introducción y objetivos

El objetivo de esta práctica es construir un sistema de Recuperación de Información basado en ElasticSearch, una herramienta para la indexación de ficheros: https://www.elastic.co/es/products/elasticsearch

En la primera parte de la asignatura se trabajó con una colección de imágenes correspondientes a carteles de películas disponibles en IMDB. Esas imágenes van acompañadas de información adicional sobre la película en cuestión, como director, actores, fecha de estreno, etc. También se incluye una sinopsis sobre la película e incluso críticas o comentarios de personas que la han visto. En la siguiente imagen se muestra un ejemplo.

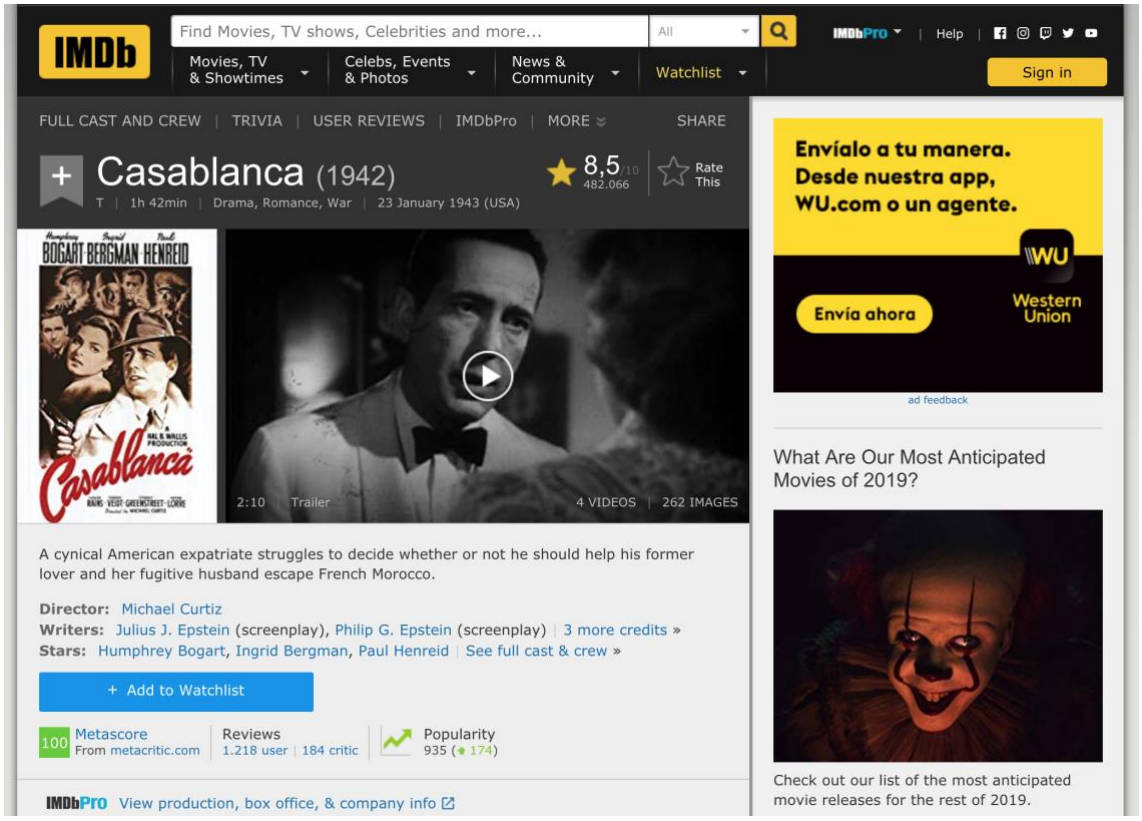

Ilustración 1. Ejemplo de información de una película disponible en IMDB

El sistema de búsqueda a desarrollar debe permitir la recuperación de información a partir de este contenido textual.

2. Descarga de Elasticsearch

Se puede descargar para diferentes plataformas de esta web:

https://www.elastic.co/es/downloads/elasticsearch

En las aulas informáticas se encuentra instalada una instancia de ElasticSearch pero se puede instalar en equipos personales.

Los pasos para realizar la instalación pueden encontrarse en la web de descarga.

3. Desarrollo de la práctica

La implementación del sistema de búsqueda se dividirá en varios pasos o fases.

3.1 Paso 1

Definición del esquema de índices de ElasticSearch para el almacenamiento de los datos de las películas teniendo en cuenta las consultas indicadas en el Paso 3.

Este esquema y su descripción deberán incluirse en la memoria de la práctica en formato PDF que forma parte de la entrega.

3.2 Paso 2

Sobre una instancia de ElasticSearch, desplegar el esquema de índices definidos en el Paso 1.

Indexar las páginas HTML con los datos de cada película. Conviene tener en cuenta que el fichero Excel con la información de las imágenes de las portadas de las películas contiene también el enlace a la página HTML correspondiente a esa película.

NOTA: En el proceso de indexación podría resultar útil el filtro HTML Strip Char disponible en ElasticSearch[1] , la librería Tika[2] o la librería de Java JSoup[3].

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-htmlstrip-charfilter.html
[2] http://tika.apache.org
[3] https://jsoup.org/

3.3 Paso 3

Implementar las consultas necesarias para ejecutar las siguientes búsquedas sobre la colección de documentos:

1. Películas sobre animales producidas desde el año 1950 en adelante.
2. Actores con más películas de aventuras.
3. Películas de temática "policiaca" disponibles en la colección. Además del listado, mostrar el número de películas por año.
4. Películas que traten sobre temas sociales en España y Latinoamérica.

Es necesario incluir en la memoria en PDF las consultas en ElasticSearch y los resultados que se han obtenido para cada una de ellas (limitado a los primeros 50 resultados por consulta).

4. Entrega y Defensa de la memoria

Importante: para la evaluación de este ejercicio, es necesario subir a Aula Global un archivo comprimido (preferiblemente zip) que contenga lo siguiente:

- Un documento en formato .pdf que debe contener las respuestas a las preguntas formuladas en las actividades presentes en el apartado Desarrollo. Incluir en la portada al menos está información: nombre, apellidos, NIA de cada componente del grupo, asignatura y curso académico o fecha.

- Un directorio que contenga todos los archivos requeridos para la cada uno de los pasos, incluyendo esquemas de índice, código, etc.

Nota: Es suficiente con que un único miembro del grupo cuelgue el proyecto en Aula Global.

La entrega se realizará mediante Aula Global y se defenderá presencialmente el último día de clase (17 de diciembre).