Universidad Carlos III

Curso Multimedia 2022-23

Proyecto Final

Curso 2022-23

Sistema de Recuperación de Información

Carlos Montero Gómez de las Heras - 100405884

Álvaro Morata Hontanaya - 100405846

Fecha: **12/12/22**

GRUPO: **81**

Índice

Esquema de índices

A continuación se incluye una tabla del esquema de índices:

Índice	Nombre	Tipo	Se puede buscar	Se puede agregar
movies	Title	Text	SI	NO
	Year	Long	SI	SI
	Director	Text	SI	SI
	Genres	Keyword	SI	SI
	Actors	Keyword	SI	NO
	Guionist	Keyword	SI	SI
	Sinopsis	Text	SI	NO
	Summary	Text	SI	NO
	Score	Keyword	SI	SI

En los documentos se ha agregado un JSON con el mapping del índice de ***movies***.

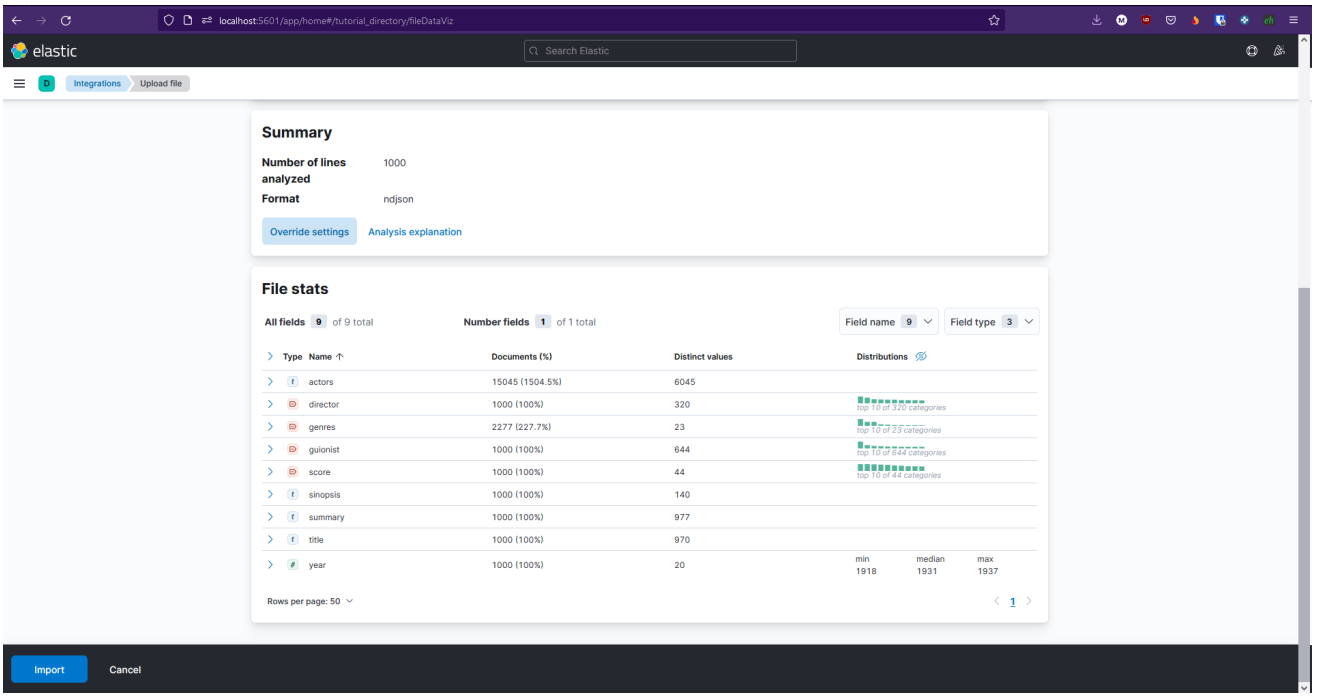
Proceso de indexación

Para el proceso de indexación se hicieron los siguientes pasos:

I. Una vez obtenido con el crawler el JSON (movies.json), que se tuvo que parar manualmente porque el proceso estaba tardando demasiado, puesto que a partir de la película 27.513 la URL obtenida del Excel nos devolvía un error 404, se colocó un corchete de cierre en el archivo para que el formato del JSON no tuviese errores.

II. Una vez hecho esto, se usó una herramienta llamada JQ para formatear el JSON de tal forma que se pudiese importar con Kibana.

III. Se usó Kibana para indexar el archivo (mediante el botón de *Upload a file* de la pestaña principal), puesto que, al estar trabajando desde Windows 10, el uso de CURL nos estaba dando ciertos problemas, incluso cambiando las comillas dobles (") por el carácter de escape (\"). A continuación se muestra el proceso de indexación desde Kibana mediante una captura de pantalla.



IV. Como se puede observar, Kibana autogenera los campos del índice. Se presiona el botón de *Import* y se añade un nombre al índice.

V. Procede a la indexación de las películas, avisando de que 135 de ellas fallan dado que en esas 135 el campo *"year"* tiene el valor "Not rated".

VI. Desde Kibana se puede comprobar que el índice se ha creado correctamente mediante la pestaña de *"Stack Management"* de la barra lateral de Kibana.

Consultas sobre Elasticsearch

Consulta: películas sobre la Segunda Guerra Mundial producidas desde el año 1980 en adelante.

```
GET movies/_search
{
  "size": 50,
  "query":{
    "bool":{
      "must": {
        "multi_match": {
          "query": "world war ii",
          "fields": ["summary", "sinopsis"]
        }
      },
      "filter":{
        "range":{
          "year":{
            "gte":1980
          }
        }
      }
    }
  }
}
```

- **HITS**: 2765
- El índice de precisión es del 100%, tras revisar la muestra de 50 resultados todas las películas cumplían con los requerimientos necesarios.
- La consulta está formada por un dos cuerpos, un *must* con el que reducimos las películas a sólo aquellas que tienen la palabra mostrada en la sinopsis o resumen.
  Pueden existir películas que cumplan con estos requisitos pero de no tener sinopsis o resumen no se pueden contabilizar.
  Después se realiza un filtro en el que solo se tienen en cuenta aquellas cuyo año de lanzamiento es mayor o igual a 1980.

Consulta: directores con más películas de acción.

```
GET movies/_search
{
  "size": 0,
  "aggs": {
    "terms_director": {
      "terms": {
        "field": "director",
        "size": 50
      }
    }
  },
  "query":{
    "match":{
      "genres":"Action"
    }
  }
}
```

- **HITS**: 3894
- Presuponemos que el índice de precisión es del 100% puesto que estamos evaluando los directores y la cantidad de películas en las que aparecen como director que son de acción. Sobre la muestra de 50 directores.
  En los resultados está agregada esa muestra de los 50 directores que se obtuvieron como resultado de la consulta anterior.
- Empezamos con el campo de *size* a 0 puesto que no nos parecía relevante mostrar películas sino mostrar los directores con la cantidad de las mismas. Por ello tenemos una agregación de directores cuyo campo *genres* contenga *Action*.

Consulta: películas de temática LGBTQ disponibles en la colección. Además del listado, mostrar el número de películas por año.

```
GET movies/_search
{
  "size": 50,
  "query": {
    "multi_match": {
      "query": "lgbt lesbian gay bisexual transexual queer",
      "fields": ["summary", "sinopsis"]
    }
  },

  "aggs": {
    "lgbt_year": {
      "terms": {
        "field": "year",
        "size":50
      }
    }
  }
}
```

- **HITS**: 526
- El índice de precisión es del 100% puesto que la muestra de 50 películas ha sido revisada y todas contemplan en su sinopsis o resumen la temática LBGTQ.
  En los resultados está agregada esa muestra de las 50 películas que se obtuvieron como resultado de la consulta anterior.
- La consulta se divide en dos secciones, la primera contiene aquellas palabras relevantes para que una película sea de temática LGBTQ, así mismo se han comprobado todas ellas para que sean relevantes. La segunda sección contiene la agregación de la consulta que nos muestra cuantas películas que cumplen la parte anterior se han publicado en cada año.

Consulta: películas que traten sobre políticos corruptos en Europa y Estados Unidos.

```
GET movies/_search
{
  "size":50,
  "query":{
    "bool": {
      "must":{
        "multi_match": {
          "query": "US United States of America USA Europe Russia Ukraine France Spain Sweden Norway Germany
            Poland Finland Italy Britain UK Romania Portugal Belgium",
          "fields": ["summary", "sinopsis"]
        }
      },
      "filter":{
        "multi_match":{
          "query": "politic politicians corruption corrupt corrupted",
          "fields": ["summary", "sinopsis"]
        }
      }
    }
  }
}
```

- **HITS**: 475
- El índice de precisión es del 70% aproximadamente. Todas las películas tienen trama en países de Europa y en Estados Unidos, y alrededor del 70% tratan sobre políticos corruptos, el otro 30% se han detectado otros diferentes tipos de corrupción pero no en política.
- La consulta se divide en dos secciones, la primera obliga a que se muestren películas con tramas en países europeos y en Estados Unidos. La segunda sección de la consulta contempla aquellas películas con palabras relevantes a “políticos corruptos”.