PRÁCTICA DE RECUPERACIÓN DE INFORMACIÓN

Julian Barcia Facal julian.bfacal@udc.es

Brais García Brenlla b.brenlla@udc.es

1. Dominio de información

Toda a información que se recolle nesta práctica é relativa ao dominio dos libros. Toda esta obtense directamente dos datos subministrados por "La Casa del Libro" (https://www.casadellibro.com/). Esta páxina web proporciona una plataforma de venta de libros cun gran catálogo, tendo de cada un destes unha gran cantidade de información específica.

O obxectivo final desta práctica é a de crear un sistema capaz de recompilar información sobre un gran cantidade de libros, directamente desta páxina web e almacenala en ElasticSearch para poder ser empregada posteriormente. A información que se recolle destes será:

- Nome do libro
- Autor
- Editorial
- Categorías
- Ano de edición
- Nº de páxinas
- Idioma
- Sinopse
- Imaxe de carátula

Todos estes datos intégranse directamente en ElasticSearch nun índice común denominado *book,* onde cada obxecto interno terá cada un dos campos anteriores, conte ou non con información para cada un deles. Polo tanto, non se garda en local os datos recolectados en ningún momento. Como obxectivo final buscarase crear unha páxina web para que o usuario final, mediante unha interfa máis sinxela e accesible, poida consultar calquera destes datos. Para crear a páxina web empregarase como base React e máis concretamente a librería de SearchUi, que adapta directamente ElasticSearch para este tipo de interfaces.

2. Tecnoloxías

Python3

Linguaxe de programación dinámico. A versión usada concretamente é a 3.6.9. Empregarase maioritariamente para o codificación da parte de scrapping da páxina web.

ElasticSearch

Motor de búsqueda e analítica para o analise e visualización de datos.

JavaScript

Linguaxe de programación lixeiro, interpretado basado en obxectos e prototipos empregado maioritariamente en deseño de interfaz de usuario e páxinas web.

React

Librería de JavaScript de código aberto deseñado para crear interfaces de usuario. Permite a creación destas mediante o uso de pezas individuais denominadas compoñentes.

Search UI

Librería de JavaScript para a creación de experiencias de busca incorporadas con Elastic. Deseñado para incorporarse con React sen configuracións.

3. Configuración do entorno

Os pasos a realizar en completo móstranse no archivo README exposto en GitHub.
Os pasos para realizar a instalación son os seguintes:

- Executar o instalador de Python 3.6.9
- Para instalar as librerías necesarias para o uso de Elastic farase mediante o archivo requirements.txt (https://github.com/Julian-BFacal/RIWS/)

pip install -r requirements.txt

- Para que Elastic funcione correctamente con React débese modificar a configuración interna deste. No archivo elasticsearch.yml débense engadir as seguintes liñas:

http.cors.enabled: true
http.cors.allow-credentials: true
http.cors.allow-origin: ""*
http.cors.allow-methods: OPTIONS, HEAD, GET, POST, PUT, DELETE
http.cors.allow-headers: ""*

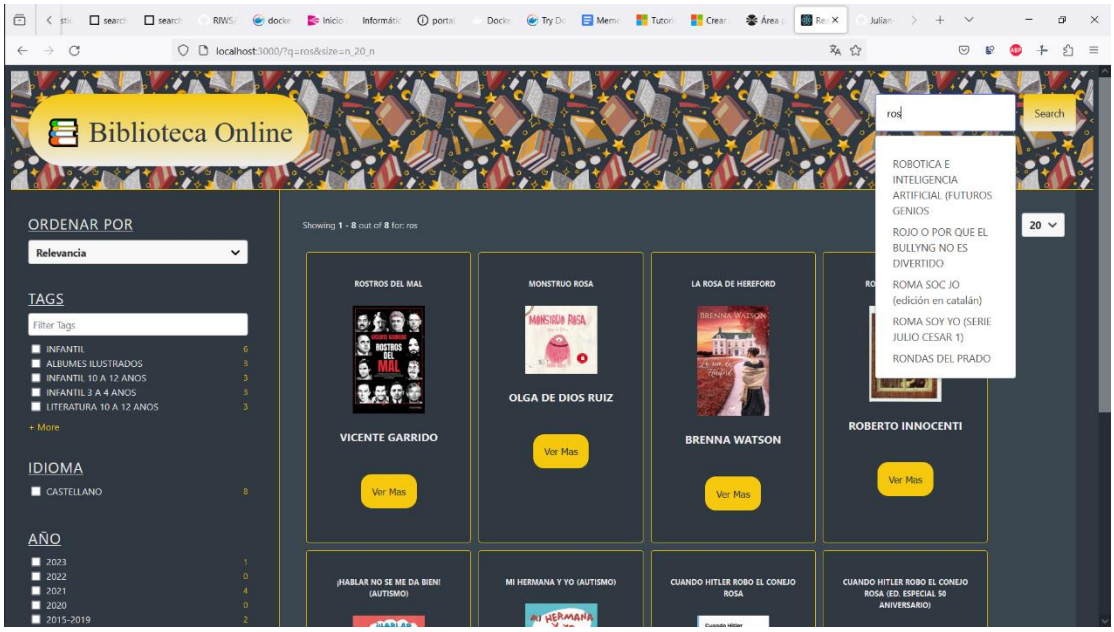
Ademais para a execución deshabilitamos a autentificación, no mesmo archivo yml:

xpack.security.enabled: false
xpack.security.enrollment.enabled: false

4. Casos de busca

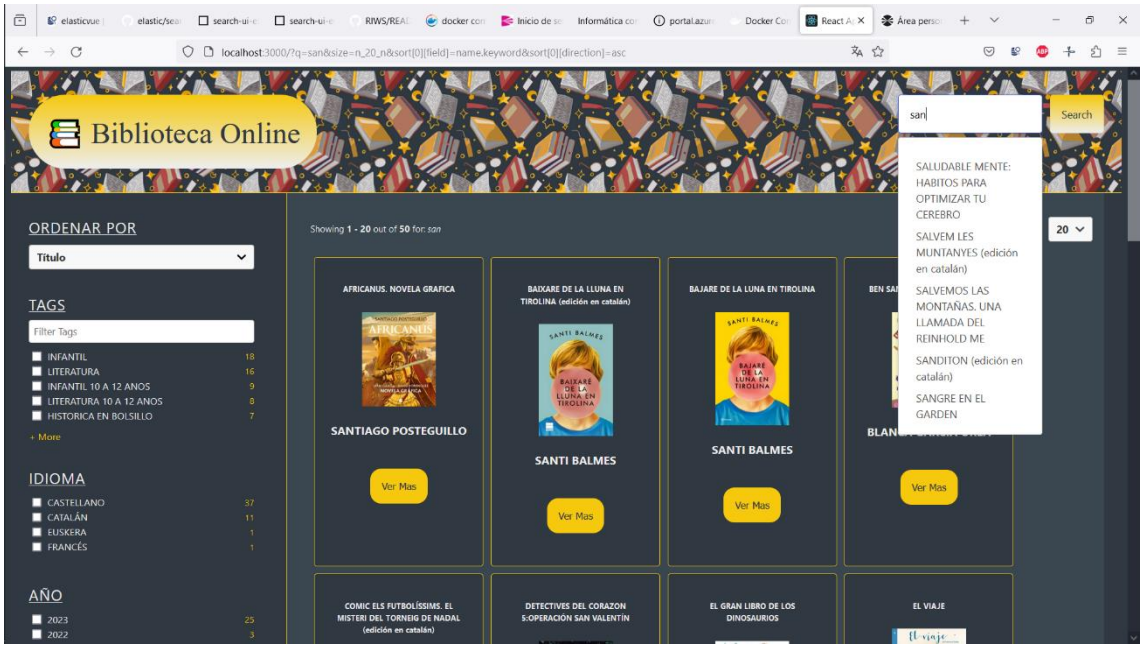
a. Busca por nome de libro

Mediante o buscador de texto da parte superior dereita pódese buscar o título dun libro. Este filtrará os resultados non só por búsquedas exactas se non que tamén filtrará a búsqueda por *sub-strings* do título, empregando cada palabra como un *keyword* sobre o que buscar correspondencias. Estas correspondencias filtrarán os resultados mentres se escriba *(search as you type*). Ademais a partir da terceira letra escrita, o buscador proporá posibles títulos similares que se poden corresponder co texto ata o momento introducido. As búsquedas realizadas neste *searchbox* non están condicionadas nin ao uso de empregar maiúsculas e minúsculas, nin ao uso de acentos.



b. Busca por autor

Mediante o mesmo buscador de texto anterior, o usuario pode introducir o nome do autor para filtrar os resultados. As búsquedas fanse sobre calquera *sub-string* do nome. Por exemplo, se un autor e Santiago Oreiro, con introducir "san" aparecerá na lista de resultados. Para o caso de autor, non se mostran suxerencias. As búsquedas realizadas neste *searchbox* non dependen nin de diferenciar entre minúsculas e maiúsculas, nin de acentos e seguen o modelo de *search as you type* xa explicado.



c. Ordenación dos resultados

Grazas ao despregable da esquerda *ORDENAR POR*, os resultados que se mostran na páxina pódense ordenar de diversas maneiras.

- *Relevancia:* Por defecto, todos os libros presentes no índice, móstranse na orde que foron introducidos no índice de Elastic. No caso de realizar unha búsqueda específica, a ordenación por relevancia priorizará aqueles que conten cun *score* de búsqueda máis alto.
- *Autor:* Tal como o nome indica, este tipo de ordenación mostrará os resultados segundo o orde alfabético dos nomes dos autores.
- *Título:* Neste caso a ordenación dos resultados farase seguindo o orde alfabético dos títulos dos libros.

d. Filtro por categorías ou tags

Cada libro conta cunha ou varias categorías ou tags aos que pertence. Debido a importancia destas, o usuario pode filtrar os resultados directamente sobre estas categorías. Isto farao mediante a sección *TAGS.* Esta sección mostra de forma predetermina as 5 categorías máis comúns, ampliable ata 20 no caso de pulsar *More*.
A búsqueda sobre estes *tags* e restritiva e aditiva (*AND Based Facet Filter*), polo tanto só se mostrarán aqueles resultados que teñan exactamente todas as categorías seleccionadas. Por exemplo no caso de seleccionar *infantil*, as posibles categorías adicionais a seleccionar redúcense a só aquelas nas que o resultado conte coa categoría *infantil* xa seleccionada.
Para axudar ao usuario, sobre a lista de seleccionables móstrase un buscador que permite realizar unha búsqueda sobre os tags. Non obstante, a búsqueda só se realiza sobre aquelas tags xa presentes na lista de 20 e non sobre todos os posibles.

e. Filtro por Idioma

Mediante a sección *IDIOMA* o usuario pode filtrar os resultados mediante os idiomas que seleccione. Esta selecciona é aditiva pero non restritiva (*OR Based Facet Filter*), pódese filtrar por varios idiomas ao mesmo tempo, mostrando aqueles libros que conten cun ou outro idioma dos seleccionados.

f. Filtro por Ano de Edición

Debido a grande variedade de datas que poden existir, a búsqueda que se fará sobre os anos de edición estará condicionada a un rangos prefixados. Os filtros tamén serán *OR Based* polo tanto, tras seleccionar un rango, as opcións a seleccionar nesta sección mantéñense. Co fin dar maior importancia a un filtrado máis concreto aos anos máis próximos os rangos son os seguintes:

- 2023: Todos aqueles libros publicados en 2023
- 2022: Todos aqueles libros publicados en 2022
- 2021: Todos aqueles libros publicados en 2021
- 2020: Todos aqueles libros publicados en 2020
- 2015-2019: Todos aqueles libros publicados entre 2015 e 2019 ambos incluídos.
- 2010-2014: Todos aqueles libros publicados entre 2010 e 2014 ambos incluídos.
- 2000-2009: Todos aqueles libros publicados entre 2000 e 2009 ambos incluídos.
- 1990: Todos aqueles libros datados de entre 1990 e 1999, ambos incluídos.
- 1980-: Todos aqueles libros previos a 1989 con este incluído.

g. Filtro por Nº de Páxinas

Ao igual ca no caso do ano de edición, entendeuse que ante a ampla variedade de opcións o mellor para o usuario era dividir as opcións por rangos. Estes rangos tamén serán *OR Based,* e polo tanto as seleccións son acumulativas, mostrando aqueles resultados que cumpran, unha ou todas as condicións seleccionadas. Os rangos predefinidos son os seguintes:

- 0 - 50: Libros entre 0 e 50 páxinas. Principalmente, libros infantiles e de poesía.
- 50 - 100: Libros entre 50 e 100 páxinas.
- 100 – 200: Libros entre 100 e 200 páxinas.
- 200 – 300: Libros entre 200 e 300 páxinas.
- 300 – 400: Libros entre 300 e 400 páxinas.
- 400+: Libros con máis de 400 páxinas, sen límite.

h. Cantidade de resultados da query

A páxina implementada conta cun modelo de paxinación dos resultados. Isto provoca que a query realizada sempre se limite a unha cantidade máxima de resultados. No noso caso as opcións de paxinación son de 20, 40 e 60 resultados por páxina, podendo calquera destes no despregable da dereita.

5. Código fonte

Todo código fonte está subido a https://github.com/Julian-BFacal/RIWS/