Information Retrieval and Web Analytics, 2025

Universitat Pompeu Fabra Barcelona

Project Part 4: RAG, User Interface, and Web Analytics

Team: G_004

Alex De La Haya Gutiérrez, 268169

Marc Guiu Armengol, 268920

Nil Tomàs Plans, 268384

Github URL: https://github.com/niiltomas/irwa-search-engine-G_004.git

Tag: IRWA-2025-part-4

Index of Contents

1. User Interface

1.1. Search algorithms:

It is implemented inside *myapp/search/algorithms.py -> search_in_corpus()*. We have implemented the search using TF-IDF indexing and BM25 algorithm. We do standard preprocessing (including stopword removal), the text fields we have considered are title and description. The parameters that we have used are the following: k1=1.5, b=0.75, we boost product rating by 1.2x and in-stock by 1.1x. Finally we return top-k 20 results.

1.2. Results Page:

You can find it in *templates/results.html*

Each record has the standard properties: Title, description, selling_price, discount, average_rating, URL(link to document on original website).

We have also implemented an **additional property** that shows whether the item is available or not.

1.3. Details Page

It can be found in *templates/doc_details.html*

It is a page that displays the documents information (title, brand category…).

2. RAG

The model receives a query from the user and generates an answer. The answer is AI powered. The prompt used for the response is very generic / basic.

e.g For the query “blue jeans” this is the AI-generated summary:

“- Best Product: JEAFS2K2YZTV7MFV Slim Women Blue Jeans - Why: This product is the best fit because it matches the user's request exactly and has the most relevant product name, title, and description for blue jeans without specifying gender or slim fit. - Alternative: JEAFESNDV9TYNZJH Slim Women Blue Jeans could also work if you want more options for women's slim blue jeans, but this one might be the more preferred based on the name alone.”

We have identified several **weaknesses**. The model assumes that the best product is the one that matches the query. It only returns the “best product” and an alternative. There is no deep attribute information mentioned in the response (such as price, rating, real fit, stock…). The information given is very generic and does not provide value to the buyer. The LLM should also adapt the language to the receiver (e.g. avoid saying *matches the user’s request* when you are addressing to the user).

Improvements:

1. Force the LLM to take into account specific product data (price, discount, rating, stock) to return a short Top-3 justification.
2. After BM25 retrieval, we rerank the top-K results using both textual relevance and commercial signals so the Top-3 is not only a text match but also a good buying choice.

e.g For the query “blue jeans” this is the improved AI-generated summary:

“*1. JEAFESND4PHGYF3F - True Blue Slim Men Jeans - Why: Best price among the slim-fit jeans options with a high rating of 5.0 and available in stock. 2. JEAFQFGMNEXH2NWC - Levis Skinny Men Blue Jeans - Why: Highly rated (5.0) skinny-fit jeans with a significant discount of 50.0% and available in stock. 3. JEAFSKYHHDQ2SDZQ - Ecko Unltd Tapered Fit Men Light Blue Jeans - Why: High rating of 4.7 and available with a limited 10% discount in stock.”*

As you can observe we have accomplished our goals.

Universitat Pompeu Fabra Barcelona

3. Web Analytics

1. Data collection:

We have collected the following data:

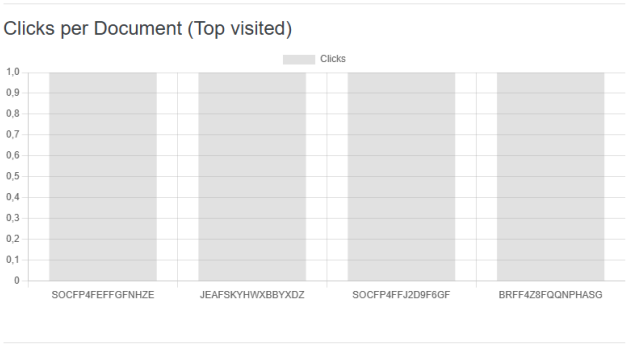
HTTP: Requests data, clicks, HTTP Sessions data. Queries: Number of terms, order. Results: Clicks on documents, to what query they are related, their rank and dwell time. User context: Browser, OS/computer/mobile, time of the day, date and **IP address**.

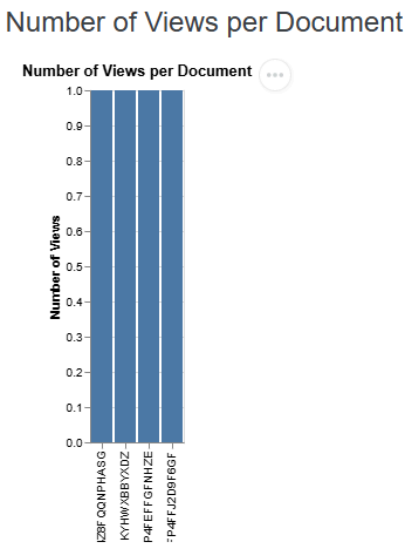
2. Data storage:

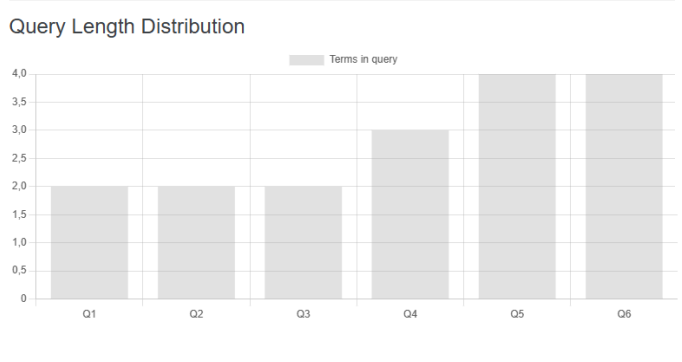
We have designed a data model to store the information above.

3. The analytics dashboard:

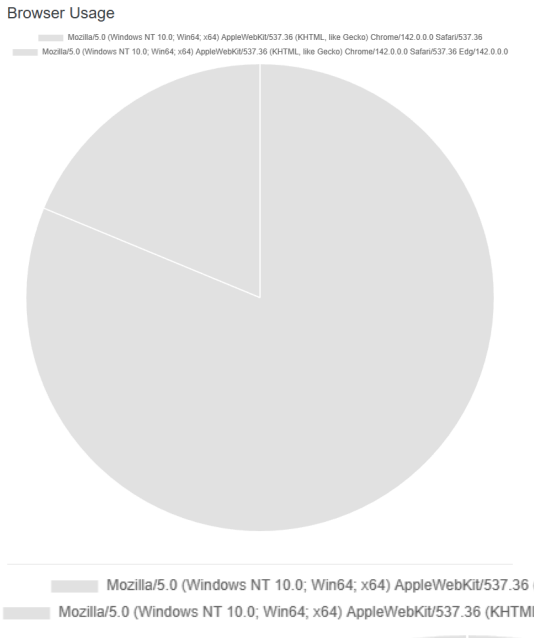
We display different usage statistics such as clicks per document, query length distribution, number of views per document, browser usage and click rank position.

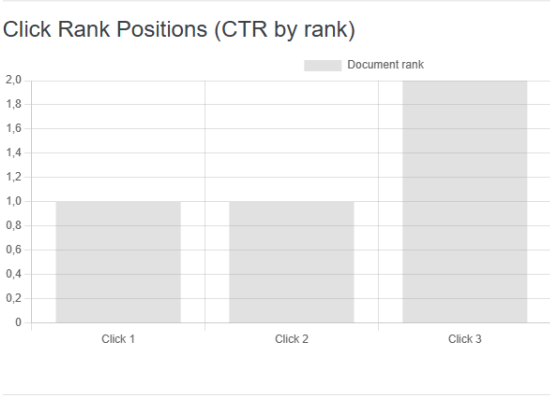






Universitat Pompeu Fabra Barcelona





Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/142.0.0.0 Safari/537.36

Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/142.0.0.0 Safari/537.36 Edg/142.0.0.0