Sergio Pérez: U188428 / Edu Rodriguez: U189644 / Maria Ackermann: U228427

IRWA Report Part 4

User Interface and Web Analytics

Github repository: https://github.com/eduro0/IRWA-2023.git

TAG: IRWA-2023-part-4

Objective:

Creating a realistic search engine

1. General Improvements:

- We made the back function perform its purpose instead of directing to an error by rerendering the previous image.
- We made the pages, specially doc_info, more visually appealing by rewriting the html code.
- If the query has no results it doesn't yield an error.
- We implemented more algorithms: tf-idf, our_score (from the previous practice session) which we can select before searching along with the amount of results you obtain for that search.
- We implemented a sentiment analysis page where the standard is to have the last visited document be the input for the analysis, though the user can write anything he/she wants

2. Data Organization

When it comes to retrieving user data, we decided to organize it in a hierarchical structure of 3 tables with 1-to-N relationships between them. Each of them is saved at intermediate points in the execution as well as when the web app is closed.

The first and highest level entity is the session table, which contains data on the user itself. Namely, information such as the total time with the search engine running, the IP address, operating system and such. This entity is only saved at the end of the execution.

The second table, which references the session involves taking information on the specific searches that are performed by the users and the performance in terms of time and results of the searches. It is saved every time a new search is performed (considering searches with the same query, algorithm and top-k to be the same) or a new session is established.

The last table contains information about the documents visited, namely, the already defined Document object corresponding to that document, as well as the sentiment analysis for it. Every time we leave *doc_info.html* we save the previous document.

All these entities have the time spent in the session, search, and document, which will give us insights on the performance.

3. Insight Research

All the previous entities are implemented as pickle compressions of dictionaries. We then felt the need to create yet another class which rewrites the information as pandas dataframe to aid in the analysis.

In this class we define functions to process the dataframes and create interesting plots, such as the time spent by document, where we can interpret that, if the average time is very low for a specific search, we will assume they have clicked away, if it is too long we will assume they have lost interest and if it is within both ranges, we will see it as a successful recommendation.

Unfortunately, we were unable to finish the dashboard display on time, even though we created a proof of concept. We would have also liked to create more valuable insights given that the data pipeline is perfect and without leakage.