Paula Farrás 254614
Sergi García 251425

GitHub Repo: https://github.com/sergigf03/IRWA-2024

Create a Search Engine:

Project Part 4 IRWA

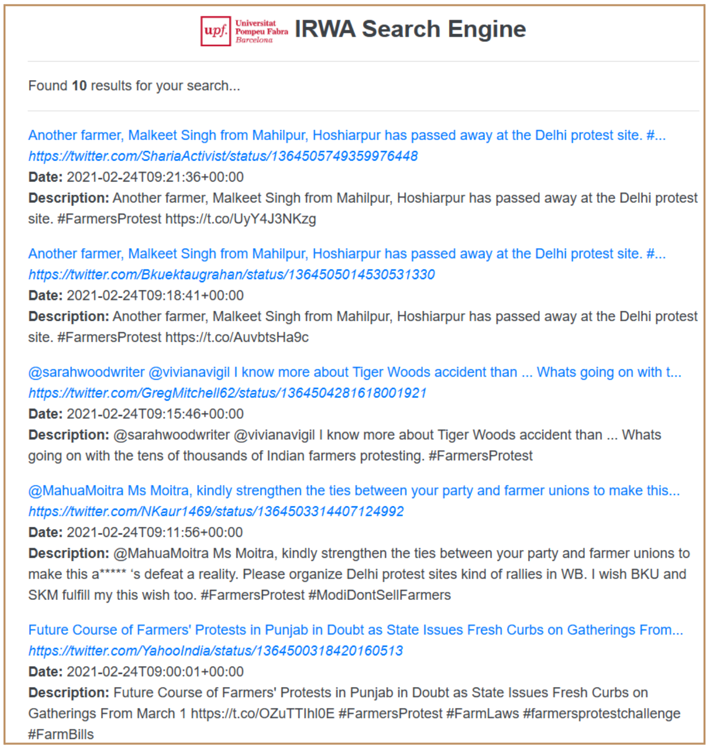
1. Introduction

In this project, our goal is to create a search engine based on a given query and displaying the top N results thanks to adapting our ranking algorithm from the past lab.

2. Creating the website I (Search Engine):

We first downloaded the search_engine directory from the given git repository. From there, we started coding.

We completed load_corpus.py, reusing functions from our lab3, and loading correctly the .json file, successfully creating the corpus. With that, we coded the function: **search_in_corpus**. We had to slightly modify the **create_inverted_index** function to adapt to the newly created corpus. Once algorithms.py was completed, we used **search_in_corpus** to complete the SearchEngine class. We first used that function and later created an instance of ResultItem to store the top 10 results of the search given a specific query. Let's dive into the statistics section now.



IRWA Search Engine

Found **10** results for your search...

Another farmer, Malkeet Singh from Mahilpur, Hoshiarpur has passed away at the Delhi protest site. #...
https://twitter.com/ShariaActivist/status/1364505749359976448
Date: 2021-02-24T09:21:36+00:00
Description: Another farmer, Malkeet Singh from Mahilpur, Hoshiarpur has passed away at the Delhi protest site. #FarmersProtest https://t.co/UyY4J3NKzg

Another farmer, Malkeet Singh from Mahilpur, Hoshiarpur has passed away at the Delhi protest site. #...
https://twitter.com/Bkuektaugrahan/status/1364505014530531330
Date: 2021-02-24T09:18:41+00:00
Description: Another farmer, Malkeet Singh from Mahilpur, Hoshiarpur has passed away at the Delhi protest site. #FarmersProtest https://t.co/AuvbtsHa9c

@sarahwoodwriter @vivianavigil I know more about Tiger Woods accident than ... Whats going on with t...
https://twitter.com/GregMitchell62/status/1364504281618001921
Date: 2021-02-24T09:15:46+00:00
Description: @sarahwoodwriter @vivianavigil I know more about Tiger Woods accident than ... Whats going on with the tens of thousands of Indian farmers protesting. #FarmersProtest

@MahuaMoitra Ms Moitra, kindly strengthen the ties between your party and farmer unions to make this...
https://twitter.com/NKaur1469/status/1364503314407124992
Date: 2021-02-24T09:11:56+00:00
Description: @MahuaMoitra Ms Moitra, kindly strengthen the ties between your party and farmer unions to make this a***** 's defeat a reality. Please organize Delhi protest sites kind of rallies in WB. I wish BKU and SKM fulfill my this wish too. #FarmersProtest #ModiDontSellFarmers

Future Course of Farmers' Protests in Punjab in Doubt as State Issues Fresh Curbs on Gatherings From...
https://twitter.com/YahooIndia/status/1364500318420160513
Date: 2021-02-24T09:00:01+00:00
Description: Future Course of Farmers' Protests in Punjab in Doubt as State Issues Fresh Curbs on Gatherings From March 1 https://t.co/OZuTTIhl0E #FarmersProtest #FarmLaws #farmersprotestchallenge #FarmBills

Universitat Pompeu Fabra Barcelona

Paula Farrás 254614
Sergi García 251425

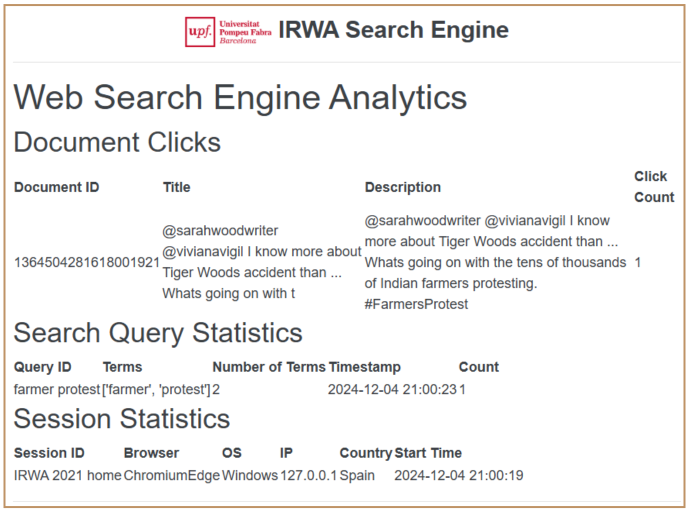
GitHub Repo: https://github.com/sergigf03/IRWA-2024

3. Creating the website II (Statistics):

First, when clicking on a tweet, we are redirected to the doc_details.html page where we are shown the details of the tweet: *description, date, likes, retweets, hashtags and url*
We were asked to perform a series of things to give the user some statistics of the results shown by the search engine. Our plan was to compute:

1. Total number of clicks by tweet: we want to know the number of clicks a certain tweet has, we create a counter that updates each time the tweet is accessed.
2. Query Data: For the query we have given to the search engine, we want to retrieve:
   - The Query ID
   - Its terms
   - Number of terms
   - Timestamp
   - Number of times that query has been searched
3. Session Data: For the current Flask session we want to retrieve:
   - Session ID
   - Browser
   - Operative System
   - IP
   - Country
   - Start Time
4. Dwell Time: we looked into ways of computing it, but we finally could not.

Each of these statistics are able to be seen in the website thanks to completing the part of stats in the web_app.py but also by creating and filling (in web_app.py) the created dictionaries in analytics_data.py.



IRWA Search Engine

Web Search Engine Analytics

Document Clicks

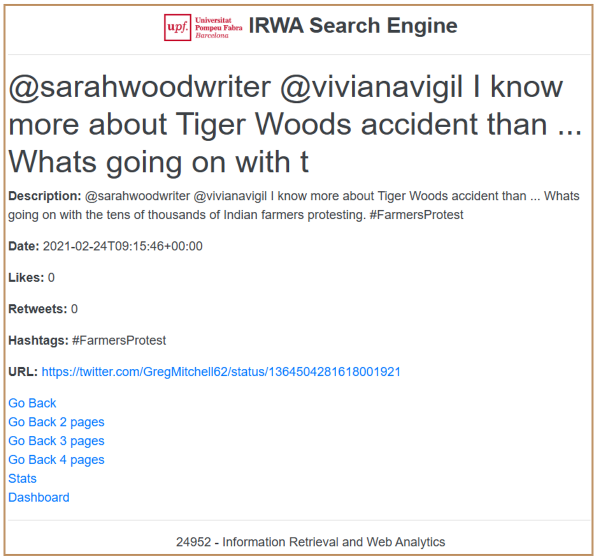
Document ID	Title	Description	Click Count
1364504281618001921	@sarahwoodwriter @vivianavigil I know more about Tiger Woods accident than ... Whats going on with t	@sarahwoodwriter @vivianavigil I know more about Tiger Woods accident than ... Whats going on with the tens of thousands of Indian farmers protesting. #FarmersProtest	1

Search Query Statistics

Query ID	Terms	Number of Terms	Timestamp	Count
farmer protest	['farmer', 'protest']	2	2024-12-04 21:00:23	1

Session Statistics

Session ID	Browser	OS	IP	Country	Start Time
IRWA 2021 home	ChromiumEdge	Windows	127.0.0.1	Spain	2024-12-04 21:00:19



IRWA Search Engine

@sarahwoodwriter @vivianavigil I know more about Tiger Woods accident than ... Whats going on with t

Description: @sarahwoodwriter @vivianavigil I know more about Tiger Woods accident than ... Whats going on with the tens of thousands of Indian farmers protesting. #FarmersProtest

Date: 2021-02-24T09:15:46+00:00

Likes: 0

Retweets: 0

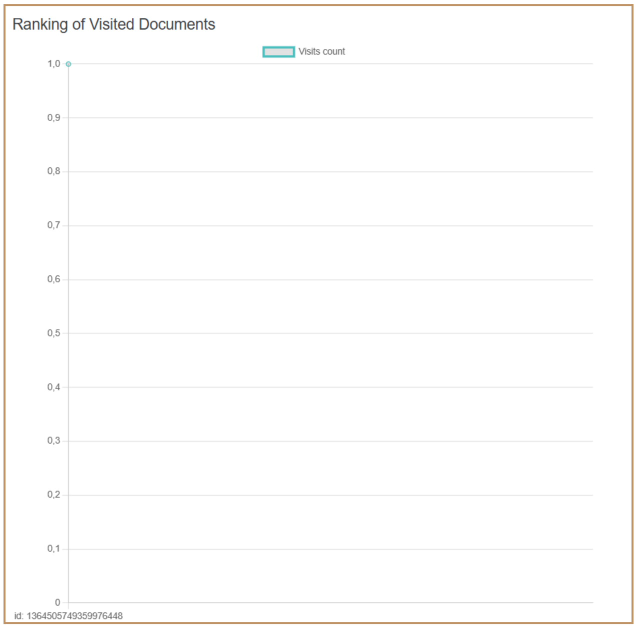
Hashtags: #FarmersProtest

URL: https://twitter.com/GregMitchell62/status/1364504281618001921

24952 - Information Retrieval and Web Analytics

4. Creating the website III (Dashboard):

To finish the web, we were asked to create a dashboard, a more visual way to show the statistics. Thanks to toying with dashboard.html and the section of dashboard in web_app.py we create a graph displaying the number of clicks by tweet and two tables: one for the query data and another one for the session data. These are some results:



5. Challenges:

It has not been an easy lab we have to admit. We are only two members after all. We have been stuck so many times, but finally we can submit more or less everything that was asked for. However, we have to make some notes on what we are lacking:

- The Flask opens correctly and the website goes to index very quickly, however, after writing a query, the search is a bit slower, at least from my computer.
- When clicking on the tweet, the doc_details.html opens correctly, then we click on stats and they are shown with no problem. The problem comes if i were to go back, or to select from the previous tab another tweet, it will update the click counter, but it will no longer show the query and session data. Same thing happens with these last two data in the dashboard.

- Lastly, also in the dashboard, the click graph works correctly but only for an individual tweet. If you try to go to the previous tab and see the dashboard as well (with two x,y entries), it gives an error.