An Analysis on a Query Dataset from Arquivo.pt Search Engine

Technical report, December 2021 (work-in-progress)
Pedro Gomes, pedro.gomes@fccn.pt

Abstract

In this document, we will describe generally the characteristics of the Arquivo.pt search engine with simple characterization statistics of the dataset. This document mainly adds new metrics to the user data analysis in addition to tools existing in our service (i.e., Google Analytics and Awstats) as well as to see their effectiveness and precision in the results.
The goal of this document is to study the searching behaviour of the users in search engines whose focus is on web archive content. The result of this study will be crucial to improving the ranking functions. In this document, we analyzed the query logs from Arquivo.pt, covering three months in 2021, between 1 June 2021 until 31 August 2021.
The document will be divided into three chapters. The first chapter describes the Arquivo.pt search engine. Then, the second chapter describes all the metrics used in this first analysis. Finally, the last chapter will include the conclusions and future work for this document.

Index

Arquivo.pt search engine

The Arquivo.pt is a non-profit service that mainly preserves information published online related to the Portuguese and European community for research and education purposes.

Arquivo.pt have more than 12 billion files collected over 829 TB of information and provides the following resources:

- Comprehensive crawls of the Portuguese Web;
- Crawls of the world web with research content (e.g., European Elections 2019);
- Search by terms, (i.e., page and image search);
- Search by URL with the possibility the see all the versions for the given URL;
- Advanced search (i.e., Advanced page and advanced image search);
- The possibility of automatic computation of the archived data for research purposes through our APIs;
- Supporting researchers with incentives to create new applications (e.g., Arquivo.pt Awards);
- The integration of external resources of importance to the user (e.g., oldweb.today and contamehistorias.pt);

As reference before, Arquivo.pt allows users to do advanced searches. In the Arquivo.pt page advanced search, we can add new terms, search results without a specific term, change dates, and select parameters related with the type of page to carry out a more refined search.

This type of search can be useful when (i) the user does not have the necessary knowledge to make a simple query; or (ii) the user knows exactly what he is looking for and selects a set of parameters to quickly retrieve a certain page or image.

In this document, we will analyse the query logs from Arquivo.pt, covering three months in 2021, between 1 June 2021 until 31 August 2021. The log system from Arquivo.pt has two types of logs:

- Apache Common Log Format, in which each record in the log is an interaction between a user and the search engine through an HTTP request;
- Apache Log4j, which is a Java-based logging utility used to log the intermediate calls to the APIs (i.e., all the calls and parameters from APIs are stored);

Preprocessing step

Since our goal is to understand how users interact with Arquito.pt we create two datasets (a) the query log dataset and (b) the wayback dataset to have a more convenient and faster process.

To minimize the possibility of bots, there are several preprocessing steps done to make the dataset cleaner:

- Remove intern requests (e.g., *10.0.0.3*);
- Remove robots based on the user agent (e.g., *Googlebot*);
- Remove empty user agents (e.g., “-”);

- Remove request from Cloud servers (e.g., IP from AWS);
- Remove empty queries and clicks with any associated query;
- Remove requests made by .jsp (since we only use API now);

We also add more information to better characterize the user:
- Identify the Country, City and Province, and ISP of the request;
- Identify the user's device (i.e., desktop or mobile);

In this document, language-specific lists of stop-words or stemming algorithms were not applied, and all queries were put in lowercase to better show how real users search in Arquivo.pt (i.e., including punctuation and misspelling).

Although it is important to analyze the requests coming from automated processes hosted in the cloud or other platforms, at this moment it will not be our focus.

Challenges

The first challenge relates to the fact that Apache Common Log Format and Apache Log4j do not feature unique user identifiers, instead of containing only cookie identifiers in a fraction of the records, plus information on source IP addresses and user-agents (i.e., identifiers for the type of Web browser in which the query was submitted). In these cases, the logs may feature queries from different users appearing interleaved in chronological order, all associated with the same IP address (e.g., from a common Internet proxy). For instance, the users that work in the same government institution will have the same IP address and also the same user agent.

The second challenge is that the query logs do not give an insight into the user's difficulties at the moment, since it is not possible to catch every interaction or doubt of the user.

Simple Characterization Statistics (Query Log Dataset)

The query log dataset used in the experiments was collected from Arquivo.pt search engine from 1 June 2021 until 31 August 2021 (i.e., three months), containing a total of 35 528 records (i.e., the number of queries and clicks), with 2 696 unique users.

Then, after a preprocessing step, each entry has the following attributes:
- **ip_address**, which consists of a unique string of characters that identifies each computer using the Internet Protocol to communicate over a network;
- **request**, which is the parameter “REQUEST” from apache;
- **user_agent**, which consists in a user agent is a computer program representing a person;
- **trackingid**, consists of a unique value that identifies the user, the search, and the session;
- **timestamp**, the instant when the user submitted a request (i.e., query or click);
- **Year, month, day, hour, minute**, which consists of values generated through the **timestamp** column;
- **type_search**, indicates the type of search;

- **query**, the set of keywords submitted by the user;
- **page**, is the page of the click;
- **maxitems**, is the value of maximum items per SERP page;
- **page_search_response**, the time to return a response from Page Search API;
- **image_search_response**, the time to return a response from Image Search API;
- **page_search_results**, the list of results returned to the user for a given query on page search (only field when there is a click);
- **image_search_results**, the list of results returned to the user for a given query on image search (only field when there is a click);
- **session_id**, the session id from jsp;
- **position**, which is the position clicked by the user;
- **country**, the country based on IP address (https://ipinfo.io/);
- **city**, the city based on IP address (https://ipinfo.io/);
- **isp**, the isp based on IP address (https://ipinfo.io/);
- **province**, the province based on IP address (https://ipinfo.io/);
- **timezone**, the timezone based on IP address (https://ipinfo.io/);
- **hostname**, the hostname based on IP address (https://ipinfo.io/);
- **type_device**, the type of device based on User Agent (https://pypi.org/project/user-agents/);
- **browser_family**, the browser family based on User Agent (https://pypi.org/project/user-agents/);
- **browser_version**, the browser version based on User Agent (https://pypi.org/project/user-agents/);
- **os_family**, the operating system family based on User Agent (https://pypi.org/project/user-agents/);
- **os_version**, the operating system version based on User Agent (https://pypi.org/project/user-agents/);
- **device_family**, the device family based on User Agent (https://pypi.org/project/user-agents/);
- **device_brand**, the device brand based on User Agent (https://pypi.org/project/user-agents/);
- **device_model**, the device model based on User Agent (https://pypi.org/project/user-agents/);
- **absolute position**, the absolute position in the SERPs;

Thus, whenever there is a click on a result the attribute **POSITION** records the position of that click (on the page that the user is viewing). So, the **POSITION** will always be a value between 1 and **MAXITEMS**. **MAXITEMS** is the number of results per page because users can choose to view 10 results per **PAGE** or 24 for example. Second, the user can click on the first **POSITION** on the fifth **PAGE**. Therefore, the **POSITION** will be 1 and the **PAGE** will be 5. So, to be easier I create a new column **ABSOLUTE_POSITION**, which is the clicked position within the results returned by the API with the formula is equals to (**PAGE***MAXITEMS)+**POSITION**.

If **PAGE** is 1, **MAXITEMS** is 24 and **POSITION** is 17, the attribute **ABSOLUTE_POSITION** will be 41. That is, the user clicked on the second page (with 24 results per page) in position 17, which is the fortieth first position (41).

If you need to see an example of real query log data you can look at the following google sheet:

- https://sobre.arquivo.pt/wp-content/uploads/Query-Logs-Sample-Data-sample.csv

Metrics

In this section, we will describe the metrics in which together can bring ways to characterize Arquivo.pt users and help management and developers see what is the performance of Arquivo.pt.

1) Type of search

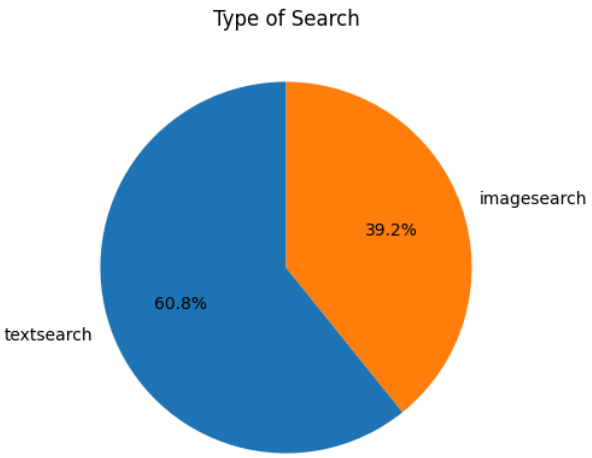

Figure 1

Figure 1 shows the difference between the percentage of queries made from a page search versus the percentage of queries made from an image search. When the user interacts with the Arquivo.pt, 60.8% of the requests are a page search while only 39.2% are an image search, in which the result is expected since our users mainly use Arquivo.pt to search for a particular version from a page.

2) Geografic

One of the big problems with geographic user identification is the high volatility of the IP address assigned. For the same ip, for instance 83.240.225.13 we have different cities:

- Viseu (reported 12/2021)
- Santarém (reported 12/2021)
- Viseu (reported 12/2021)
- Beja (reported 12/2021)

Thus, we will use ipinfo.io, since it is used by ebay and craigslist, which are search based systems.

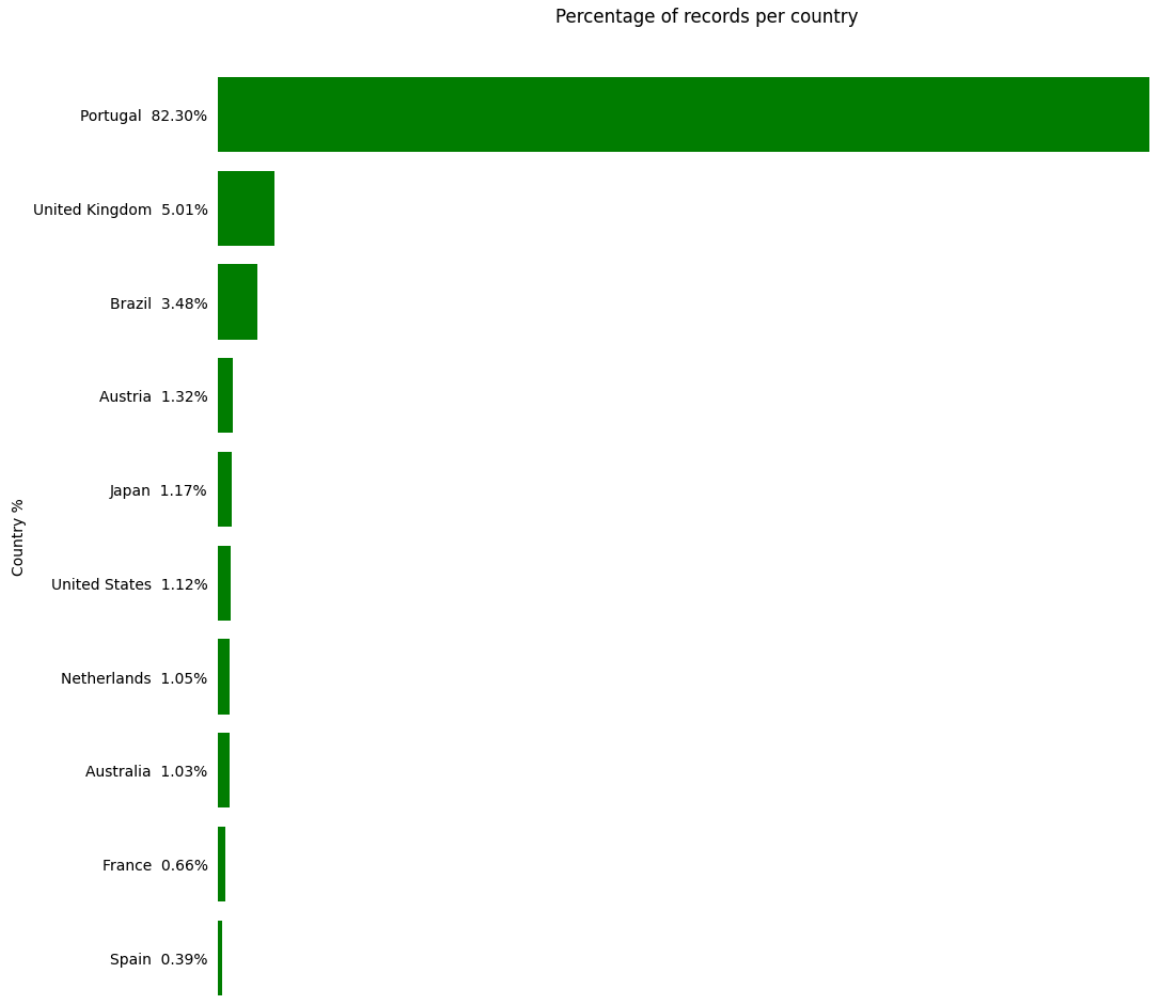

Figure 2.1

Figure 2.1 shows the percentage of requests by country. The countries with the most requests are Portugal with 82.30%, United Kingdom with 5.01%, and Brazil with 3.48%. As expected, most of the requests are made from Portugal. Interestingly, Japan has about 1.17% of the requests, which need to be analyzed.

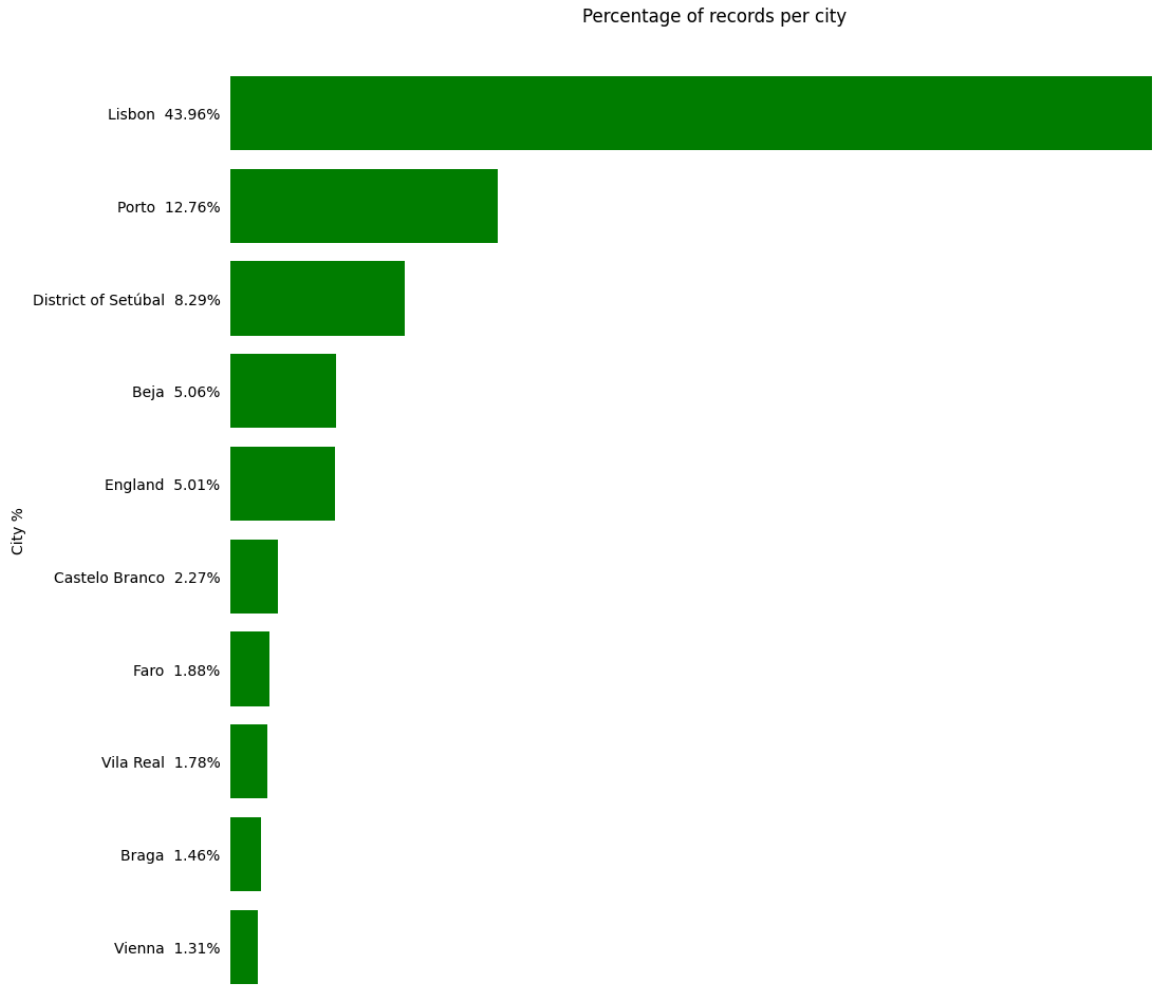

Figure 2.2

Figure 2.2 shows the percentage of requests by city. The cities with the most requests are Lisbon with 43.96%, Porto with 12.76% and Setúbal with 8.29%. Interestingly, Vienna has about 1.31% of the requests.

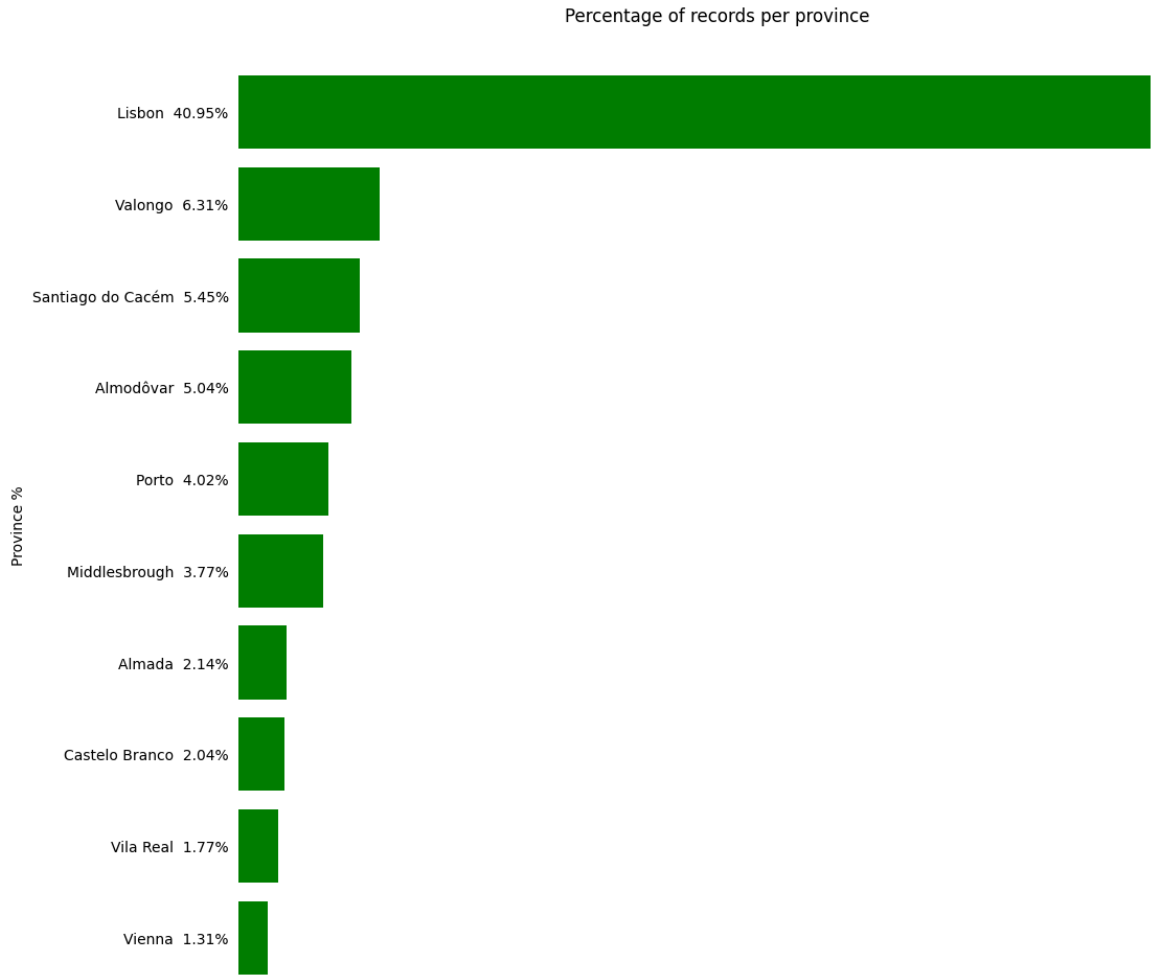

Figure 2.3

Figure 2.3 shows the percentage of requests by province. The provincies with the most requests are Lisbon with 40.95%, Valongo with 6.31%, and Santiago do Cacém with 5.45%. Interestingly, Middlesbrough has about 3.77% of the requests.

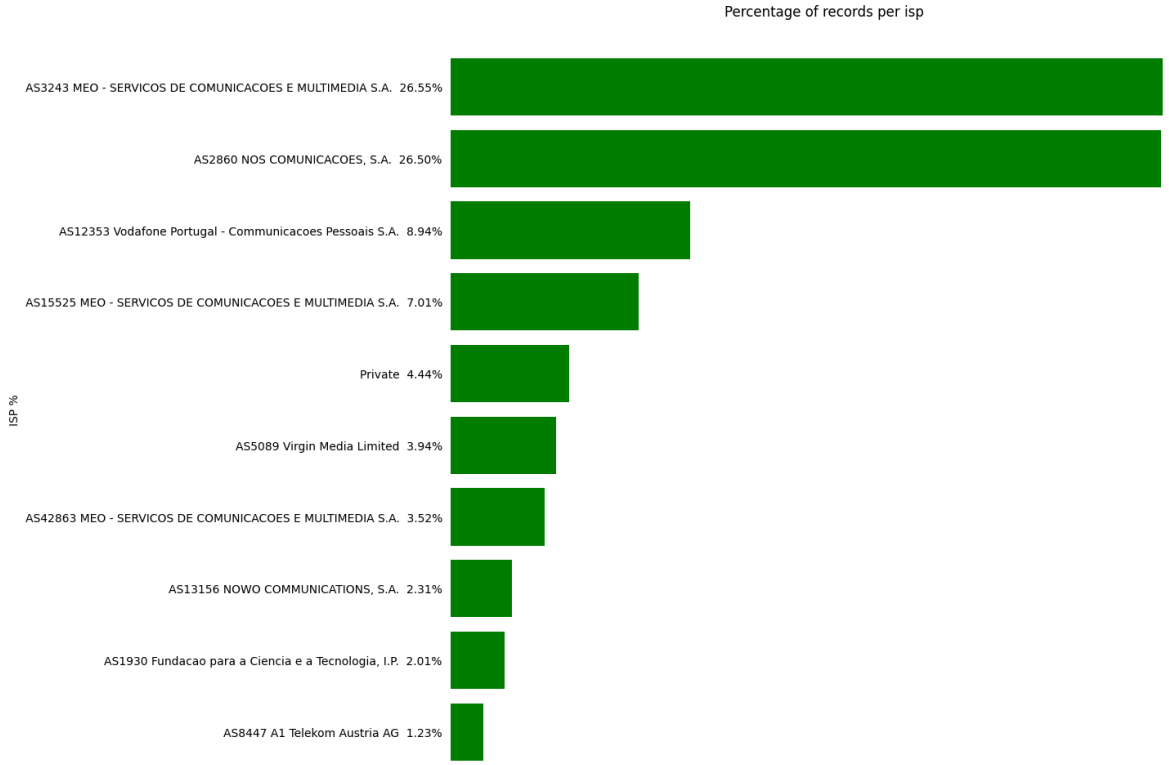

Figure 2.4

Figure 2.4 shows the percentage of requests by ISP (i.e., internet service provider). Based on the results, we will need to remove the request with the ISP equal to “Private” and “Fundacao para a Ciencia e a Tecnologia”. Interestingly, there are three types of ISP for the company “MEO”, which needs to be analyzed.

3) Type of devices

Knowing which devices are most used to access Arquivo.pt will help understand which screen resolutions are most used by users.

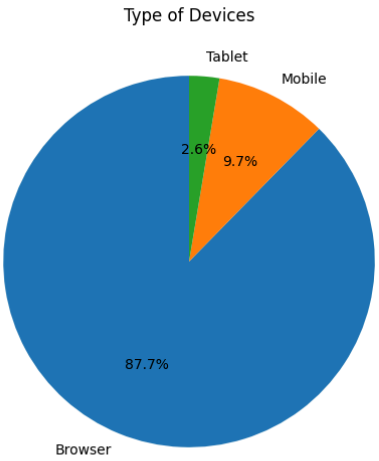

Figure 3.1

Figure 3.1 shows the percentage of requests by type of device. As expected, the device with the most requests is the browser with 87.7%.

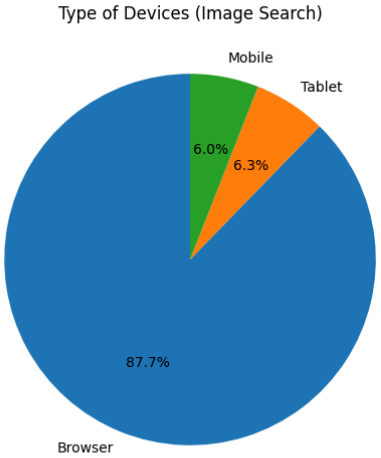

Figure 3.2

Figure 3.2 shows the percentage of requests by type of device (image search). As expected, the device with the most requests is the browser with 87.7%. Interestingly, Tablet has about 6.3% of the requests.

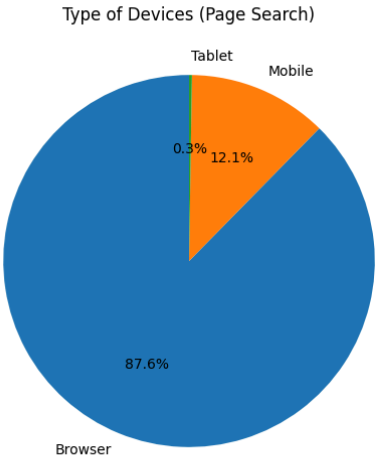

Figure 3.3

Figure 3.3 shows the percentage of requests by type of device (page search). As expected, the device with the most requests is the browser with 87.6%.

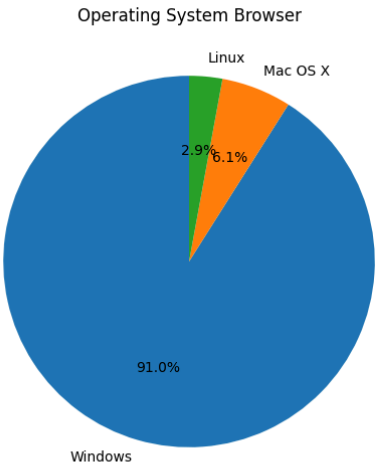

Figure 3.4

Figure 3.4 shows the percentage of requests by the OS (i.e., operating system) on browsers. As expected, the device with the most requests is Windows with 91%.

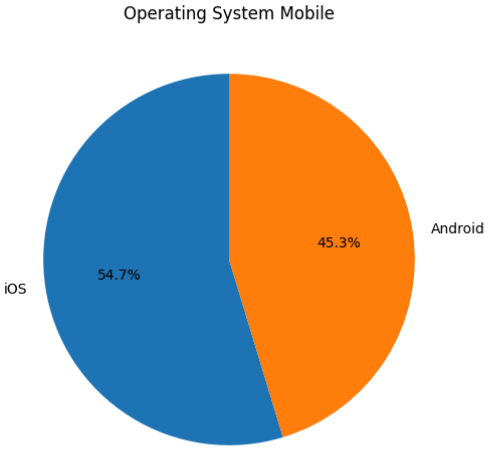

Figure 3.5

Figure 3.5 shows the percentage of requests by the OS (i.e., operating system) on mobile. Interestingly, IOS and Android devices have about the same percentage of requests.

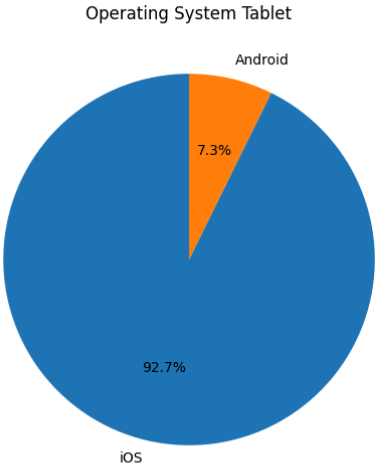

Figure 3.6

Figure 3.6 shows the percentage of requests by the OS (i.e., operating system) on tablet. The devices with the operating system IOS have the most requests with more than 92%.

Based on this section, we can conclude that most users use Windows as an operating system, while mobile users use IOS and Android as operating systems.

4) Absolute Position

The percentage of clicks in the first position can be a good indicator for the quality of the search. If the user clicks on the first result it is because he analyzed the spinnet, the title, and\or the URL, which motivated the user to click on the result. If the results are extremely bad, the users would not click on any result. However, this metric can be influenced by several aspects:

- Different types of users (e.g., beginner or expert);
- Interface changes (e.g., colour scheme or the number of SERPs results);
- Different ranking function;
- New content (e.g., add new collection);

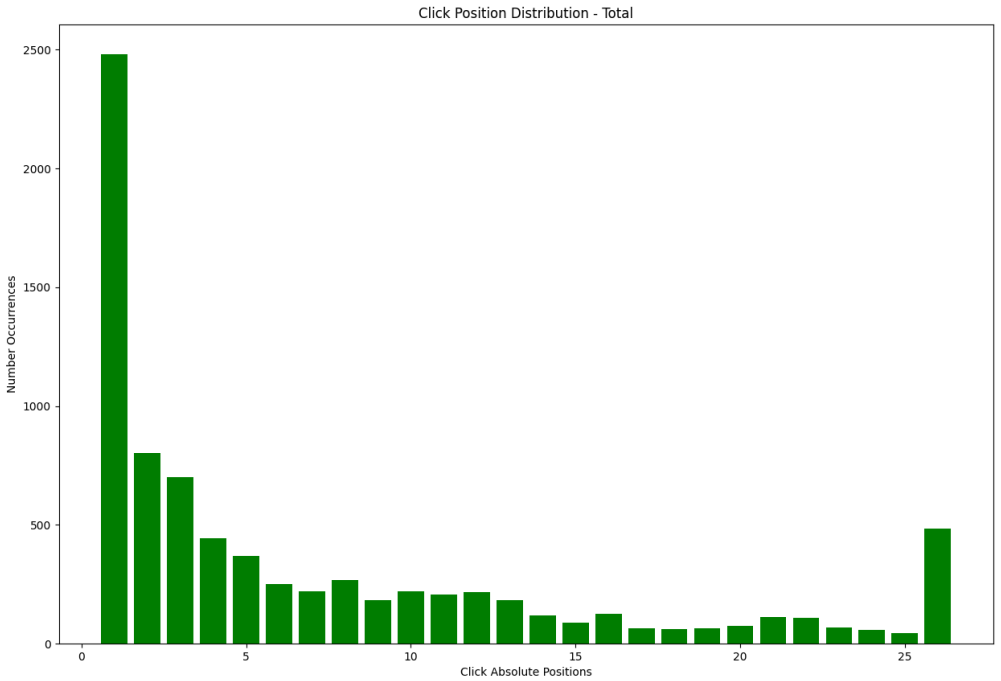

Figure 4.1

Figure 4.1 shows the distribution of positions clicked by the user (page and image search). As expected, users tend to click on the first five positions. The 26 bar reflects the clicks in the 26 position or further.

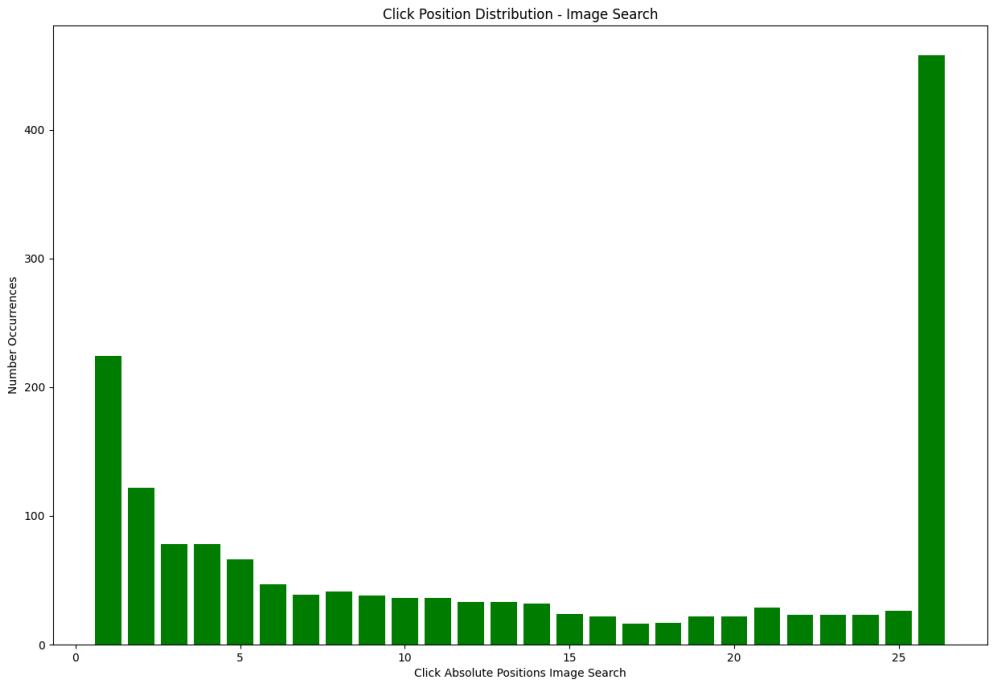

Figure 4.2

Figure 4.2 shows the distribution of position clicked by the user (image search). The 26 bar reflects the clicks in the 26 position or further. Interestingly, users tend to click on results beyond the first page. This problem can occur due to the fact that for some queries there are results with duplicate images.

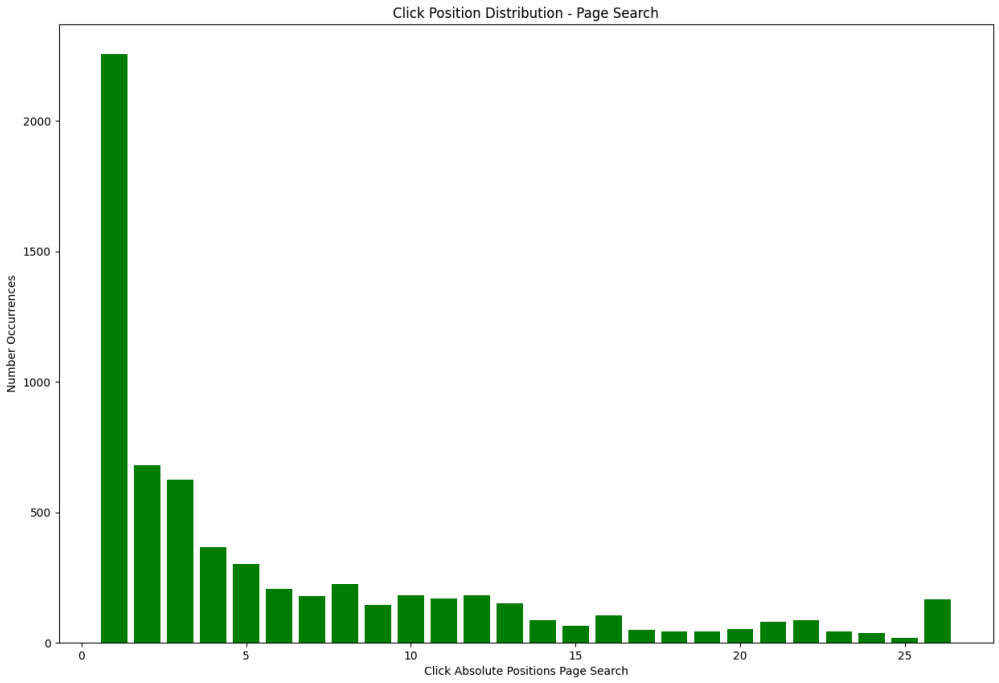

Figure 4.3

Figure 4.3 shows the distribution of position clicked by the user (page search). The 26 bar reflects the clicks in the 26 position or further. As expected, users tend to click on the first five positions.

5) Top Queries

Knowing what users are searching is essential to understand the intentions when using Arquivo.pt. Based on this data, we may improve the themes of our collections, the ranking or adding new features.

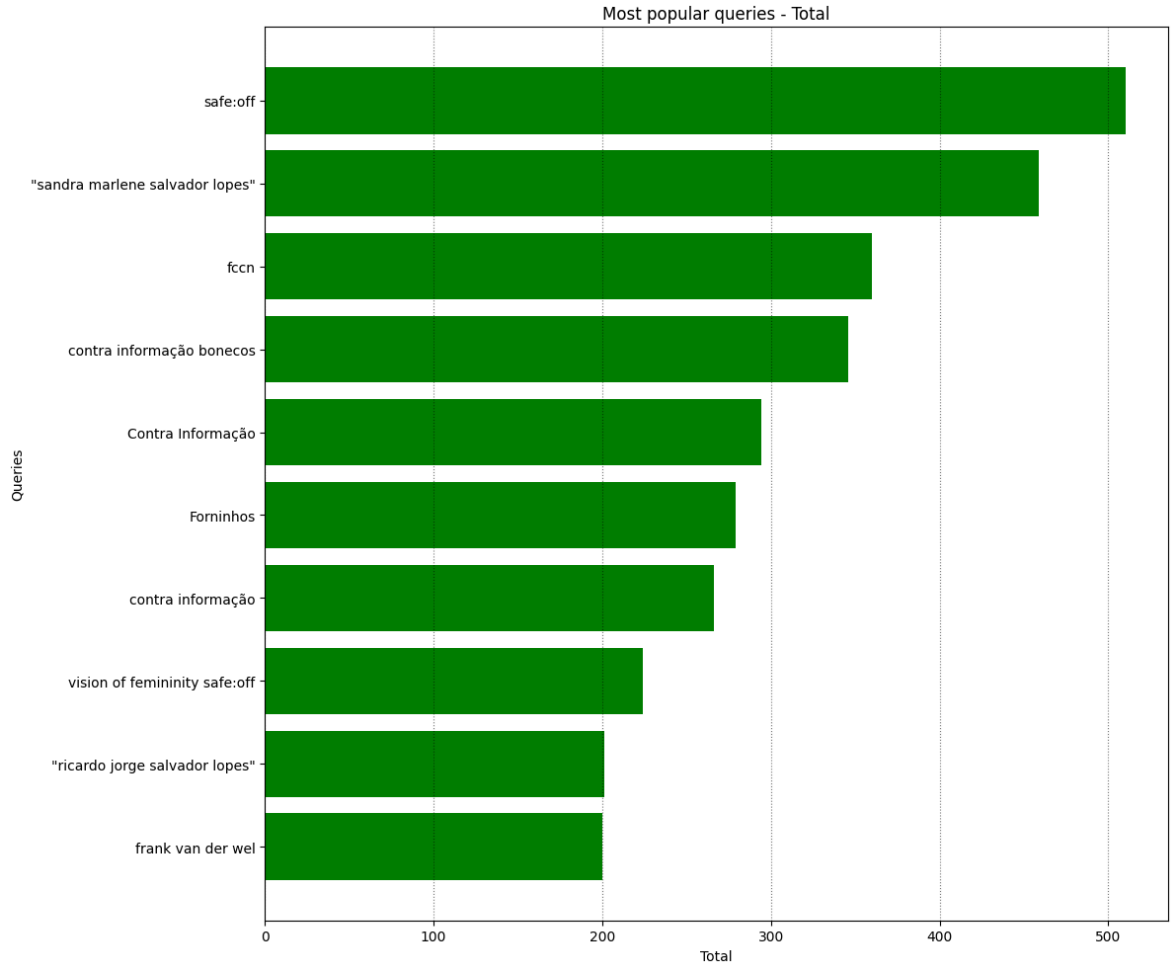

Figure 5.1

Figure 5.1 lists the top 10 queries most popular (page and image search). “safe:off” (i.e., advanced parameter to turn off the NSFW filter) is the most searched query, while “sandra marlene salvador lopes” (i.e., personal name) is the second. As expected, people tend to search for their own name in Arquivo.pt. The next step is to extract the top 100 most popular queries to do a more detailed analysis of the topic search in Arquivo.pt.

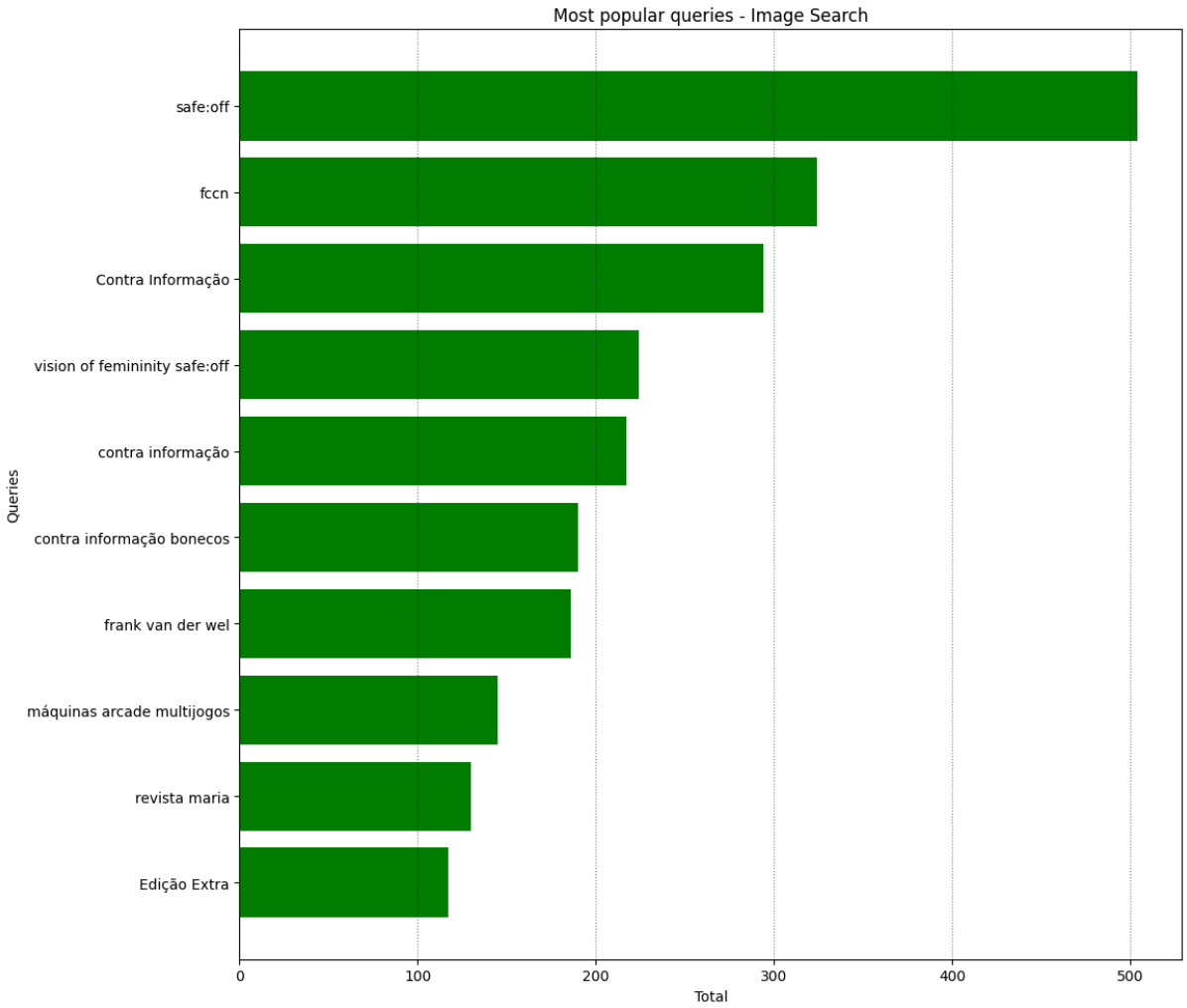

Figure 5.2

Figure 5.2 lists the top 10 queries most popular on image search. “safe:off” (i.e., advanced parameter to turn off the NSFW filter) is the most searched query, while “fccn” is the second. Interestingly, there were few searches with proper names.

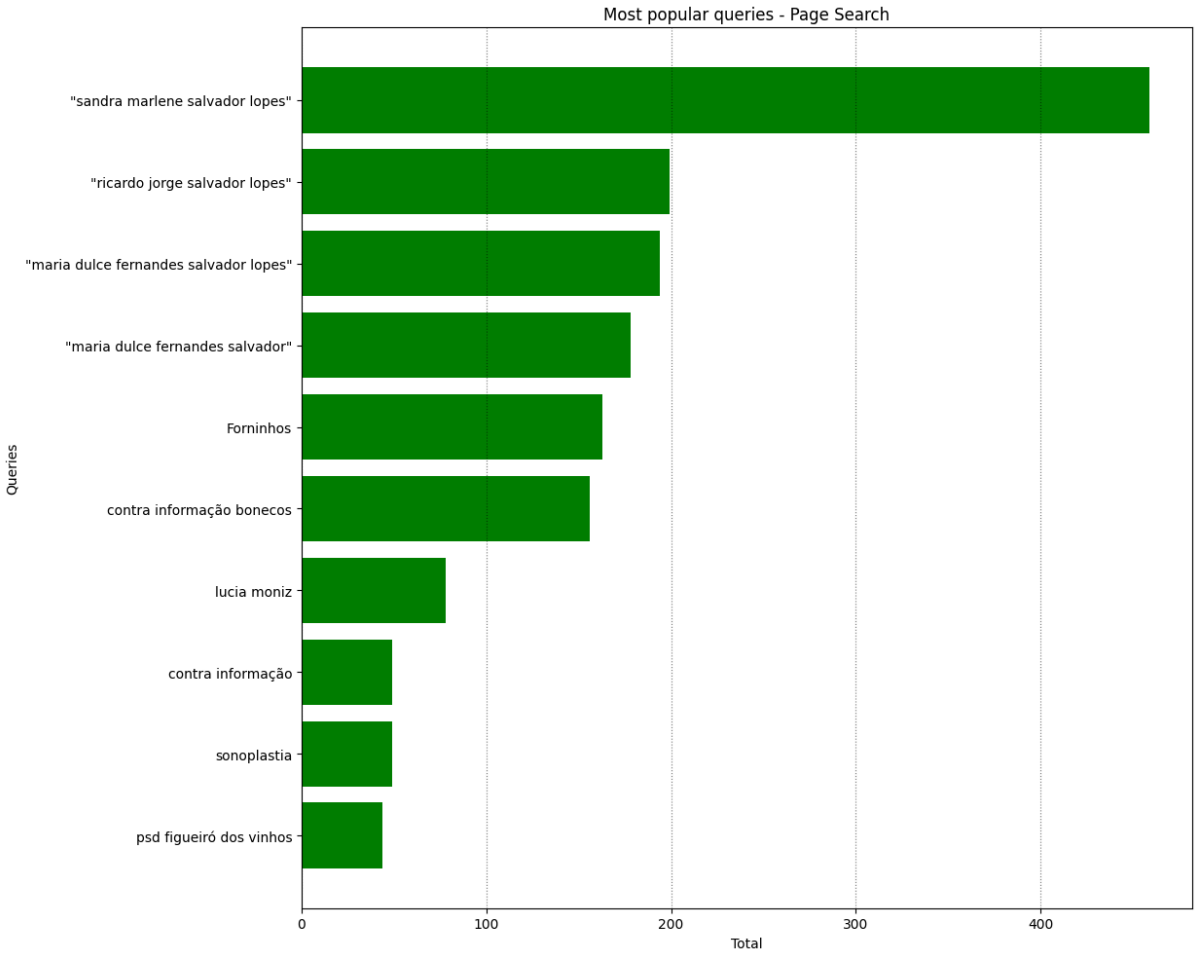

Figure 5.3

Figure 5.3 shows the most popular queries (page search). “Sandra marlene salvador lopes” is the most searched query, while “ricardo jorge salvador lopes” is the second. These results are interesting because they are made by the same set of people (i.e., probably by family members).

6) Number of unique users per month

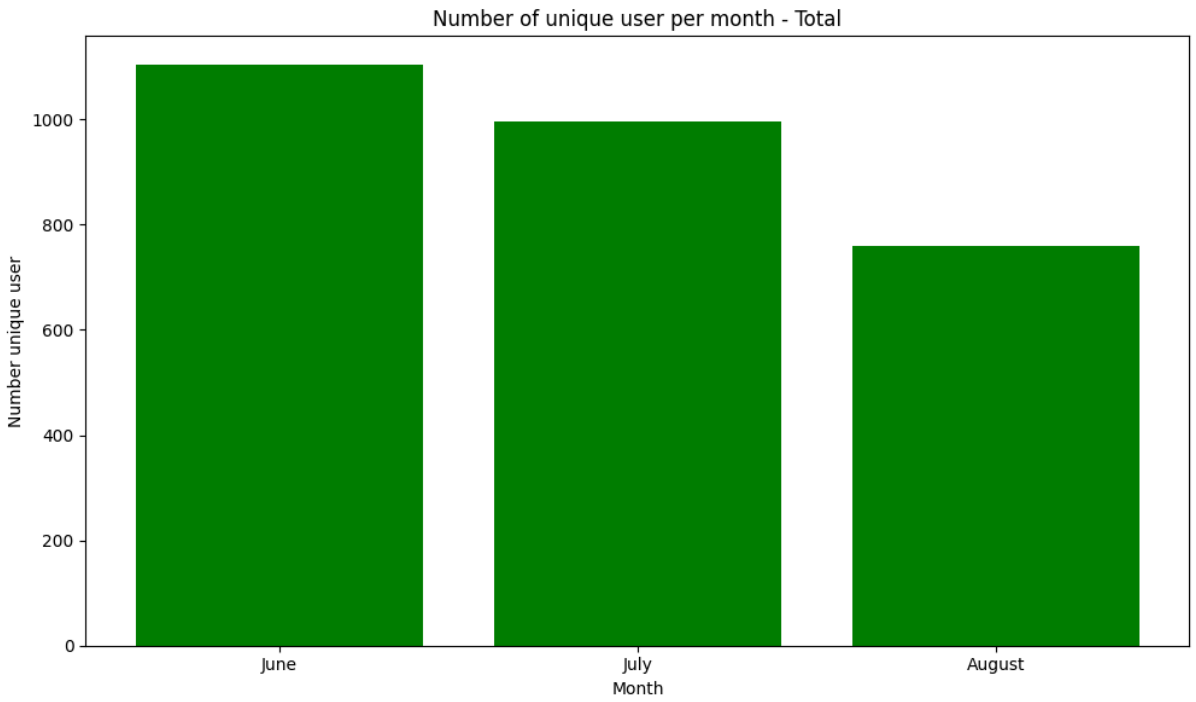

Figure 6.1

Figure 6.1 shows the number of unique users per month. As can be observed, the number of queries is decreasing. One of the reasons for this event is the fact that these three months are the vacation season in Portugal since it represents 82.30% of the requests.

7) Number of unique queries per month

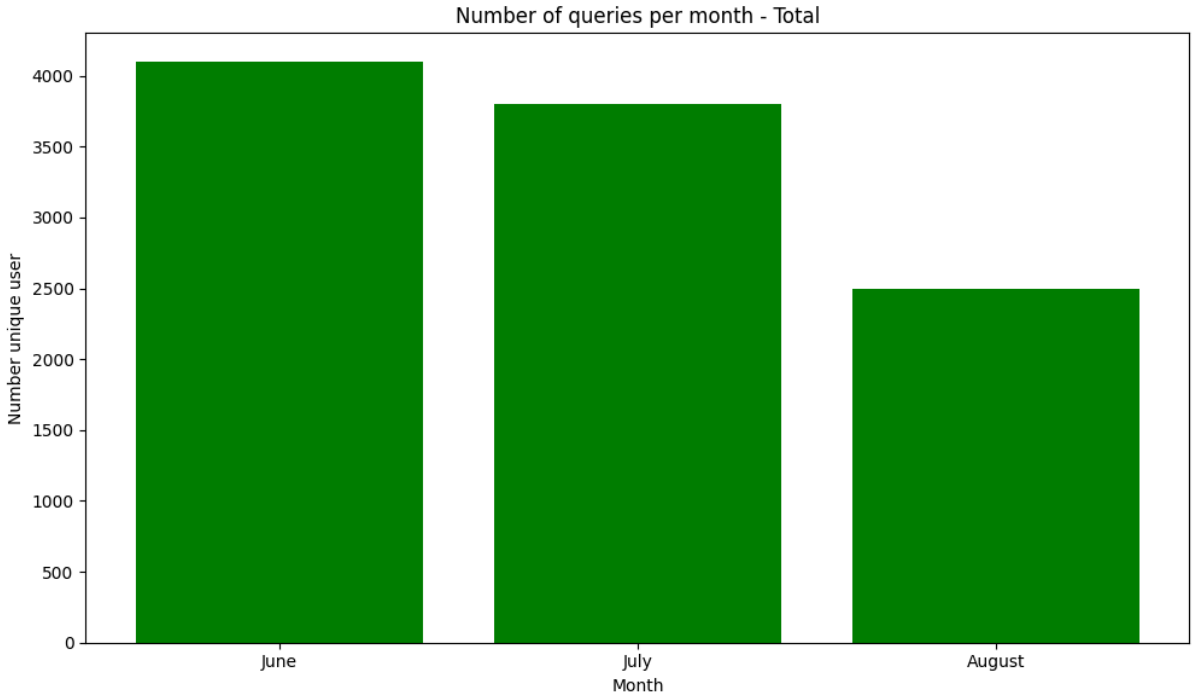

Figure 7.1

Figure 7.1 shows the number of unique queries per month. As can be observed, the number of queries is decreasing, in which it is aligned with the number of unique users per month.

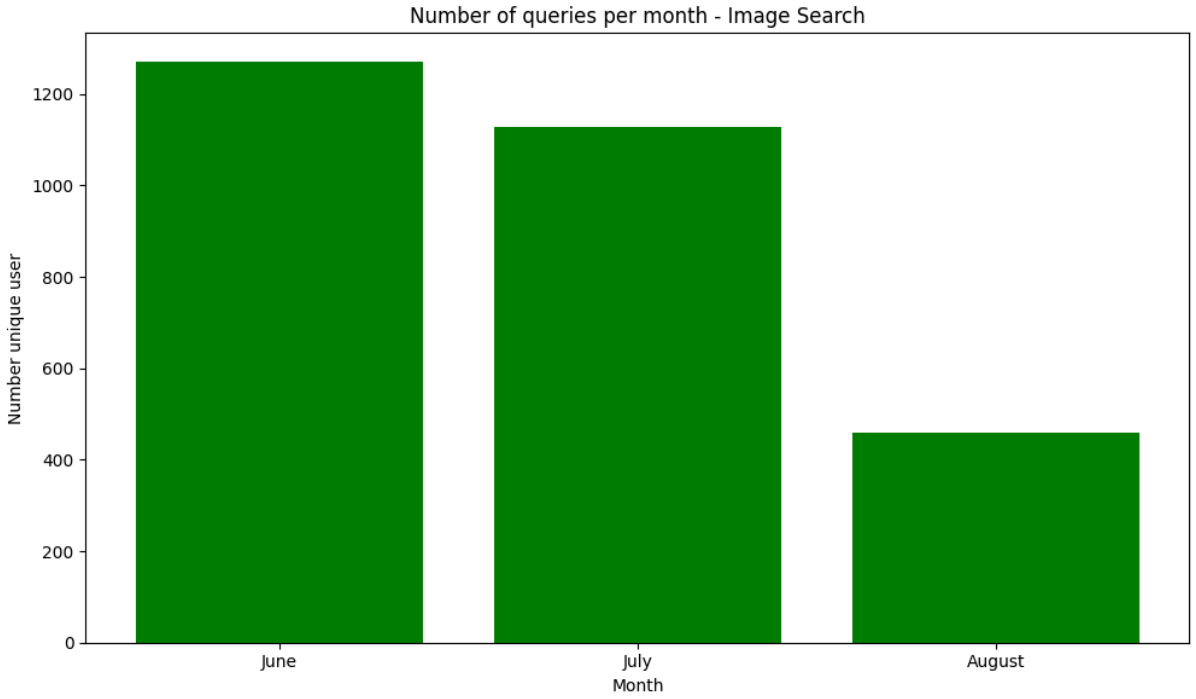

Figure 7.2

Figure 7.2 shows the number of unique queries per month (image search). As can be observed, the number of queries is decreasing. However, the decrease in the number of queries from July to August is much more pronounced than Figure 7.1.

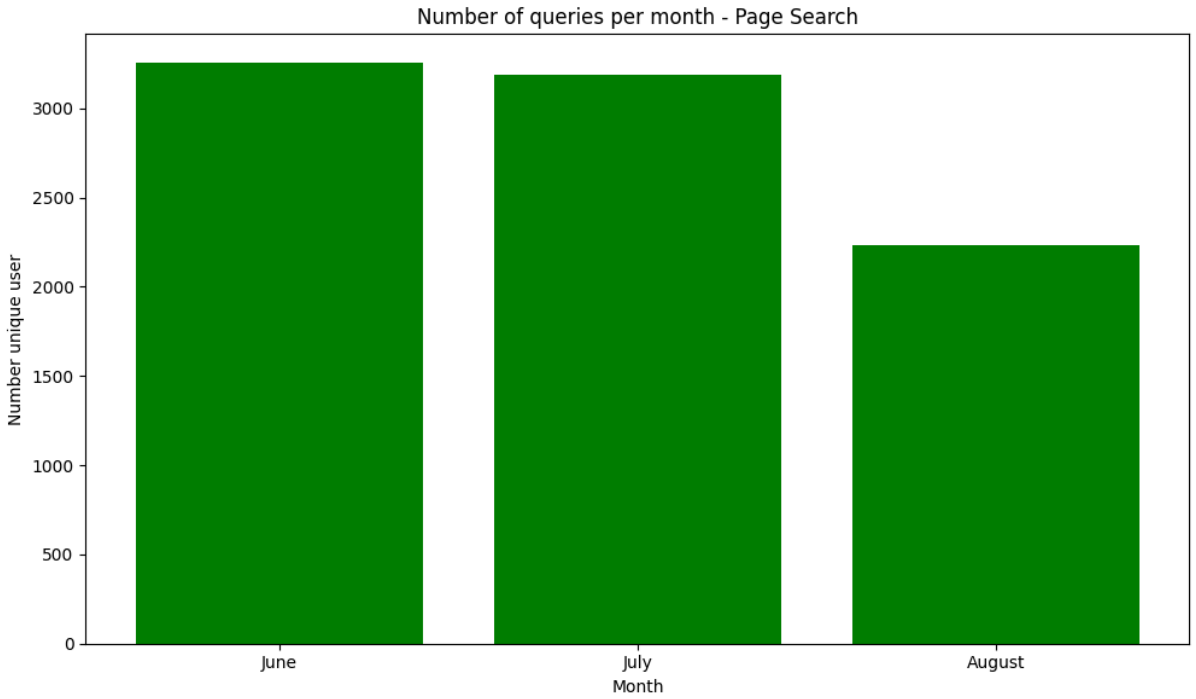

Figure 7.3

Figure 7.3 shows the number of unique queries per month (page search). As can be observed, the number of queries is decreasing. However, the difference between June and July is not as pronounced as in the image search.

8) Response Time (APIs)

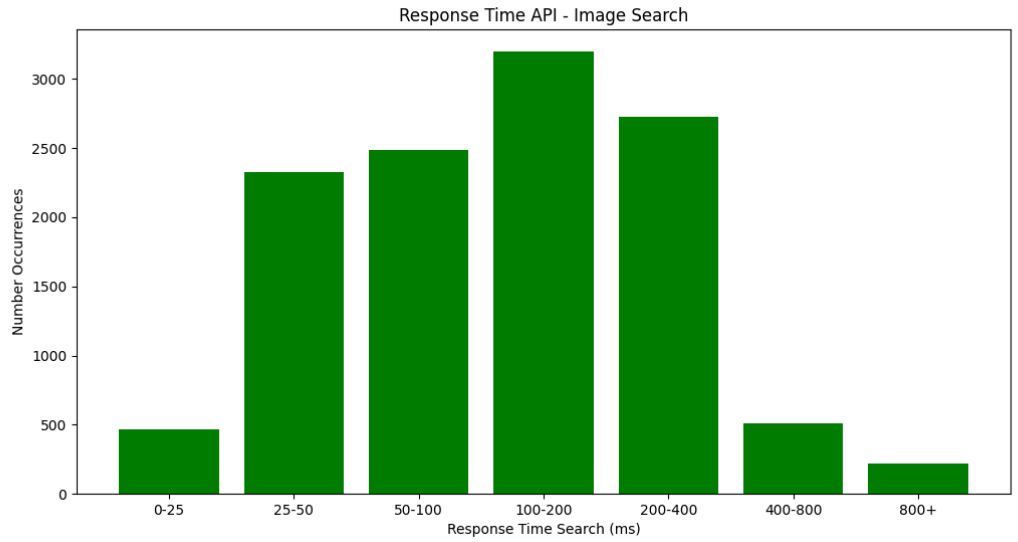

Figure 8.1

Figure 8.1 shows the response time of the API (image search). In this case, the response values are acceptable for a response to queries that can be heavy.

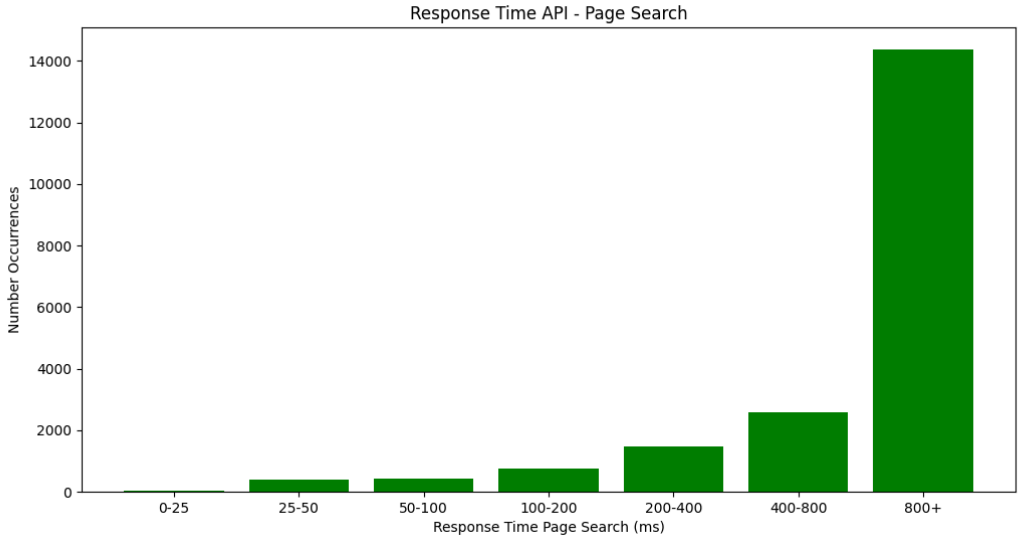

Figure 8.2

Figure 8.2 shows the response time of the API (page search). In this case, the response values are not acceptable for a query response, which shows a big difference between technologies for page search vs image search (i.e., Nutchwax vs SOLR). It will be necessary to better evaluate the necessity of an upgrade.

Future work

To improve the work done, in addition to the points covered in each section, the following advancement will be needed:

1. You will need to perform the study for a larger query log dataset:
    a. “””Retrieving logs from a large time range can bring several advantages into the analysis of the logs: (i) it is possible to see the evolution of users (ii) it allows to see if there is a particular seasonal search pattern in the logs (e.g., when the President of the Republic approves the state budget); and (iii) the analysis will be less affected by bias. “””
2. Create a mechanism that generates the views in real-time.
3. Improve our methods to detect bots and geolocation:
    a. https://github.com/monperrus/crawler-user-agents;
    b. https://www.abstractapi.com/guides/how-to-geolocate-an-ip-address-in-python;
    c. Can we change this website isbot.js.org to python?
4. Compare if the number of queries rises relative to other metrics (e.g., number of users). For instance, if the number of users does not increase relative to the increasing number of queries, maybe the user has difficulties doing a specific task and needs to do more queries or the user is spending more time in Arquivo.pt. So, this metric needs to be correlative with other metrics. We also need to have in mind that increasing the number of queries and decreasing the number of users does not always mean that users are doing more query reformations.
5. Make a study of the most used functionalities and parameters (page and image search).
6. Percentage of clicks on the query suggestion. However, there are some possibilities that we need to keep in mind:
    a. Does the person do the search and then clicks on the query suggestion?
    b. Or every time they click on a query suggestion it is a hit? even if it is on page 10? (Went through the page and only then saw that it was better to click on the suggestion).
7. Compare the results with other search engine:
    a. https://www.internetlivestats.com/google-search-statistics/
    b. https://www.wordstream.com/blog/ws/2019/02/07/google-search-statistics
8. Collect and analyze data from the references for Arquivo.pt. However, there may be some problems: https://twitter.com/dcgomes77/status/1442532297543147523 will log the following “refer”:
    a. https://t.co/F3rUuXoug7?amp=1.
9. Realize the number of times there are changes between image and page search. Every time we change from page search to image search the "seach_id" of the tracking ID changes (e.g., 1b0983abf0438ac2b439_aec61d47c30f2d78d693 → 1b0983abf0438ac2b439_6b9f7f7f898fcd676b22).
10. Register when the user is on the replay page and clicks on a different version.

    a. We can use the “/wayback/” request, in which the front-end needs to call again the “/wayback” (however, can be heavy).
    b. We can add a new path like “/wayback/view” to do the logging.

Decisions

1. We will remove entries with the following ISPs:
    a. Private
    b. Microsoft
    c. Facebook
    d. Google
    e. Amazon
    f. Fundacao para a Ciencia e a Tecnologia
2. Entries using search.jsp do not have “tracking_id”. So, it will be removed.
3. Queries with the same “tracking_Id” and “request” are removed, because they are triggered by the refresh of the page.
4. If we have three queries in a row from the same user where the first and the last have the Session_ID, “1234”, and the middle one has no Session_ID, then the middle one also has the Session_ID, “1234”?