Wildfires in Portugal

Milestone 1: Data Preparation

Pedro Angélico
FEUP, M.EIC
up202108866@up.pt

Sofia Pinto
FEUP, M.EIC
up202108682@up.pt

Tomás Gaspar
FEUP, M.EIC
up202108828@up.pt

Tomás Palma
FEUP, M.EIC
up202108880@up.pt

Abstract

Portuguese wildfires leave a trail of destruction every year. Our goal is to condense all the news related to these events in a single place, plus link them to the particular region where they occurred. For this purpose, we collected several news articles from the web, totalling up, for now, 3509 links. However, only 1914 of these ended up being useful. Our pipeline is divided into steps in order to ensure that the created documents precisely match the topic of wildfires. Firstly, we retrieved multiple links from various sources. Then, we transformed some of them into documents following a predefined structure. Lastly, we filtered them by noticing if any Portuguese municipalities or parishes were mentioned, plus if the topic of the article was truly a wildfire. This dataset will serve as the foundation for building a search engine designed to extract multiple insights from its data and hopefully help in the prevention of future strikes.

Keywords

Wildfires, fires, information retrieval, data pipeline, data processing, data characterization

1 Introduction

Every year, Portugal is victim of multiple harsh wildfires that destroy ecosystems and harm the safety of the people. With the intent of having a better understanding about the history of wildfires, we decided to create a search engine that would link various news from the Portuguese media to the region where the advertised fire happened.

2 Data Pipeline

Our pipeline consists of multiple tasks, each split into separate scripts to ensure that elements with distinct responsibilities remain independent. One is responsible for extracting the links from the data sources (*linkGetter.py*), another fetches information about the local municipalities from the public Wikipedia API [10] (*wikiPlaces.py*), a third is in charge of scraping the corresponding websites to the links (*linksToDocs.py*), and, finally, another filters the documents based on the relevance of the terms present in the content (*filter.py*).

To further illustrate this process, we created the diagram in Figure 1, which depicts what we will detail in the following subsections.

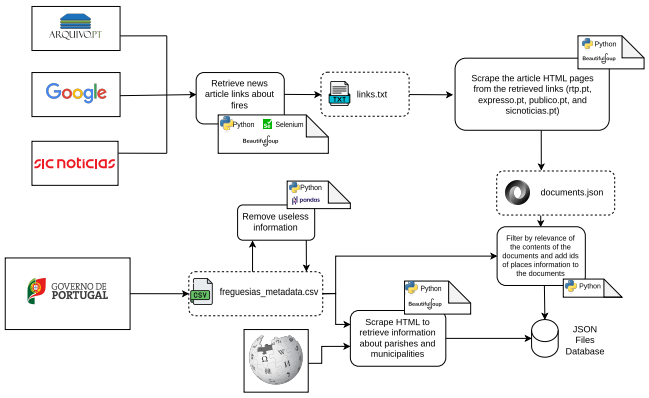

Figure 1: Pipeline Diagram

2.1 Data Sources and Collection

We collected data from three different sources, each presenting its own challenges. This process resulted in a dataset comprising news articles about wildfires in Portugal, alongside some unrelated articles that were later removed using a filtering method (see Data Processing section for details).

Firstly, we used the Arquivo.pt API [3] in order to extract links to articles from Portuguese news websites, including Expresso [4], Público [7], and RTP [8]. This information is made available under the GNU General Public License [5]. Instead of relying on a single search to meet our criteria, we made several searches using a set of keywords related to fires within the time range of 1991 to 2024. The keywords included: "incêndio" (fire), "incêndio florestal" (wildfire), "fogo" (fire), "queimada" (burnt land), "ardeu" (burnt), and "ardidos" (burnt area). It is important to mention that Arquivo.pt [1] stores URLs as they were scraped, meaning some older links, particularly from RTP [8] (due to changes in their URL structure), were no longer accessible. We addressed this issue by converting the outdated links to their current formats. We extracted a total of 1806 links.

Secondly, we extracted articles' links from Google News [6] using web scraping techniques by automating the search and pagination process to extract the HTML content. After that, Beautiful Soup [2] was used to parse the HTML and extract 60 unique links.

Thirdly, SIC Notícias [9] offered a section for wildfire-related news. The articles' links were obtained via requests to its public API by iterating over the page results. This website generated 1703 links.

PRI 2024, October 2024, Porto, Portugal
2024.

Although Google News [6] holds copyright licensing agreements, the news' links accessed through this technique forbid copies of their information. The same happens with SIC Notícias [9]. In order to solve this issue, we are planning to contact these companies as soon as possible, as explained in Prospective Search Tasks.

Finally, we joined all the links in order to delete repeated results, ending up with 3509 links. The following process includes creating the documents from these articles and filtering them with the purpose of keeping only news about Portuguese wildfires.

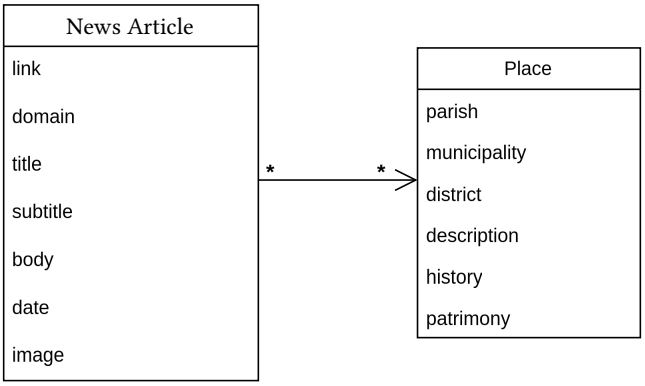

Figure 2: Conceptual Data Model

2.2 Data Processing

Given the links from Sic Notícias [9], Expresso [4], RTP [8], and Público [7], we coded a scraper for each of the news websites to create the wildfire documents with the relevant information. We detail the structure of documents in the Data Characterization section.

As the scraping operation is time-consuming, it saves progress incrementally to a file, allowing the process to resume from any point in the event of an error, such as a network failure.

Additionally, using a CSV file from the Portuguese government's public database, listing all of the districts, municipalities, and parishes in Portugal, we used Wikipedia's API [10] to gather information about each municipality and parish, and created documents for them.

After that, we applied a filter to retrieve only the documents related to wildfires in Portugal. With respect to this process, to ensure we analyse news strictly about Portugal, the documents are first filtered to check if their contents mention one or more places in Portugal. If no mentions to places in Portugal are found, then the document is discarded, otherwise a list of identifiers of the places mentioned is added to the document.

Finally, we differentiated documents mentioning wildfires and other types of fires, by assigning a positive score to words related to the first type and a negative score to the second. If the final score is greater than zero the document is kept, otherwise it is deleted.

3 Data Characterization

We have two types of collections. The first consists of news about wildfires, and the second one contains information about Portuguese municipalities and parishes. These are represented in the conceptual data model, below (Figure 2).

Each document in the news collection is composed by the following fields: "link" (text), "domain" (text), "title" (text), "subtitle" (text), "body", "date" (in ISO Date Time Format), "image" (text), and "places" (list of text). The field "places" has IDs that reference documents from both counties and parishes collections.

On the other hand, each document in the places collection contains the following fields: "municipality" (text), "district" (text), "description" (text), "history" (text), "patrimony" (text), and "parish". The latter must be *null* if the document corresponds to a "municipality" or *text* if it is, in fact, a parish.

We will focus our data characterisation study on the first type of documents. We managed to gather 3509 links from our sources, from which only 2522 could be turned into documents. Additionally, we filtered these results in order to achieve a collection related solely to Portuguese wildfires and, therefore, reduced the number of documents to 1914. Moreover, we opted to analyse the distribution of our results by domain across the three different stages of our pipeline.

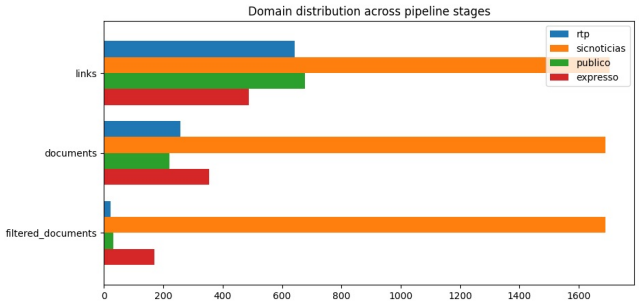

Figure 3: Domain distribution across pipeline stages

From Figure 3, we can see that SIC Notícias [9] was the most reliable, as it had a dedicated section to wildfires. Other results, mostly coming from Arquivo.pt API [3], had less valuable information compared to SIC Notícias [9], due to their search engine's ranking system.

These news articles span the period from 2004 to 2024, as can be seen in Figure 4. In this histogram, there is a clear upward trend in

the number of articles over time, mostly due to the challenges we faced while scraping older news from Arquivo.pt API [3].

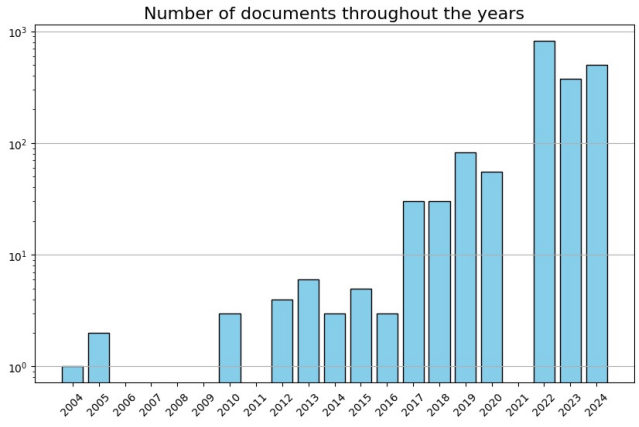

Figure 4: Number of documents throughout the years

In Figures 5 and 6, we present a list of the ten most common words in both the title and the body of the news. As expected, they are mostly terms related to fires ("incêndio" (wildfire), "fogo" (fire)) or to the means applied to fight them ("bombeiros" (firefighters), "meios" (means), "proteção" (protection)).

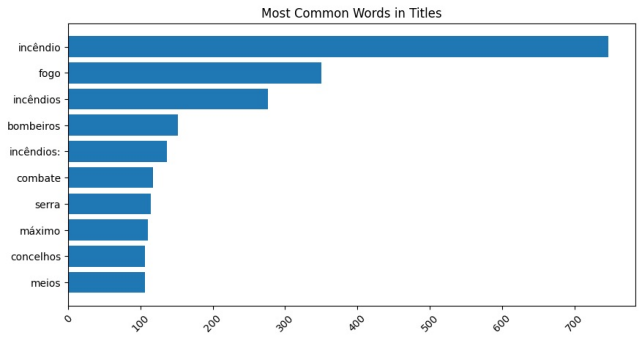

Figure 5: Most Common Words in Titles

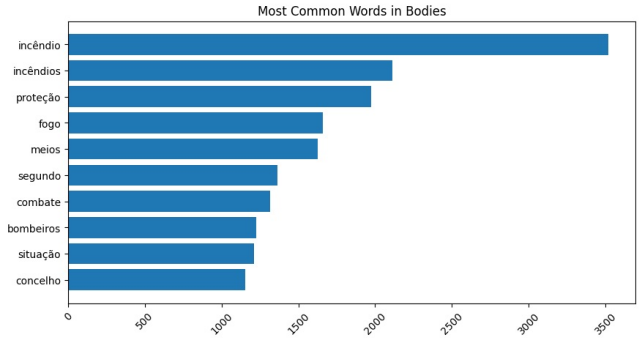

Figure 6: Most Common Words in Bodies

4 Prospective Search Tasks

This study aims to satisfy information needs on several key aspects of wildfires, namely:

- Infrastructure destroyed by wildfires - this is crucial to understand the extent of the damages caused, plus the economic and social impacts.
- Wildfires leading to displacement of people.
- Preventive measures - this is a cornerstone in combating wildfires. We would like to know information either about mandatory or recommended measures.
- Investigations of suspected arson - this is most often rumoured to be the main cause of wildfires in Portugal, and these investigations are key to holding responsible parties accountable and preventing future wildfires.

Furthermore, we plan to reach out to news organisations to request authorisation for using their contents for research purposes, ensuring full compliance with intellectual property regulations.

5 Conclusion

In conclusion, we have detailed the process of collecting news about wildfires in Portugal. By extracting links and scraping several Portuguese news websites, applying a filter to isolate wildfire-related content, and enriching the dataset with location data, we produced a comprehensive collection of documents.

For now, our collection holds close to 2000 documents, which we are expecting to be able to increase. These news articles cover a 20-year period, from 2004 to 2024. Due to our filtering system, we have refined the collection to ensure the documents precisely match our criteria. This is evident from the strong relevance of the keywords from both the titles and the bodies, which are closely tied to the topic of fires.

This dataset will be the basis for developing a search engine aimed at extracting critical insights from the data. Such a tool would be fundamental in understanding wildfire patterns and supporting future strategies for wildfire management and prevention.

References

[1] Arquivo.pt 2024. https://arquivo.pt/.
[2] BeautifulSoup 2024. BeautifulSoup. https://pypi.org/project/beautifulsoup4/.
[3] Vasco Rato André Mourão Daniel Bicho Vítor Gouveia Daniel Gomes, Francisco Esteveira. 2024. Arquivo.pt API. https://github.com/arquivo/pwa-technologies/wiki/Arquivo.pt-API.
[4] Expresso 2024. https://expresso.pt/.
[5] GNU. 2023. GNU General Public License. https://www.gnu.org/licenses/gpl-3.0.html.
[6] Google News 2024. https://google.com.
[7] Público 2024. https://www.publico.pt/.
[8] RTP 2024. https://www.rtp.pt/.
[9] Sic Notícias 2024. https://sicnoticias.pt/.
[10] Victor Vasiliev Bryan Tong Minh Sam Reed Brad Jorsch Yuri Astrakhan, Roan Kattouw. [n. d.]. Wikipedia API. https://pt.wikipedia.org/w/api.php.