Data Preparation - News articles

Pedro Galvão, Rodrigo Reis, Tiago Alves

Faculty of Engineering, University of Porto
Master in Informatics and Computing Engineering
Information Processing and Retrieval
Prof. João Damas

November 18, 2021

Abstract

In the last few years, the growing amount of news keeps it impossible to be completely updated of everything, being necessary to have tools which can process large amounts of data and search accurately. This project deals with the collection, refinement and respective analysis of the news website SAPO24. Firstly, an initial research was made in order to find the best Portuguese news data set. The group decided it would be more interesting to scrap the news instead of choosing a data set, since the news were much more relevant in this website. After understanding the data using Python and different libraries, it was necessary to clean it, remove null values, duplicates between other actions. Afterwords, it was necessary to do some Data characterization, creating some graphs of the most relevant information, use NLP to better analyse the text and create some statistics. A pipeline of the whole process was made in order to easily understand the whole process of handling the news articles from SAPO24 website.

Keywords

news, article, data, SAPO24, data refinement, information retrieval, information system, scrapping, natural language processing, data characterization

1 Introduction

The SAPO24 news website is one of the most relevant Portuguese sources of news, being a website which keeps the user updated of the most important news, whilst keeping the user secure of fake ones. However, due to the ever growing number of news which appear everyday, it is necessary to be able to process large amounts of information and to search accurately in order to access the most relevant ones. The aim of this project is to develop and information system, including the data collection of news data, preparation, information querying, retrieval and finally a functional efficient search system. Firstly we will present the general pipeline, giving an overview of the whole process of handling the data, followed by thoroughly description of how the data was obtained in the Data Collection section. Secondly, the conceptual model will be explained and in Section 5 the process of refining the data, such as removing duplicates and null values will be described, followed by how the data was stored. Then it will be shown how the data was characterized and the some analysis and graphs of the most relevant data will be presented. Finally in section 7 some possible queries which the search system will be able to handle are presented, followed by the conclusion.

2 Data Preparation Pipeline

Figure 1 depicts a representation of our pipeline. In order to set our goals and all the steps of our project we create a pipeline, describing each main component of this milestone:

- **Data Collection**, in which we collect data from the SAPO24 website through scrapping.
- **Data Cleaning**, to remove duplicates and invalid data, and change the format of some attributes.
- **Data Storage**, using a csv file.
- **Data Exploration**, performing quantitative analysis on its attributes.

In the following sections we describe in detail each of these steps.

3 Data Collection

To obtain data from the website we make HTTP requests using Python and we scrap everything with the with Beautiful Soup library.

We begin by making a request to to one of the news page that contain a list of news in the website. From this page we extract the URL to a set of article pages, each one containing exactly one article. Afterwords we make requests to each of these URL's and extract the data we need and write it in one row of a csv file. Finally we proceed to the next page in website and

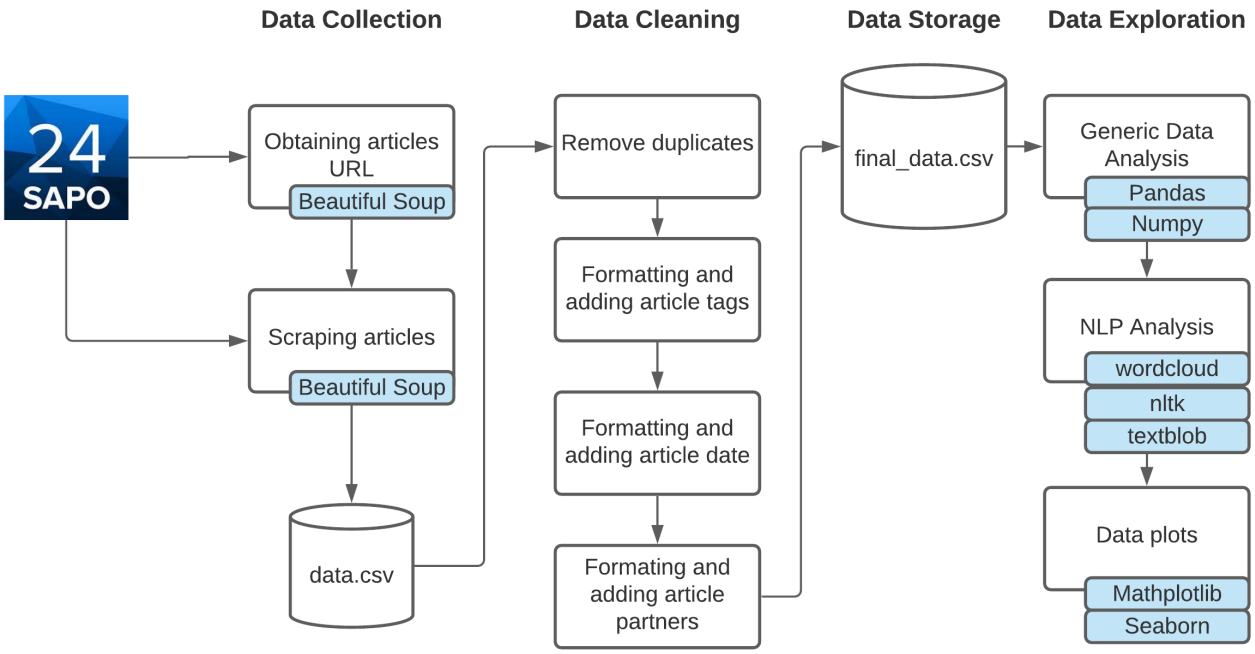

Figure 1: Pipeline

repeat all the process. Scrapping from SAPO24 is not a difficult task because we can indicate the news page as an attribute in the URL, however lots of news are added every day. So, if we want to keep our project always updated in the future we need to implement a process that gets data periodically and compares it with the existing data in our database. To handle that issue, we create an hashed id based on the article title, that allows a process that is always scraping compare the scrapped article id and verify if that exists in a database.

4 Data Model

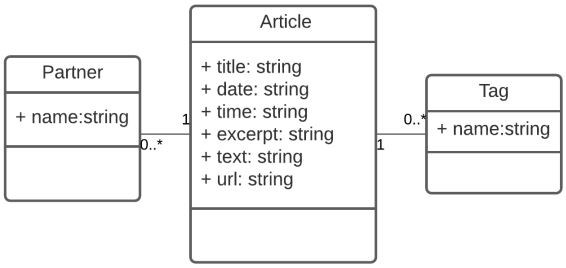

Figure 2: Conceptual Model

The main class is the Article class and the other ones represents additional information. We consider the attributes in the Article Class the ones that are relevant in an article, but some people can ask why this articles doesn't have an author. SAPO24 usually work with news agencies, so the website only has the agency information, that they call partners. The articles can also have some tags that represent usually a subject, a name or an identity. The excerpt is a small portion of the text which can be visualized before opening the article itself.

5 Data Cleaning

Since our data was directly scrapped from sapo website and since the company is responsible for its news quality, the data was pretty decent. Another advantage of scrapping is that we could download the data exactly as we wanted, not being necessary to remove extra columns or to rename them. Despite of this, there were still some improvements that were applied to the data. Sometimes, during the scraping phase, the website was updated in the middle of the process and it parsed the same same news article more than once. There were also some repeated news in the website itself. To handle this problem, we removed duplicate rows from the results. After further analysing the data, we realized there were also some null values, which should be traded by the correct data structures, such as an empty list. Regarding the columns tags and partners, since all the columns were a string it was necessary to correctly parse the data and transform it to a list, so that it would be possible to analyse them individually. Finally it was necessary to parse the data to a more appropriate data structure, such as the dates.

6 Data Characterization

We scrapped 12245 news articles from SAPO24. They were all published in the period between April 7th and November 16th of 2021. After the data refinement, there were 11885 unique news, 6 partners, 7043 tags and 195 dates. It was also possible to differ there were

as much as 5 repeated titles. Considering the word count, there is an average of 429 words per new, being the max of 18103, and the 25, 50 and 75 quartiles being 244, 366 and 523 respectively. Considering the date statistics, the mean of the overall dates was in 2021-08-19 at 14:24:34, and the median in 2021-08-24 at 11:46:00. In order to better analyse the collected data, some graphs of the most relevant information were created.

6.1 News per Month

The chart in figure 3 shows the of number of news per week within this time frame. Notice that the number of news in April is smaller in part due to the fact that we did not perform scrapping for every day of these months.

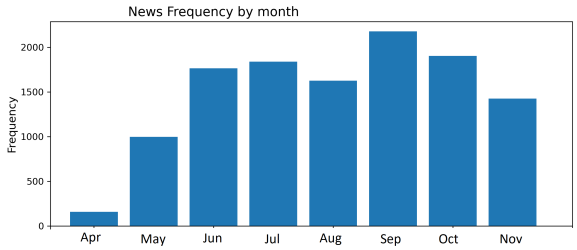

Figure 3: News per Month

6.2 Text Length Characterization

We analysed the length of the text in number of characters. We found that the average text size is 2671 characters and the median is 2309. The longest text in the dataset has 54071 characters and the shortest one has 23. The graph in figure 4 shows the distribution of text length in the dataset. This is a usefull analysys to show to the user longer or smaller news.

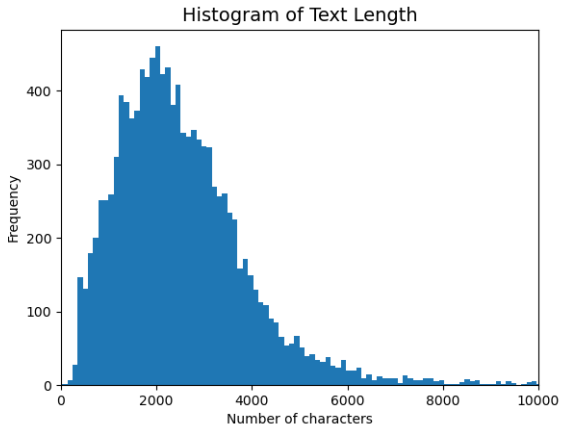

Figure 4: Text size

6.3 Partner Characterization

Considering the partners of the website, MadreMedia and Lusa are by far the most frequent ones, having more than 10000 of the news. Although there were as much as 6 partners, "7Margens" only contributed in 3 news and "The Next Big Idea" only has 4, which can be seen as an exceptional partnership. We decided to also create a Heatmap, to see how the partners were correlated between them, getting the chart in figure 6, which shows clearly that MadreMedia and Lusa contribute together in most of the news. AFP also has some contributions with MadreMedia, although it is hard to see due to the huge scale difference.

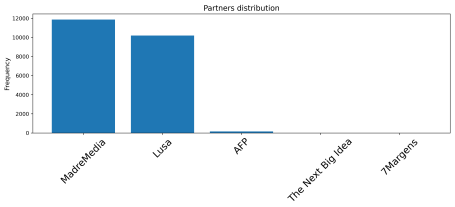

Figure 5: Partner Distribution

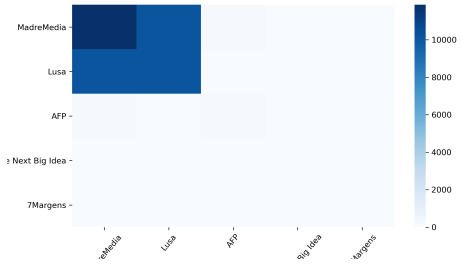

Figure 6: Heatmap partners

6.4 Tags Characterization

In the Tags distribuition graphic (figure 7)., it is easy to see that in the last few months the COVID-19 was the most talked topic by far, having more than 2500 news about it.

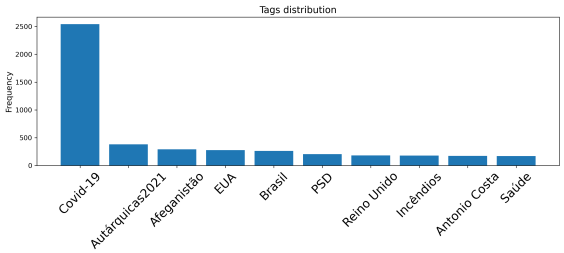

Figure 7: Tags Distribution

To facilitate the visualization of the proportion between the other tags, we also created a graph excluding the Covid-19 tag, giving relevance to the

Autárquicas2021, which happened in 12 December of this year. Another very relevant topic was news related to Afghanistan, which can easly be explanied by the retreat of the US Troops from the country and the subsequent takeover of the Taliban, a government which removed the rights of millions of people, being a perfect topic for a news website. This also explains the USA being in the third place.
(figure 8).

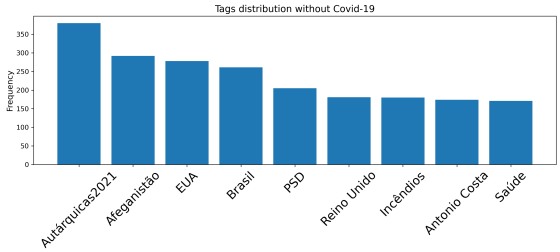

Figure 8: Tags Distribution without Covid-19

6.5 Natural Language Processing

In order to better analyse the content of the news, some Natural Language processing was used. Firstly, it was necessary to clean the text for this objective, removing the punctuation, extra spaces and even the most common not relevant portuguese words. One of the most graphic pleasing representations are of the most common words, being perfectly represented by the next word cloud. The predominance of Covid-19, pandemics, and related topics is also evident in the word cloud, accompanied by politics.

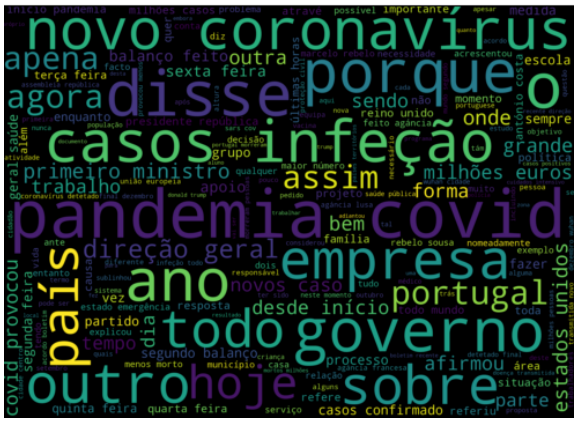

Figure 9: Wordcloud

7 Search Tasks

After processing and refining all the data, it was possible to run and think of some queries which the information system will be able to process. Here are some examples of possible queries to be made to the database:

- Which news are related to technology in the last month?
- What are the most important themes in the last week?
- Which news are related to Donald Trump and Barack Obama?
- Which news have as a partner MadreMedia?
- At what time do most publications get published?
- What small news can I read related to COVID?
- What were the most used words in the last year?
- When was António Costa trendy in the last year?

8 Conclusion

This paper describes the first milestone of Information Processing and Retrieval course. After analysing all the proposed objectives of this Milestone, we think we were able to accomplish all the different topics successfully, although there are numerous of things that were though about in order to improve the project, but were not created due to the lack of time. We also believe we created a strong foundation for the next milestones, but that even this one will be improved during this course.

References

[1] "Atualidade - SAPO24" 2021. SAPO24. https://24.sapo.pt/atualidade.

[2] "Beautiful Soup Documentation — Beautiful Soup 4.4.0 Documentation". 2021. Beautiful-Soup-4.Readthedocs.Io. https://beautiful-soup-4.readthedocs.io/en/latest/.

[3] "User Guide — Pandas 1.3.4 Documentation". 2021. Pandas.Pydata.Org. https://pandas.pydata.org/docs/user$_guide/index.html user-guide$.

[4] "NLTK :: Natural Language Toolkit". 2021. Nltk.Org. https://www.nltk.org/.

[5] "Matplotlib — Visualization With Python". 2021. Matplotlib.Org. https://matplotlib.org/.

[6] "Wordcloud For Python Documentation — Wordcloud 1.8.1 Documentation". 2021. Amueller.Github.Io. https://amueller.github.io/word$_cloud/$.