Portuguese Political Parties in ARQUIVO.PT

João Afonso M. D. Andrade
up201905589@edu.fe.up.pt
M.EIC
Faculty of Engineering of the University of Porto
Porto, Portugal

Maria José V. S. Carneiro
up201907726@edu.fe.up.pt
M.EIC
Faculty of Engineering of the University of Porto
Porto, Portugal

Miguel Azevedo Lopes
up201704590@edu.fe.up.pt
M.EIC
Faculty of Engineering of the University of Porto
Porto, Portugal

ABSTRACT

ARQUIVO.PT is a research infrastructure that allows searching and accessing web pages archived since 1996, with the goal of preserving information published on the Web for research purposes.
The aim of this project is to collect, prepare and process data retrieved from ARQUIVO.PT about all of the portuguese political parties currently holding parliamentary seats, throughout history, in order to develop an information search system of all of their websites.
Multiple tasks were performed as to achieve the final search system such as data collection and preparation, information querying, retrieval and it's evaluation.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Structured text search**.

KEYWORDS

datasets, information processing, information retrieval, full-text search

1 INTRODUCTION

ARQUIVO.PT provides an extensive amount of data collected from the web throughout the years, about multiple different topics and hosts a contest each year that rewards the most innovative projects that analyses and explores their preserved data. Inspired by their mission of maintaining and archiving the information from the web that would inevitably be lost, we decided to explore the content published by all of the current political parties with parliamentary seats on their websites and create a search system that would allow to find each party's statements and evolution throughout history. Considering the misinformation regarding politics and the rise of fake news, having a search system that allows to clarify each political groups opinion and stance about specific issues will allow people to make more conscious decisions about who they elect and choose as their representative. That is our biggest motivation and what makes us passionate about this project.

2 DATA COLLECTION

In order to retrieve every item of content published on each political party's website, we took advantage of the API provided by ARQUIVO.PT

2.1 ARQUIVO.PT's API

Our initial expectation was to collect all the content that concerned the party's website via the API, but after thorough analysis of the API's specification we concluded there was no direct way to do this and so we started developing other strategies to achieve this goal. ARQUIVO.PT's API is a search API that presents a maximum of 2000 results per request, as such, the first approach we considered, consisted of doing an empty query search restricted to the party's web domain and iterating through each page of results. However, after running our first test-run we concluded that the API had some functional issues (confirmed by issues raised by other people in the issues page at Github) that wouldn't allow us to go beyond the first page of results. As such, we moved on to another approach. Since we could only retrieve 2000 results at a time, we decided to do the same search, but instead of searching the whole archive we limited the results to 2000 entries per month. The final API call we used is presented bellow:

Endpoint: "https://arquivo.pt/textsearch"
Parameters: (Parameter Name | Parameter | Value Used)

- Query | q | empty string
- Site search | siteSearch | Party's website (eg: www.ps.pt)
- Start date | from | Start month being queried
- End date | to | End month being queried
- Maximum items per response | maxItems | 2000

The response provided by the API is a list of links to the archived webpages, as well as some information regarding the date of collection and format of the page (whether it's an HTML page, a link to a PDF file, an image, etc). Also included in each of the results was a link to a page that provided the extracted text from that page. This proved very useful as it saved us the trouble of having to extract the text from every page.

2.2 Dataset Size and Associated Constraints

We anticipated that we would have big datasets for each of the political parties, namely the ones who have existed for over 20 years. Partido Social Democrata, for example, has had a webpage for 26 years, which multiplied by 12 months at 2000 results per month could translate to a staggering 624 thousand webpages. This large amount of data raised a couple concerns: how we would store

Conference'17, July 2017, Washington, DC, USA

the data and how long it would take to retrieve it.

To address the storage concern, we decided to store each political party's data in a separate JSON file. As for the time it would take to retrieve all of the data, we initially considered a multi-threaded approach, but soon realized the bottleneck of the process was not the time it took our computers to process the requests but rather the request limit associated with the ARQUIVO.PT's API (250 requests per minute). The API limits and the associated time constraint ended up being one of our biggest obstacles in the first delivery. Our calculations estimated that at most we could have almost 3 million webpages to collect (2 952 000 pages), at 250 requests per minute this would translate to 197 hours of runtime. For this reason, in this first delivery we decided to reduce the range of data by only collecting data of the 4 biggest parties (by number of representatives in the Assembleia da Republica), from the past 5 years.

In the next delivery we intend to include all the data from all the parties who are currently elected.

2.3 Political Parties Website's Domains - Changes Over Time

Before collecting the data through the process described earlier, we conducted research to determine when each of the political parties first created their websites and checked whether or not that website's domain name had changed over time. Table 1 describes our findings:

Political Party Name	Domain Name	Active Period
Partido Socialista (PS)	www.ps.pt	1999 - 2018
	ps.pt	2001 - current
Partido Social Democrata (PSD)	www.psd.pt	1996 - current
Chega (CH)	partidochega.pt	2019 - current
Iniciativa Liberal (IL)	iniciativaliberal.pt	2017 - current
Partido Comunista Português (PCP)	www.pcp.pt	1996 - current
	pcp.pt	2001 - current
Bloco de Esquerda (BE)	www.bloco-de-esquerda.pt	1999 - 2003
	www.bloco.org	2003 - current
	bloco.org	2005 - current
Livre (L)	livrept.net	2015 - 2018
	www.livrept.net	2015 - 2018
	partidolivre.pt	2018 - current
Pessoas–Animais –Natureza (PAN)	pan.com.pt	2013 - 2015
	www.pan.com.pt	2013 - current

Table 1: Political Parties Website's Domains Over Time

Apart from checking domain name changes, we also noticed some parties use both "www" prefixed domains and non-prefixed domains, and so we checked the date for the first occurrence of both www prefixed domains and non-prefixed domains in the ARQUIVO.PT's archive. The content of all these domains was collected and later merged onto a single file (per party).

3 DATA PREPARATION

In order to obtain a cleaned data set in the next stage of the project, we firstly need to prepare and analyse the original data set. Therefore all the data, especially the text one, should be normalized to fit better search criteria. We firstly began by examining the data collected to better grasp it's extent and content and find the best way to adopt in order to correctly organize and clean the data frame.

3.1 Data Pipeline

As we can see, the pipeline presented in figure 3 (Section 7 - Annex) represents the whole process of data collection as a sequence of events. It starts with the decision we made of selecting political parties to collect data from, using the API available from ARQUIVO.PT. After we collected the data we proceeded with the cleaning and refinement of the data, obtaining refined and more precise data frames in the end for each political party. In order to further improve the quality of the data, we also did some data exploration in the end.

3.2 Data Cleaning

After we finished analysing and preparing the data frame we decided to improve the quality of our data in order to obtain the best results possible. In order to accomplish this goal we noticed that most of our text fields had multiple unrecognized characters and escape sequences that were unuseful for the better comprehending of the text provided, so we removed them.

3.3 Data Refinement

After we proceeded with the data cleaning we then decided to more thoroughly refine and treat some more specific aspects of our data frame. For this we removed all the data which had no text field (web pages that provided no information) and we also proceeded to removing any data with corresponding text fields (removing all the duplicates) since there would be no more extra information from those. In the end we ended up with a much more refined and cleaned data frame being much easier to comprehend it.

3.4 Conceptual Data Model

As we can see in the picture below, our conceptual model consists of a single main class:

- party_page: each political party page has a date that corresponds to the date when it was published, a link to the corresponding web page, the contentLength, which indicates the length of the text obtained, the type of the file and finally the obtained text

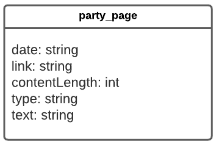

Figure 1: Domain Conceptual Diagram

4 DATA CHARACTERIZATION

Data characterization was made to analyze information about the data collected and to provide a better understanding of each political parties web content. The exploration was initially performed using a Jupyter Notebook, but to ensure larger granularity in the makefile and an organized storage of the outputs we also performed it using python scripts.

All the characterizations done were applied to each party's dataset and can be obtained by running the correspondent scripts, but for the scope of this report we chose to only show, in the case of single party analysis, the results for IL, since we had all of the party data throughout it's genesis and it has a more representative size that CHEGA, for example.

4.1 Null Analysis

Globally, the data retrieved had no null values since it came directly from the API's responses and we only stored the successful ones. Considering that the goal of our information system is to search the webpage content of each political party, it was important to explore the empty text content present in each dataset. As we can see in table 2, CHEGA was the party with the highest percentage of pages with empty text at 68%, contrasting with PSD, which had the lowest, at 23%.

	Missing Data	Percentage
PS	22922.0	51.590106
PSD	6994.0	22.885377
CH	2380.0	68.058336
IL	6378.0	47.650355

Table 2: Missing Data per Party Webpage

We also wanted to explore the correlation between the webpage type and the empty text values, as that could explain why a lot of the information retrieved had no text. As we can observe in figure 2, almost all of the application/json webpages have empty text, and only a small amount of text/html webpages are empty, so only the ones that are of a textual type are worthy of performing text and word analysis.

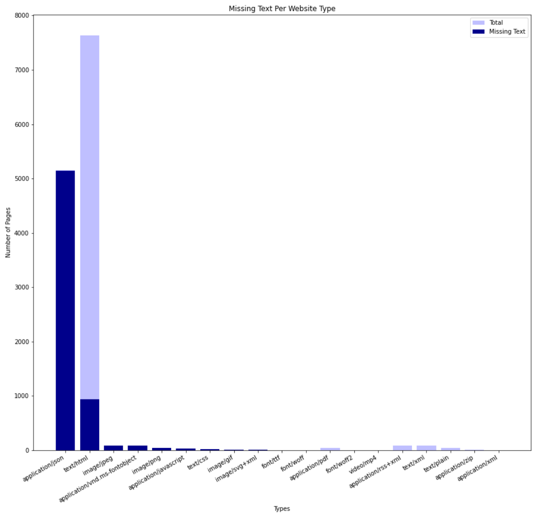

Figure 2: Missing Text In IL

4.2 Text Analysis

In order to grasp the amount of data collected for each party and compare the text between political groups, we analyzed the amount of pages and the text length for each one. This improved our understanding of the data and it's quality.

In table 3 we can see the average number of webpages retrieved for each party per year, and their minimum and maximum values. CHEGA is a good example of the ARQUIVO.PT limitations because it is a relatively new party, with genesis in 2019, and since ARQUIVO.PT lacks in updated information in recent years, searching and obtaining data about new parties is rather difficult. In contrast, older parties like PSD, have larger amounts of data.

	Avg Number of Pages	Min	Max
PS	2637.8	6.0	4967.0
PSD	3481.8	1.0	7203.0
CH	117.4	0.0	457.0
IL	1095.6	0.0	2628.0

Table 3: Pages per Party per Year

In Figure 3, we can see the total amount of pages per year by each party and the limitation highlighted before is clearly present in all of them, as there is barely any data present in 2021. PS and PSD are the most active in each of their webpages, and CHEGA is the least active.

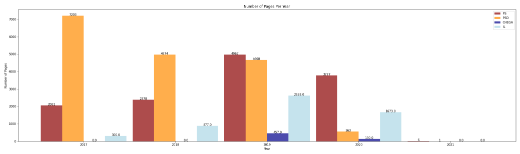

Figure 3: Pages per Year

The average webpage length allows us to understand some of the text quality in each party. For example, CHEGA is the party with the smallest amount of pages, but the second highest in text length , which shows a tendency in writing less but more extensive pages by this party. IL is the party with the smallest text length, contrasting with PSD, which is the highest.

	Avg Length	Min	Max
PS	4679.444214	1	735766
PSD	66950.572294	2	1185007
CH	19425.709830	39	153538
IL	4400.423930	2	418327

Table 4: Text Length per Party per Year

Textual analysis makes us differentiate the style of broadcasting information by each party, as well as recognize gaps in our data and their future implications in the long run.

4.3 Words Analysis

When characterizing a dataset that includes text it can be relevant to see the frequency of the use of each word in the text. In our particular context, we expected the most frequent words to point out a theme in the discourse of each party. To do this analysis we plotted a word cloud, as can be seen in Figure 4.

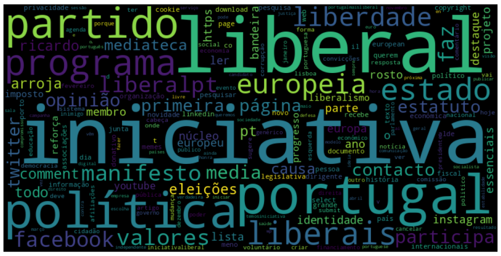

Figure 4: WordCloud IL

Still on the topic of word analysis, since these parties are in constant competition, we considered it would be interesting to analyse the number of times the other parties are mentioned in each party's website. The results can be seen in the example expressed by Figure 5.

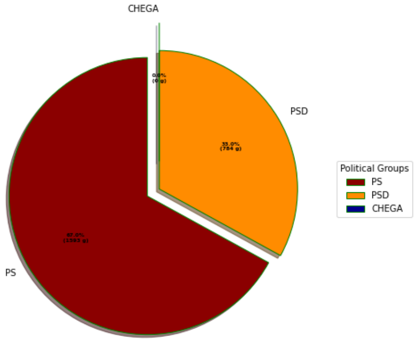

Figure 5: Mentions of Other Parties in IL

5 PROSPECTIVE SEARCH TASKS

We conclude this first milestone with a clean and characterized dataset that includes the past 5 years of webpages of the top 4 political parties by number of seats in the Assembleia da República. In the future we plan to extend this dataset to every webpage of all the parties represented in the Assembleia da República since their websites inception date.

6 ANNEX

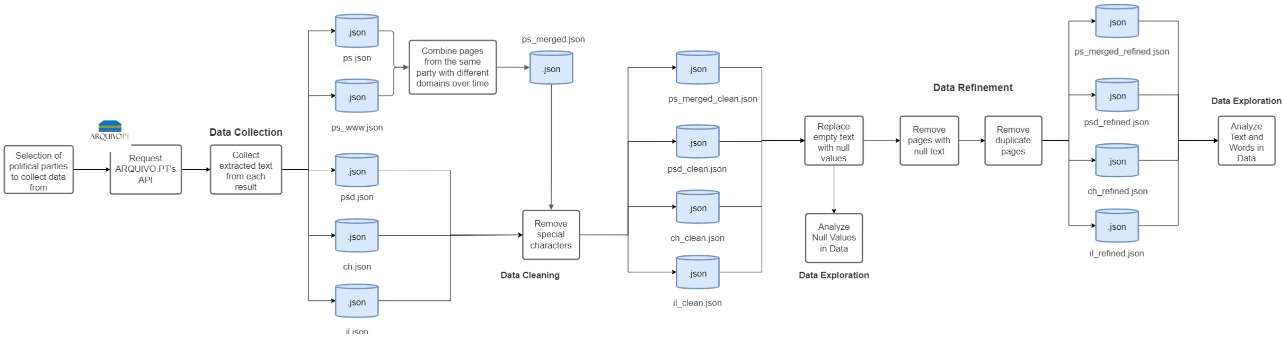

Figure 6: Pipeline Diagram