News search engine

Faculty of Engineering of the University of Porto, Portugal

Vítor Cavaleiro
up202004724@edu.fe.up.pt

Rodrigo Figueiredo
up202005216@edu.fc.up.pt

Diogo Fonte
up202004175@edu.fe.up.pt

Sofia Rodrigo
up202301429@edu.fe.up.pt

ABSTRACT

This report intends to document the process of collection, preparation and processing of a specific unstructured news dataset, for a later development of a search system. The data was collected from trustworthy sources such as Components and ML Resources.

The data preparation phase was conducted throughout a pipeline and subsequently analyzed for a graphic understanding. The next step was to index the information to enable its retrieval using queries. These results have been thoroughly evaluated and they are presented later in this report.

KEYWORDS

Dataset, data, news, pipeline, indexing, query

1 INTRODUCTION

In today's rapidly evolving information landscape, the ability to access and extract relevant information from the vast ocean of digital data has become paramount. The digital age has ushered in an unprecedented era of information abundance, where a continuous stream of news and updates inundates our screens, complicating the task of discerning what truly matters. The constant flow of information, generated by a multitude of sources, from social media platforms and news websites to academic journals and research repositories, underscores the pressing need for robust information retrieval tools.

This article introduces a cutting-edge news search engine, specifically designed to meet the challenges of information retrieval. By combining a data preparation and cleaning methodology, a further understanding of it through graphic analyzers and an indexing and retrieval tool such as Solr, this search engine promises to deliver a tailored and insightful user experience, making it an indispensable tool for individuals and researchers.

The report is divided in two principal sections.

The first one gathers the data preparation process, which includes how the datasets were selected, what they contain and the explanation of the pipeline carried out to obtain data prepared for its analyzing. This first milestone lays the groundwork for a simpler information needs retrieval.

The second section involves information indexing using Json schemas and algorithms, as well as a profound evaluation of the different proposed queries and their results.

2 MILESTONE 1: DATA PREPARATION

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for exploring and visualizing the data. Using specialized data preparation tools is important to optimize this process [4].

2.1 Data collection

Data collection is the first important step to produce a reliable search system. It involves choosing and exploring the content of the dataset, as well as exploring its quality.

2.1.1 Dataset choice. Before selecting the dataset, it was essential to reach a consensus on the theme. After deliberating the pros and cons of various topics, we concluded that news would be a highly suitable subject for an information search system, given the possibility of finding multiple datasets meeting our desired quality criteria.

Once the theme was decided, we initiated the search for datasets that met the necessary requirements, particularly concerning data size and quality. The chosen dataset encompassed 204,000 articles from 18 American publications, collected from Components [3]. However, in order to incorporate data from another source and introduce some complexity, we opted for an additional dataset, this time consisting of 2,555 documents sourced from the BBC website, collected from ML Resources [1].

The first dataset has an MIT license, while the second one is open for academic purposes.

2.1.2 Dataset content. The first dataset contains 204,135 articles from 18 American publications. Includes date, title, publication, article text, publication name, year, month, and URL (for some). Articles mostly unevenly span from 2013 to early 2018, with a smattering pre-2013.

The second one consists of 2225 documents from the BBC news website corresponding to stories in five topical areas from 2004-2005, divided in 5 categories: business, entertainment, politics, sport and tech, and includes raw text data, from where we extracted the needed attributes.

To merge the datasets, we standardized both of them into the same format, ensuring the following attributes:

- `title`: Title of the article
- `author`: Author of the article
- `date`: Date of publication in format YYYY-MM-DD
- `content`: Textual content of the article
- `publisher`: Publisher of the article
- `source`: From where the article was extracted

- `category`: Describes the topic or subject matter of the article.
- `url`: Source webpage of the article

2.1.3 Data quality. The data was sourced from Components and ML Resources, two well-known and reputable platforms. It also has a large volume, is diverse and well documented.

There were some null values and missing data for some fields, but everything was handled the way it should and the data was standardized into a format we specified.

2.2 Data preparation

Data preparation is the principal phase of this first milestone. Data ingestion and data cleaning ease the management of indexing tools that will be concreted afterwards.

2.2.1 Data ingestion. It is important to clean, reduce or filter the data before its analysis and processing.

Firstly, the first dataset that comes in a .db file is exported to a JSON file using SQlite studio. Then, a python script converts through simple functions and dataframes the JSON file to a CSV file.

A similar process takes place with the BBC database. In this case, the data is collected from BBC and BBCSports articles written in various txt files, separated in 5 folders according to the categories. Afterwards, another python script combines them into a CSV file.

Finally, we bind both CSV files into one, which will be the final format of the data for further usage.

2.2.2 Data cleaning. With the help of the pandas library of python, we manipulate the CSV file throughout dataframes.

For the first dataset we start by removing irrelevant parameters such as the first column, the *id*, the *digital* and the *section* in order to transform our data into a more useful information source. Moreover, we eliminate redundant columns such as *year* and *month* since it is already detailed in the *date* parameter. Another important change was to rename some columns to more descriptive words for the actual meaning of the columns. *Publication* parameter was changed to *publisher*, *category* to *source*, and *section* to *category*.

The second dataset was exported as raw text files divided into categories, so we extracted the relevant attributes from the text and merged all the categories into one csv file, leaving it in the same format as the first dataset, so we could merge both datasets.

After merging both datasets, we had one csv file with all the data and there were some things we needed to do in order to clean the data. The first thing was removing all duplicates, as well as replacing empty strings with NaN. We also decided to remove every row with missing title or content as that is valuable information that needs to be present. There was also a problem with the author names, as a lot of them were coming with "\n" before and after their names, so we removed that as well.

2.3 Data analysis

At the start of this stage we decided to create a new column refering to the keyphrases of each article. We used the rakt-nltk library for that.

In order to extract more information about the chosen dataset and with the help of python libraries such as matplotlib.pyplot or seaborn, we have developed different data analysis plots to represent some parameters. Each of the following tries to obtain some characterization of the data to help us understand the information we are handling:

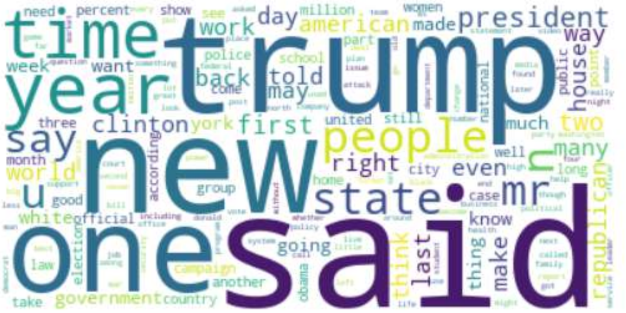

Figure 1: Articles wordcloud

Figure 1 represents a wordcloud, which is generated from article content using keyword extraction to identify the most common words or phrases. We could conclude that our data is mostly dedicated to political issues.

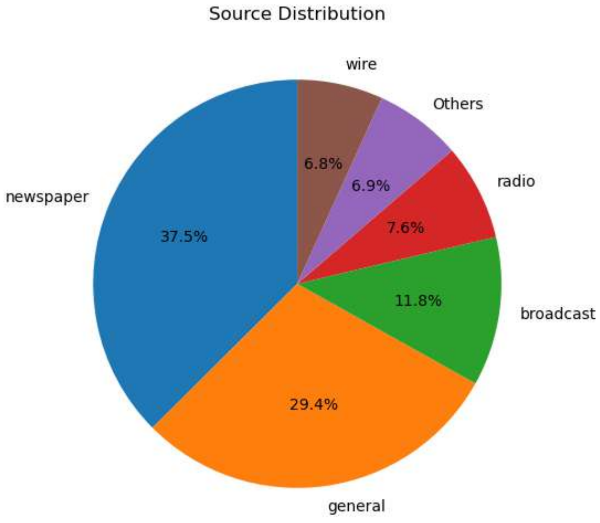

Figure 2: Sources pie chart

Figure 2 represents a pie chart, which visualizes the proportion of articles from each publication source within the dataset. The plot clearly emphasizes the newspaper as the most used source of news release. The general category includes articles that were published in several platforms (newspaper, website, etc).

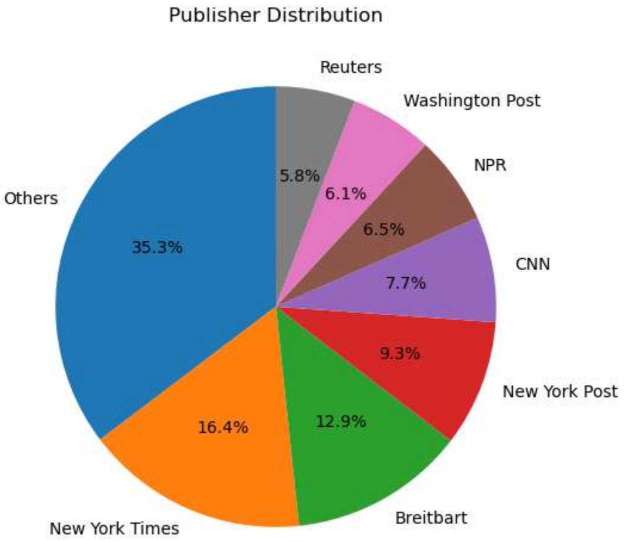

Figure 3: Publishers pie chart

Figure 3 represents another pie chart, which shows the percentage of articles from each publisher within the dataset. The most common ones are The New York Times, Breitbart and The New York Post.

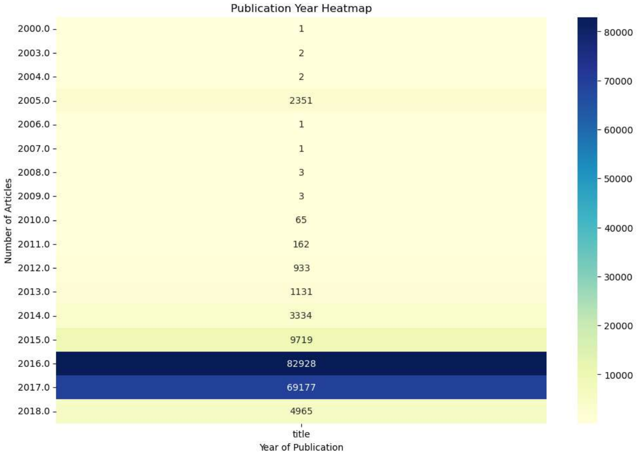

Figure 4: Heatmap of articles per year

Figure 4 represents a heatmap, a multivariable plot generated to visualize the number of articles published per year. 2016 and 2017 are clearly the years that cover the most amount of written articles.

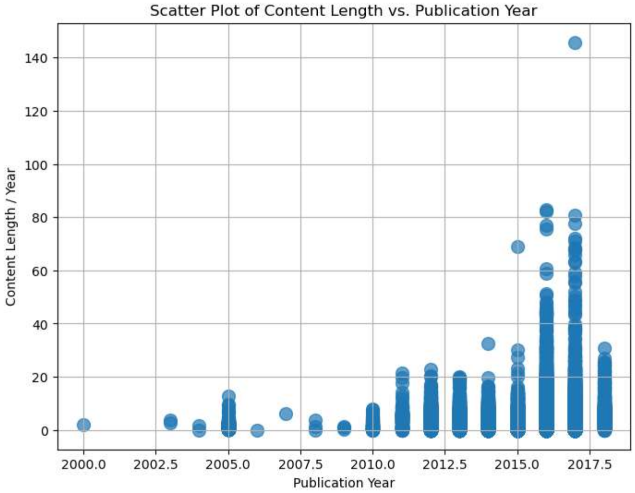

Figure 5: Scatter plot of content length per year

Figure 5 describes a scatter plot, which explores the correlation between the length of the articles and the publication date. It exposes 2015, 2016 and 2017 as the years with longer articles and also with the most amount, just as shown in the previous plot.

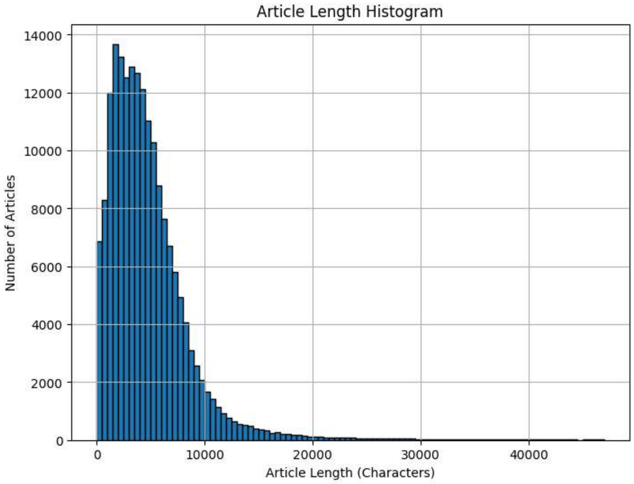

Figure 6: Article Length Histogram

Figure 6 represents a histogram of the articles length, and it can be observed that it has a right-skewed distribution with the majority of articles having a length of less than 10000 characters.

2.4 Domain data model

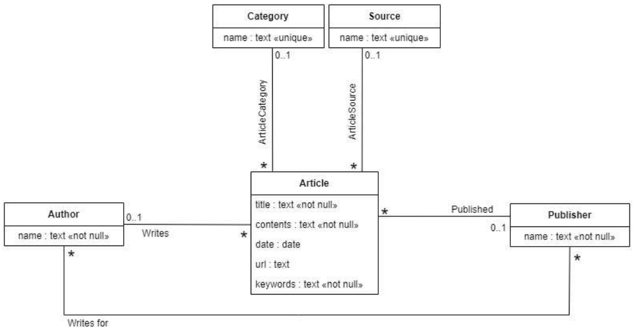

Figure 7: Domain data model [2]

The data model comprises four primary classes that represent key entities within the domain: Article, Author, Publisher, and Category. These classes encapsulate essential aspects of the domain and are interrelated to facilitate effective data organization.

The Article class serves as the central entity and embodies written content within the domain. It possesses several attributes that define articles, including a title, content, publication date, URL, and keywords. Articles have associations with the Author, Publisher, and Category classes, enabling the linkage of articles to authors, publishers, and categories, as appropriate.

Authors represent individuals responsible for creating articles and the sole attribute for authors is their name. An Author class is associated with the Publisher class, illustrating the collaborative relationship between authors and publishing entities.

Publishers signify the organizations or entities responsible for publishing articles, and the only attribute is the name of the publisher. This class maintains a connection with the Author class, demonstrating the affiliation between authors and their respective publishers.

Categories are used for classifying articles based on various themes or topics, and their only attribute is also the category name. Categories are directly connected to the Article class, enabling the categorization of articles into relevant topics.

Sources are connected to the articles as they specify where a given article comes from. Their only attribute is the source name.

2.5 Data processing pipeline

The whole process of this first milestone is developed in Jupyter Notebook.

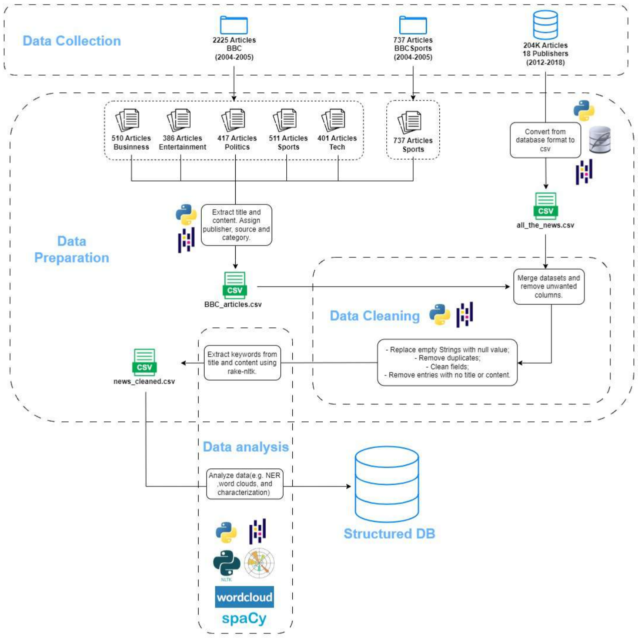

Figure 8: Data flow diagram of the pipeline [2]

2.6 Prospective search tasks

The search system that is going to be developed in the next milestones requires a previous task of information needs search. These are some examples:

- Find news articles where Trump spoke on the immigration crisis
- Find news about LeBron's good performances in lost games
- Find articles related to homicides investigated by the FBI in 2017
- Find news articles regarding the conflicts between republicans and democrats about gun ownership

REFERENCES

[1] [n. d.]. *BBC dataset from ML resources.* http://mlg.ucd.ie/datasets/bbc.html
[2] [n. d.]. *Diagrams.net.* https://app.diagrams.net/
[3] [n. d.]. *News dataset from Components.* https://components.one/datasets/all-the-news-articles-dataset
[4] [n. d.]. *What is data preparation?* https://aws.amazon.com/what-is/data-preparation/?nc1=h_ls