Scientific Articles Search System

BEATRIZ SANTOS, SÉRGIO ESTÊVÃO, and SÉRGIO DA GAMA

As part of the curricular unit of Information Processing and Retrieval, this report gathers the development of the first milestone or our project. In this project, our goal is to refine a dataset chose by us and create an information processing and retrieval tool. However, for this milestone, we only had to analyse the data and create a successful pipeline to process and transform it. With this in mind, we start by searching a dataset and we end up choosing one of scientific articles. This choice was mainly based on the attribute summary that had a good amount of text to be indexed in next milestones and on the attribute link, which allow us to connect each entry to the actual entire article. So, after having the dataset we start processing it, removing and creating some attributes or just transforming some of them. The combination of this operations resulted in the overall pipeline. This last one is accomplished and described in the makefile that we created to process the data all at once. Finally, we also developed a classes diagram that visually represents our dataset structure.

Additional Key Words and Phrases: Information Processing and Retrieval, M.EIC, dataset, arXiv, papers, statistics

1 INTRODUCTION

Within the scope of the curricular unit of Information Processing and Retrieval, a database search system was requested. This project required the group to perform data collection and preparations, querying and retrieval, and retrieval evaluation.

At beginning of the project, we started by choosing a theme for the dataset, common themes were politics, news, movies, and music. We decided to elaborate the project based on scientific papers, because there would be vast amounts of textual information associated with them and they are easily categorized, since we could use authors and fields of investigation to batch and select specific articles. Both of these properties are essential in creating an ideal information retrieval system.

2 DATA PREPARATION

The project's first milestone is preparing and characterizing the dataset. The dataset, *ArxivData.json*, was obtained through Kaggle. This dataset is a collection of scientific papers and their corresponding information from the website *arXiv* [1], maintained by Cornell Tech.

In this milestone we modify the dataset to satisfy our needs, analyze it and develop its data domain conceptual model.

2.1 Data Pipeline

In order to get the most relevant information from the chosen dataset, our pipeline suffered some alterations.

The first step was to aggregate all of the **authors** from an article into a list. After that, the normalization of the publication **date** was created. This is useful to search for a paper from a specific month or date. In addition, some unnecessary links, that the original data set kept, were removed and only the **link** that has all of the information about the article, such as the PDF file, was kept.

Lastly, each element of the dataset had a sub-element containing **tags** that characterize the area of the article. After some observation, the group noticed that not all of the information from that sub-element mattered to the project, dropping the unnecessary information. Besides that, in order to get the full name and normalize the tags, some **web-scrapping** had to be done.

Starting the scrapping phase, we confirmed that the dataset had outdated tags from the websites. Fortunately, concatenating these outdated tags with the *arXiv* [1], it redirected to the web pages of its corresponding updated tag. To confirm we had the updated one, we scrapped the web pages of each tag to confirm we had the most recent ones.

After confirming we had all the tags, we could begin the process of extracting the additional information from the tags. By scrapping the page of each tag, we got its respective **field** and **subject**, and by scrapping the main website page we got the **area of study** of each tag.

This ends the process of organizing the original dataset to retain only what we need for the proposed work.

After this process, the dataset was ready to be converted into a Pandas dataframe and the existence of null values was searched for, such as the removal of duplicates.

Figure 1, shows this process of pipeline formation.

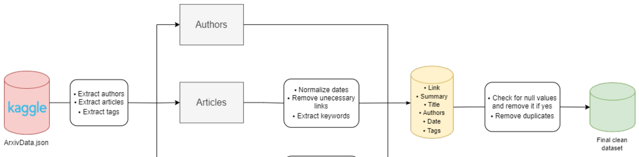

Fig. 1. Pipeline formation

Note: we used a cache system to save the corresponding information of each tag, reducing the number of requests to the website and consequently speeding up the refining process.

2.2 Data Characterization

Our dataset focuses on the scientific **articles**, which are then accompanied by their respective **authors**, as well as other useful information.

The dataset itself has **41 000 entries**. However, there are more authors than entries (articles), because an article can have more than one author, making up to **59 888 authors**, of which 10 are represented in Figure 2. The authors in the said figure are the ones with more articles published, sorted in ascending order.

Authors' address: Beatriz Santos, up201905680@up.pt; Sérgio Estêvão, up201905680@up.pt; Sérgio da Gama, up201905680@up.pt.

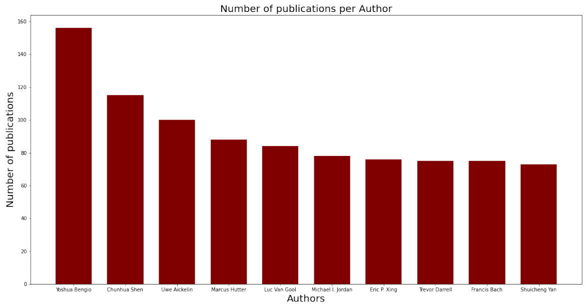

Fig. 2. The 10 authors with more articles

Yoshua Bengio appears to be the author with more articles published, having almost 160 publications. Therefore, a big gap exists between him and the second author, Chunchua Shen, which has about 120 articles, 40 less than Yoshua. After the second author, the number of articles from the rest of the authors start stabilizing with about 80 articles published.

Although, Figure 2 only gave us a good understanding of each author. So, to get a better grasp on how authors appeared in our dataset, we created the histogram in Figure 3, which uncovered that most of the dataset entries had from 1 to 15 authors. Besides that, the most frequent amount is 2 to 5 authors per Article. By analyzing the said graph, as the zero column is empty, we can also conclude that all the articles have at least one author.

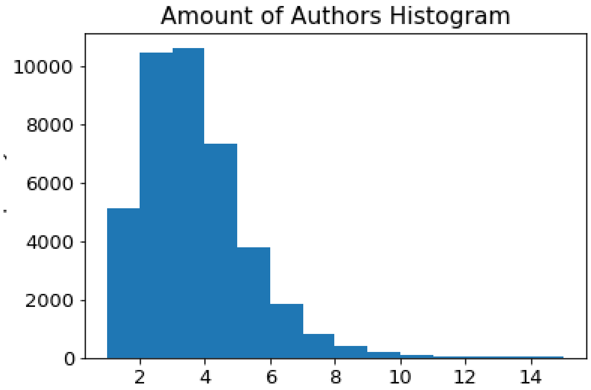

Fig. 3. Frequency of Authors per Article

Additionally, we also analyzed the frequency of subjects per article, represented in Figure 4. This allowed us to see that each Article is related to a relatively small amount of subjects, ranging in general, from 1 to 4 subjects per Article. It is also visible that all Articles have at least related to one subject.

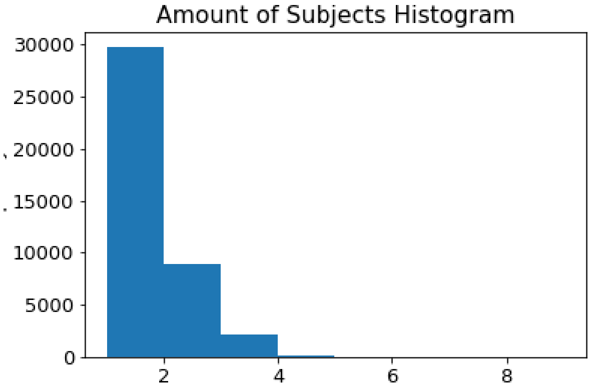

Fig. 4. Frequency of Subjects per Article

In order to place the data in time, we created Figure 5 which revealed that the oldest article in our dataset is from 1993 and the most recent from 2018. Nevertheless, the majority of the reports were published between 2018, almost having normally distributed data between those dates. The said graph also reveals that the year with more articles published is by far 2017, with over 12 thousand publications made.

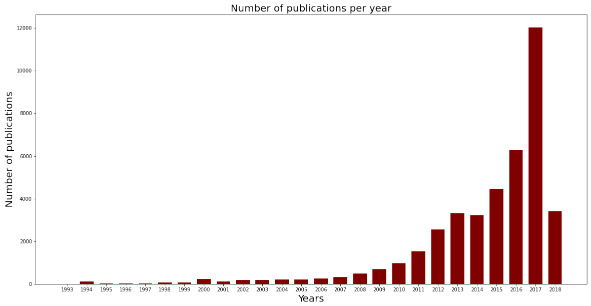

Fig. 5. Articles published by year

We decided it would be interesting to check the distribution of subjects in the papers and verify which ones were more widely common. Since the dataset has a total of **161 unique subjects**, 156 of them appear in less than 4000 articles (>0,1 % representativity), so we group them in the "Others" category. In Figure 6 we can verify that Machine Learning, Computer Vision and Pattern Recognition, and Artificial Intelligence are the most common subjects, all subjects of Computer Science.

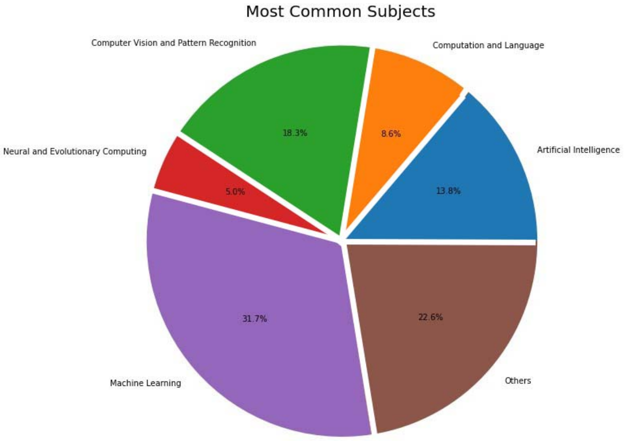

Fig. 6. Subjects of the papers

Furthermore, each article also as areas associated. In total, our dataset has **8 areas**, although, the most common areas are Computer Science, Statistics, Mathematics, etc.. The widely majority of the articles are about Computer Science, and here we can confirm that the dataset isn't clearly balanced when it comes to areas, being more focused in Computer Science. This can be observed in Figure 7.

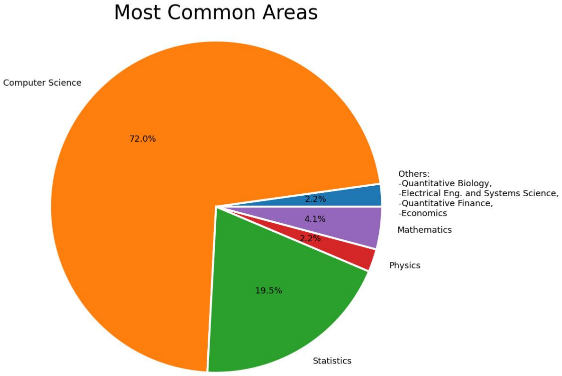

Fig. 7. Areas of the papers

Apart from areas and subjects, in our dataset each area can have multiple fields. In this case, Physics is the only area that as multiple ones, in Figure 8 we can observe the distribution of Physics fields in the articles.

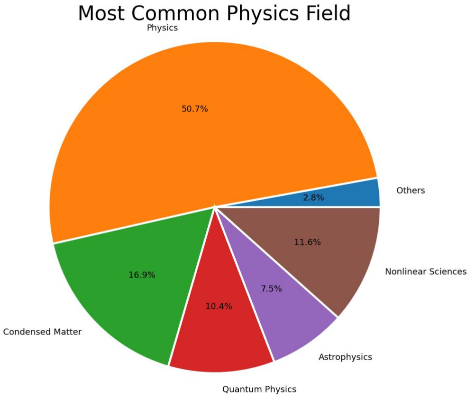

Fig. 8. Fields of physics

Analysing now the 'summary' attribute, to find the most common vocabulary used in the papers we created a graph showing the top words in the dataset. This data shows that words like "model", "method" and "network" are widely used. This statistic confirms the dataset context, since the contents of these papers are scientific, academic, and mostly about Computer Science and these top words are commonly used in those fields. I the Figure 9, we can observe this facts.

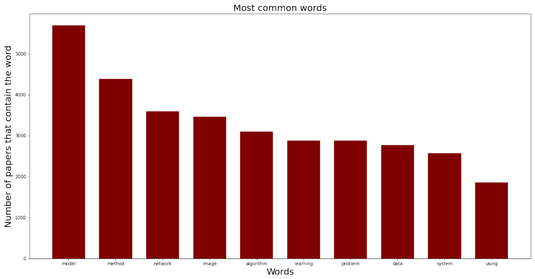

Fig. 9. Most common Words

Additionally, we also analyzed the frequency of words in the summary field per article, represented in Figure 10. This allowed us to see that most of the articles have between 100 and 200 words.

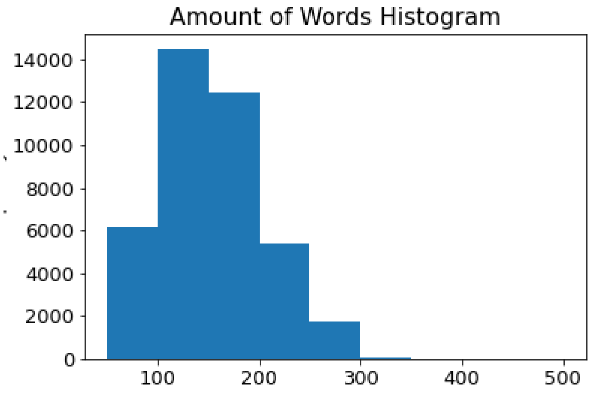

Fig. 10. Areas of the papers

2.3 Data Domain Conceptual Modal

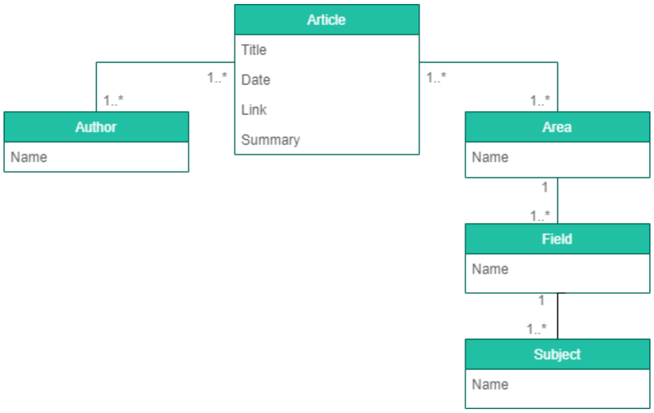

Fig. 11. Data Conceptual Model

Our data conceptual model, represented in the Figure 11, has one main class, **Article**, that has the attributes below:

- **Title**: Article's title
- **Date**: The publishing date
- **Link**: The link where the Article is available
- **Summary**: Summary of the Article's content

Each Article has one or more Authors, which are instances of the **Author** class that only has one attribute, the Author **Name**.

In addition to that, each Article also has one or more Areas, instances of the class **Area**. In turn, these last ones have one or more Fields, which are instances of the **Field** class. Finally, the fields have subjects related to them, which are instances of the **Subject** class. However, a subject can only belong to a field, and a field can only belong to an area.

3 CONCLUSION

In this milestone, we acquired, prepared, and characterized the dataset of our project. All of the goals were accomplished since the end product of the work done was a coherent and complete dataset with relevant information to start the information retrieval system. Furthermore, we developed a better understanding of the project's theme that will indeed prove helpful in the upcoming milestones.

REFERENCES

[1] Ginsparg, P. (1991). Research-sharing platform.