Information Processing and Retrieval

T03G02

Nuno França
up201807530@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

João Tomás Teixeira
up202108738@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

Isabel Silva
up201904925@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

Rodrigo Esteves
up202403070@up.pt
Faculty of Engineering, University of Porto
Porto, Portugal

Figure 1: Different species [2].

Abstract

This project focuses on biological species, where accurate and accessible information is essential for researchers and educators. The article explains the steps needed to develop a specialized search engine for species data, which is available on Wikipedia. The process involves collecting species-related data from Wikipedia, cleaning and preparing it for analysis, and conducting an in-depth review of the information. The ultimate goal is to create a robust system that enhances search capabilities, improves information retrieval, and ensures that users can easily access accurate data on a wide range of species.

Keywords

Species, Information, Datasets, Scraping, Data Retrieval, Data Preparation, Data Analysis, Data Processing, Pipeline, Data Refinement, Data Cleaning

1 Introduction

This article is part of the "Information Processing and Retrieval" (PRI) course, undertaken during the first semester of the first year of the Master's in Informatics and Computing Engineering (M.EIC) at the Faculty of Engineering, from the University of Porto (FEUP).

The focus on species as a research topic stems from its relevance in biological and environmental sciences and the wide range of characteristics it presents. The study of species and their attributes is particularly significant in today's data-driven world, offering insights into biodiversity and ecosystems while also showcasing the complexity of biological data. The topic covers varied types of information, such as taxonomy, distribution, and behavior, making structuring data more challenging and thus a suitable candidate for investigating advanced search and retrieval techniques. This theme directly aligns with the course's goal of exploring practical, real-world information retrieval systems.

In this project phase, we begin with **Dataset**, which introduces the sources of species data and assesses the data quality. Here we explain the selection, processing, and storage methodologies, ensuring a clear workflow and we evaluate and visualize the processed data, using different criteria to analyze relationships. Finally, **Conclusions and Future Work** highlights the outcomes and sets the direction for the upcoming phases of the project.

2 Dataset

The dataset produced by the data extraction and enrichment method includes important contextual and visual information in addition to species names, providing a solid basis for further research and information retrieval tasks.

The subsequent subsections will provide further details on the procedures applied during the data collection and filtering processes.

2.1 Description

The Wikipedia Species Directory, an extensive and community-maintained database of biological and botanical species, provided the species information that assembled the dataset used for this research. Entries for several species from diverse biological and botanical groups are included, along with relevant data such as scientific names, introductory information, and other characteristics.

2.2 Origin

The dataset was generated by extensively gathering all the items of data that could be found in the Wikipedia Species Directory (Wikispecies).

2.3 Pipeline

Utilizing the LXML [5] Python module, HTML data from Wikispecies was scraped and converted into a text file with 1.500.000 lines, one line for each page name. As a result, the information was filtered using this same library, and only the data relevant to the scope of this project was stored. Subsequently, BeautifulSoup4 [3] was utilized in Wikipedia, to collect complete information on the species and save it across multiple JSON files.

It is possible to view the pipeline's methodology visually in figure 2.

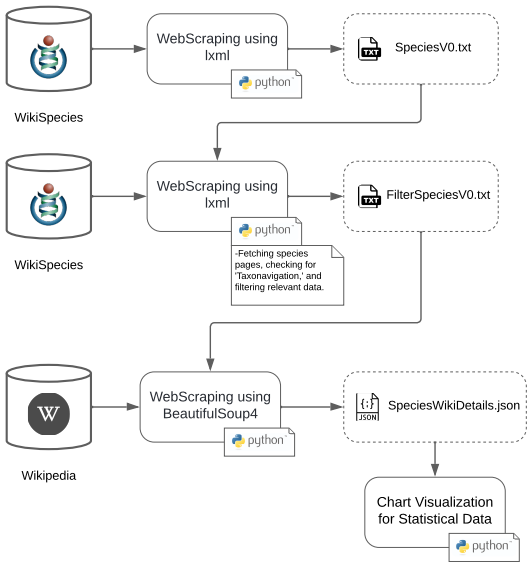

Figure 2: Pipeline.

2.4 Data Collection

The data extraction process began with scraping the Wikipedia species page to compile a comprehensive list of species. Although the page contains a wide variety of information, it also includes entries that are not specifically related to species. To ensure the dataset's relevance, filtering techniques were applied to retain only the species-related entries, resulting in a cleaner dataset.

After this initial filtering, the selected list of species was used to scrape Wikipedia for additional detailed information. For each species, several key pieces of information were gathered. The introductory section of the Wikipedia page, which offers a brief overview of the species, was extracted, along with other relevant sections detailing aspects such as habitat, behavior, and conservation status. Additionally, the primary image associated with each species was collected to provide a visual reference. The scientific classification of each species, covering the taxonomy from domain to species level, was also retrieved. Furthermore, details regarding the discoverer or the team responsible for first describing the species were obtained.

2.5 Data Cleaning

Throughout the data cleaning procedure, all entries were examined to ensure their relevance to species or whether the information gathered related to other data that was irrelevant to the project's scope.

Despite verifying that the species were included in Wikispecies, certain columns required removal due to their non-conformity to actual species.

To prepare the dataset for analysis, several duplicate entries were identified and subsequently removed to ensure data integrity and improve the accuracy of the results.

In addition, the species with a brief introduction (less than 150 words) and no sections were eliminated, as were the wiki sites lacking taxonavigation.

When the data was cleaned, it was discovered that, in certain instances, links about distinct species led to the same Wikipedia article. About 30,000 values in the dataset experienced this and were subsequently eliminated. Nonetheless, these statistics will be analyzed further in the document.

Another column was eliminated because not many species had a conservation status indicating whether or not they were in danger of becoming extinct.

2.6 Final Dataset

Following the conclusion of the data collection, and cleaning processes, the final dataset is organized and includes the relevant species information found in the Wikipedia Species Directory (Wikispecies).

Following the data cleansing procedure, the final dataset has about 120,000 records for distinct species.

There are still some limitations on the existing data in the final dataset. Some species don't have complete information, especially when it comes to the Discoverer and Conservation Status.

Because of the limited information in the source material (Wikipedia Species Directory - Wikispecies), some entries have limited descriptive details.

However, the final dataset, which is based on openly accessible data from the Wikipedia Species Directory (Wikispecies), provides a strong, organized basis for examining species diversity, classification, and associated information. Even though this corrected dataset is much smaller than the original extraction, it is nevertheless of excellent quality and relevance for further study or investigation.

2.7 Tables with the Main Fields

This project contains four distinct tables: Section, Species, Discoverer, and Conservation Status.

Information about the section title and the field containing the matching text can be found in the Section table.

The Conceptual Model discusses the various fields that constitute the species. Still, the primary fields are the image, the name of the species, and the introduction — which provides crucial details about the particular species.

The Discoverer table lists the name of the person who found the species, although this may not always exist.

Lastly, the name field is the only one in the Conservation Status table, and it may not be present for some species.

2.8 Conceptual Model

After the data preparation phase, the project's structure was built around a well-defined conceptual model:

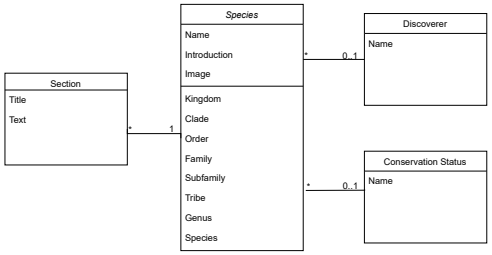

Figure 3: Conceptual Model.

A species consists of the following attributes: a name, an introduction, an image (which may not exist), and a scientific classification. Both the discoverer and the conservation status have corresponding names and can be linked to multiple species. Each Wikipedia section for a species is treated as a unique entity. Sections with the same title but different content are considered distinct and are associated with only one species.

2.9 Data Characterization

2.9.1 Statistics.

With the final dataset ready, the group created some graphs to better visualize and analyze the data that will be further explored in the following parts of the project.

To complete the mentioned task, Pandas [4] was utilized to analyze a CSV file, created with meaningful values from the JSON dataset, while Matplotlib [1] helped with the graphics creation.

Figure 4, 5, 6 and 7, are the most relevant statistics, where it can be seen distributions of size and the most relevant values for the future work of full-text search.

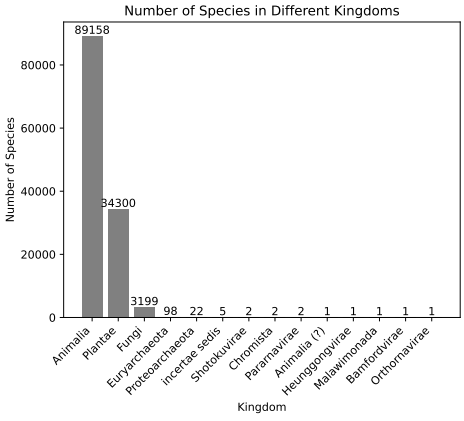

Figure 4: Species divided by kingdoms.

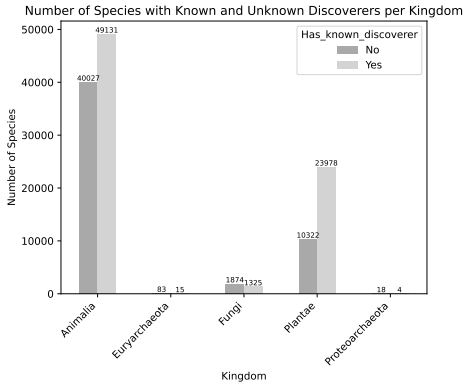

Figure 5: Known discoverer by species kingdom.

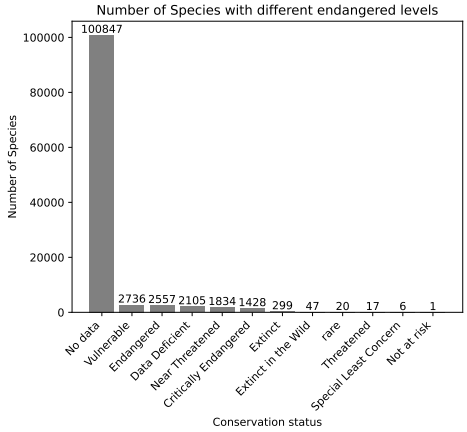

Figure 6: Species conservation status.

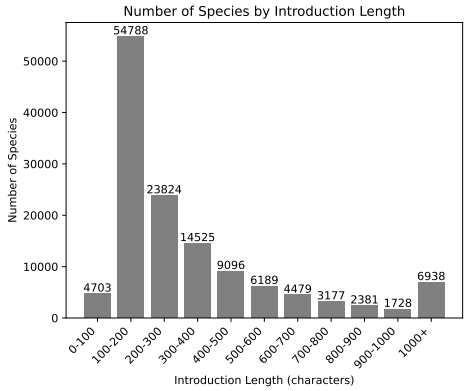

Figure 7: Species introduction length.

2.9.2 Fields/Values Comparison.

Figure 4 illustrates the distribution of species across the different kingdoms. The most common kingdoms are Animalia, Plantae, and Fungi. The discrepancy in size is so pronounced that, for the purpose of this project, the other kingdoms could be ignored.

Figure 5 shows the distribution of the number of species of a given kingdom that have a known discoverer. The number of known discoverers of species of the kingdoms Euryarchaeota and Proteoarchaeota is so low, compared with the number of known discoverers of the kingdoms Animalia, Fungi, and Plantae, that it is only possible to visualize its columns in the graphic because of the values written above said columns. When comparing the distinct columns Animalia, Plantae, and Fungi, the dispersion of the values is high. The Plantae kingdom has approximately twenty times the number of known discoverers of the Fungi kingdom, and therefore the Animalia kingdom has around twice the number of known discoverers of the Plantae. This means that the Animalia kingdom has almost fifty times more known discoverers than the Fungi kingdom, which proves the value difference amongst columns.

Figure 6 shows the big difference between species that have no data regarding the conservation status and the species that have relevant information on this field. It is also important to notice that there is a column labeled Data Deficient, which denotes that there is no known information about the conservation status.

2.9.3 Textual analysis.

The dataset has two main fields with rich textual data, the introduction field and the sections field.

The statistics related to the introduction field can be seen in Figure 7. The figure shows the distribution of species introductions relative to the number of characters in their description. As seen in the graphic, the most common length for the introduction is between 100 and 200 characters. Beyond this range, there is an exponential decrease in the number of species as the length of introductions increases.

The statistics related to the size of the section field are presented in the annexes in Figure 9. This figure shows the distribution of the total length of species sections, measured by the number of characters. It's important to note that the content themes within these sections are arbitrary and can vary deeply from one species to another. Just like the graphic mentioned above, the most common length for the total of sections sits between 1 and 500 characters. However, unlike the introduction length graphic, the number of species decreases more gradually, as the section length increases.

2.10 Research Scenarios

During our data analysis, several key themes emerged, providing valuable insights into the study of species. By examining the most commonly referenced elements within the data, we gained a deeper understanding of the critical factors that impact both biodiversity research and conservation efforts.

One of the most important aspects identified was Conservation Status. Terms such as "endangered," "vulnerable," and "extinct" frequently appeared, showcasing the importance of tracking species' population trends and the urgency of conservation efforts. Research inquiries like "species at risk of extinction" or "species with declining populations" can offer crucial information for prioritizing protection strategies.

The theme of Discovery also was identified, focusing on the individuals who have contributed to identifying new species. This highlights the historical and scientific value of species discovery and classification, encouraging research into who first described certain species. For example, exploring "species discovered by [scientist]" allows for a deeper look into the historical context of taxonomic research.

The dataset also has frequent references to Scientific Classification, such as genus, family, and order, indicating the importance of understanding a species' biological categorization. Questions like "related species within the same genus" or "species classification hierarchy" can help in comprehending the evolutionary relationships between different organisms.

Finally, detailed Species Descriptions (Sections), including characteristics like size, color, and behavior, were often noted, emphasizing the role of specific traits in identifying and studying species. With these extra sections, the dataset can also answer more broad questions, such as, "venomous animal located in Southern Europe", or "species of fish with more than 2 meters, that frequents the Atlantic Ocean".

3 Conclusions and Future Work

In conclusion, the completion of all tasks in the data preparation phase marks a significant step forward in this project. This phase completion certifies that a well-organized and structured dataset is in place, which will serve as the basis for more detailed analysis.

A major challenge encountered during this phase was correctly processing species attributes from their respective Wikipedia page, such as handling irrelevant introductions and missing values, like the absence of conservation status on some pages.

As we transition to the next stage, efforts will be focused on further improving the project. With the foundational work completed, the next phase will focus on the development of a search engine that enables the exploration of all species attributes identified during this phase.

4 Annexes

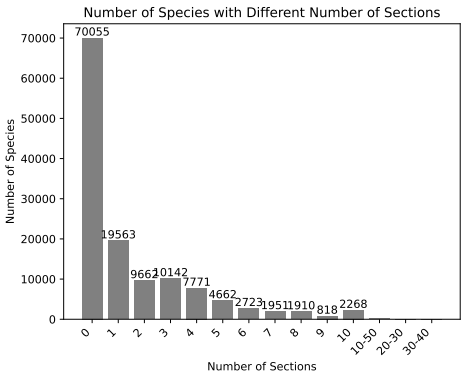

Figure 8: Species divided by number of sections.

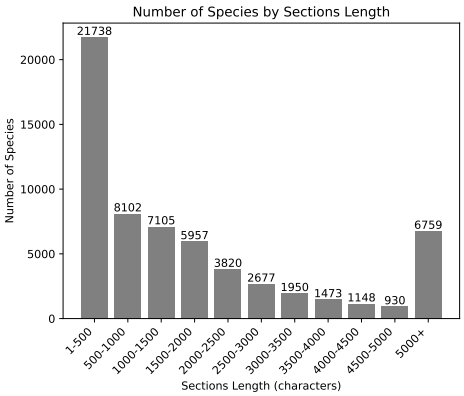

Figure 9: Species divided by total section length.

References

[1] 2012 – 2024. *Matplotlib Library.* https://matplotlib.org/ Accessed: 2024-10-08.
[2] 2024. *Banner image top 10 species.* https://www.kew.org/sites/default/files/styles/header_style/public/2024-01/Banner%20image%20top%2010%20species.jpg.webp?itok=xdgL0P4Q Accessed: 2024-10-09.
[3] 2024. *Beautifulsoup4 Library.* https://pypi.org/project/beautifulsoup4/ Accessed: 2024-10-07.
[4] 2024. *Pandas Library.* https://pandas.pydata.org/ Accessed: 2024-10-10.
[5] 2024. *Pythonic XML Library.* https://pypi.org/project/lxml/ Accessed: 2024-10-11.