Plants

Information Processing and Retrieval

Bárbara Rodrigues	Rúben Monteiro	Sofia Costa	Tiago Ribeiro
Faculdade de Engenharia	Faculdade de Engenharia	Faculdade de Engenharia	Faculdade de Engenharia
da Universidade do Porto	da Universidade do Porto	da Universidade do Porto	da Universidade do Porto
up202007163@edu.fe.up.pt	up202006478@edu.fe.up.pt	up202300565@edu.fe.up.pt	up202007589@edu.fe.up.pt

ABSTRACT

In a world overflowing with information there is the demand to find relevant data that aligns with a user's information needs.

This report describes methods for retrieving and processing data about "Plants", scrapped from Wikipedia [1]. After thorough data cleaning and preparation, the resulting dataset underwent exploration and characterization, revealing compelling and noteworthy insights.

KEYWORDS: data processing, data retrieval, data scraping, data analysis, data characterization, search system

1 INTRODUCTION

Within the scope of the college's Information Processing and Retrieval class, this report describes the work done for a three-part project. By the end of the fall semester, a working search system focused on the chosen topic has become available.

Following an extensive search for an interesting dataset rich in textual data, it was decided as a group that the topic for the project would be "plants". Subsequently, the data related to this topic was collected, prepared, explored, and processed by applying the concepts acquired throughout the semester.

2 DATA PREPARATION

The first milestone of the project, focused on the preparation and characterization of the data. This phase is highly dependent on the chosen topic's dataset which required extraction actions such as scraping.

The expected outcome from this initial phase is a well-documented and reproducible pipeline for data processing to simplify and enable subsequent milestones.

2.1 Data Selection

After having trouble deciding on a diverse range of prepared datasets, such as music lyrics or wine reviews, a consensus was reached collectively that the topic of the project would be focused on plants. This is a very text-rich and well-documented theme that is widely and easily available to anyone interested in researching it.

With this topic in mind, it was not hard to find an authentic source, as Wikipedia, the widely known encyclopedia, already had many wide lists full of information, ready to gather and use. The specific list is "List of plants by common name" [2]. Wikipedia content is typically available under a Creative Commons license, allowing for the use and redistribution of the data.

Each entry of the list provides a table of simple scientific information regarding the plant, along with multiple and diverse paragraphs written about the plant's characteristics and details. The list bears around 400 unique values, each one containing one or more paragraphs with information about the plant.

Finally, to collect all this data, the gathering process chosen was web scraping. By making use of Python [3] scripts and the "Beautiful Soup" library [4], it was possible to collect all the information relative to each plant of the list that had a link attached to it.

The information extracted was then stored in a CSV file, ready for the next part of the pipeline.

2.3 Data Processing

Upon completing the data collection phase and performing a careful data analysis, a series of processing tasks were executed to improve the quality of the data. All the steps taken to reach a clean and complete dataset are illustrated in **Figure 1 in Appendix A.1**, which represents the structured and reproducible pipeline of the project.

Firstly, the incongruence in column names was noticeable. For example, some column labels end in ":" such as the "Kingdom" column which appears as "Kingdom:", so all these values were rewritten to appear with the same representation.

Another challenge faced was the inconsistent organization of data across different articles. Specifically, similar types of information were often presented in varying formats and headings. For instance, columns "Habitat", "Habitat and range" and "Habitat and distribution”: those columns were aggregated and combined to make only one, in this specific case, the “Habitat” column. This was the main step to ensure the homogeneity of the dataset. These variations were merged and reconciled, harmonizing the data by aligning these sections under a common label. Initially, the dataset had 178 columns but after merging and transforming, it was reduced to only 63 columns.

Following data aggregation involved removing irrelevant columns for future exploration. These columns either had a significant number of “NaN” values or offered a very limited amount of data, leading to the removal of 28 unnecessary columns.

Upon closer examination of certain column contents, namely the “Introduction” and “Description” columns, details about the geographical origin of the plants were uncovered. Given this discovery, it was decided to extract this information and organize it into a fresh column named “Origin Country”.

Preceding saving the transformed dataset, the “Name” column was modified in only three cases. In contrast to other samples where the “Name” column had just the plant name, these specific instances had lengthy text. These cases were manually adjusted to include only the original name by reading the “Name” field.

The data processing phase was essential for creating a cohesive and structured dataset, facilitating subsequent analysis, and ensuring that our data remained consistent and easy to work with. Therefore, the cleanup, processing, and transformation of the original dataset was completed, resulting in a final set of 36 columns containing information that is believed to be crucial for the ongoing progress of the project.

2.4 Data Characterization

Following the data cleaning and refinement process, an exploration analysis of the dataset was conducted, to gain a deeper understanding of the data extracted. Using Jupyter Notebooks [5], it was possible to create plots such as histograms, pie charts, and word clouds for a total of 403 rows and 36 columns. Each row is equivalent to the number of plant species scrapped from the Wikipedia list [reference].

Exploring the Scientific Classification of plants seemed like an interesting task to retrieve some insights about the dataset. This classification system includes groups such as Kingdom, Clade, Order, Family, Genus, Species, and other subgroups.

Table 1 - Count of unique values for the Scientific Classification column

Group	Count
Kingdom	1 (*Plantae*)
Clade	1 (*Tracheophytes*)
Order	43
Family	87
Genus	203
Species	299

Focusing on the “Order” and “Family” groups **(refer to Table 1)**, since they had smaller values to work with, two histogram graphs were generated to illustrate the distribution of plant species for each value within each of these classifications.

By observing **Figure 2**, it can be concluded that the “Order” with most plant species is called “*Asterales*”, which according to Wikipedia, brings together “dicotyledonous plants, like marigolds, daisies, and sunflowers” [6]. Additionally, according to **Figure 3 in Appendix A.2**, the “Family” with more plant species is “*Asteraceae*” which “consists of over 32,000 known species of flowering plants in over 1,900 genera within the order *Asterales*” [7]. These results suggest that flowers are the dominating plant of the dataset.

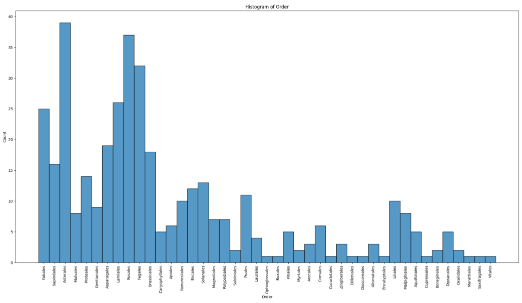

Figure 2 - Histogram for the ‘Order’ column

Since every column of the dataset is rich in textual data, the data exploration was limited to text analysis. Therefore, word clouds were generated, for any relevant column, as well as finding out the frequency of specific words.

By generating a word cloud for the column related to the plant’s name, it can be concluded which species dominates the list.

By analyzing **Figure 4**, four notable *Genus* within the “Name” column are identified, specifically *Quercus*, *Lambertia*, *Allium*, and *Acer*. These designations correspond to the first words in species names, representing the respective genera. *Quercus*, also known as oak trees, is observed 17 times in the dataset.

Lambertia, recognized as "wild honeysuckle," appears 10 times. *Allium*, the Latin term for garlic, is recorded seven times. Lastly, *Acer*, the genus for maples, is documented six times in the respective column.

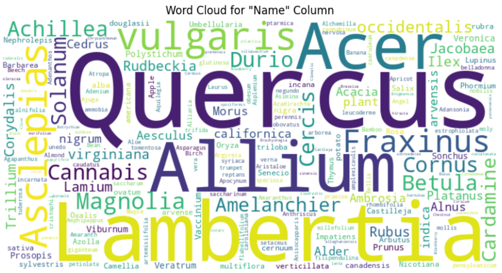

Figure 4 - Word Cloud for the "Name" column

More word clouds were generated for each column but only some of them presented relevant results. By generating a word cloud for the column related to the origin country **(refer to Figure 5 in Appendix A.3)** of each plant, it was possible to identify the continent or country where most plants originated from. North America and Australia were the most mentioned native locations.

As for the medicine column (**refer to Figure 6 in Appendix A.4**) the words that stand out the most are "antioxidant" and "disease" which can imply the connection of plants with potential therapeutic benefits and disease prevention.

Since plants can have multiple different characteristics, an interesting task for data analysis was to extract the number of mentions of specific colors. The generation of a pie chart **(refer to Figure 7)** with the count of each color mentioned led to the identification of the most predominant colors in the whole dataset, which are: red, green, and white.

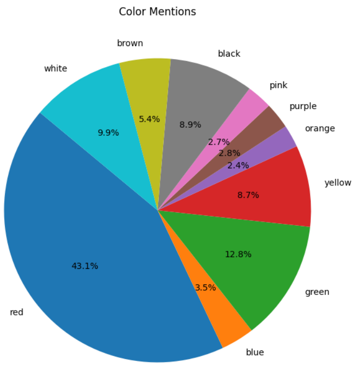

Figure 7 - Pie Chart of color mentions count

3 PROSPECTIVE SEARCH TASKS

In this section, a few prospective search tasks are outlined for a comprehensive plant information search system. Each task aims to address specific information needs the user might have, to make their search experience easier and more efficient.

- **Interest in plants with medicinal properties.**

Users will be able to find information about the therapeutic uses, active compounds, and historical applications of various plants.

- **Interest in learning about a plant's toxicity.**

Users will be able to access information related to the toxicity of different plant species. They can learn about poisonous plants, potential health risks, and necessary precautions.

- **Interest in finding out where a plant is from.**

Users will be able to discover the geographic origin and native habitats of various plant species. They can explore the natural distribution and historical context of plants.

- **Interest in learning which plants are edible.**

Users will be able to access information on edible plants, including details on culinary uses, nutritional benefits, and safe consumption practices.

REFERENCES

[1] Wikipedia, "Wikipedia, The Free Encyclopedia" Wikipedia, 2023. [Online]. Available: https://www.wikipedia.org/. [Accessed in September 2023].

[2] "List of plants by common name" [Online]. Available: https://en.wikipedia.org/wiki/List_of_plants_by_common_name . [Accessed in September 2023].

[3] "Welcome to Python.org" Python, [Online]. Available: https://www.python.org/. [Accessed in October 2023].

[4] "beautifulsoup4" PyPI, [Online]. Available: https://pypi.org/project/beautifulsoup4/. [Accessed in September 2023].

[5] "Project Jupyter" Jupyter, [Online]. Available: https://jupyter.org/. [Accessed in September 2023].

[6] "Asterales" Wikipedia, 18 March 2023. [Online]. Available: https://en.wikipedia.org/wiki/Asterales. [Accessed in October 2023].

[7] "Asteraceae" Wikipedia, 24 August 2023. [Online]. Available: https://en.wikipedia.org/wiki/Asteraceae. [Accessed in October 2023].

A APPENDICES

A.1 Pipeline

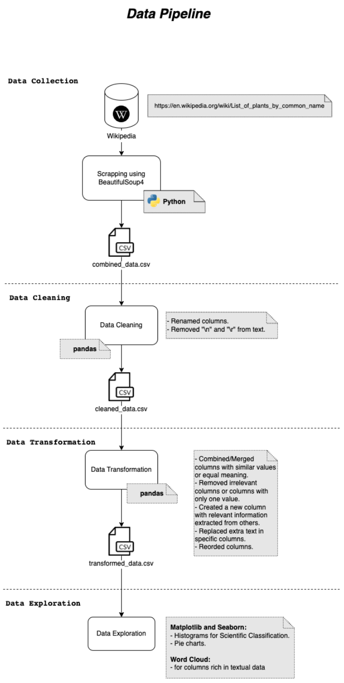
Data Pipeline
Data Collection
https://en.wikipedia.org/wiki/List_of_plants_by_common_name
Wikipedia
Scrapping using BeautifulSoup4
Python
CSV
combined_data.csv
Data Cleaning
Data Cleaning
- Renamed columns.
- Removed "\n" and "\r" from text.
pandas
CSV
cleaned_data.csv
Data Transformation
Data Transformation
pandas
- Combined/Merged columns with similar values or equal meaning.
- Removed irrelevant columns or columns with only one value.
- Created a new column with relevant information extracted from others.
- Replaced extra text in specific columns.
- Reorded columns.
CSV
transformed_data.csv
Data Exploration
Data Exploration
Matplotlib and Seaborn:
- Histograms for Scientific Classification.
- Pie charts.
Word Cloud:
- for columns rich in textual data

Figure 1 - Data Pipeline

A.2 Histogram for Family

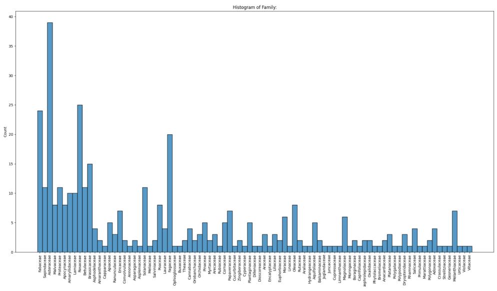

Figure 3 - Histogram for the 'Family' column

A.3 Word Cloud for the column "Origin Country"

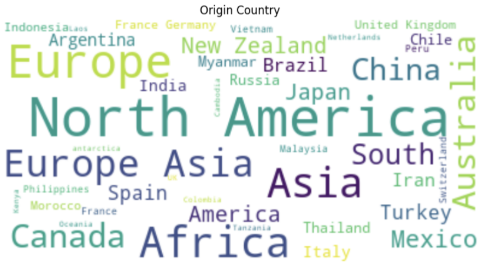

Figure 5 - Word Cloud for the 'Origin Country' column

A.4 Word Cloud for the column 'Medicine'

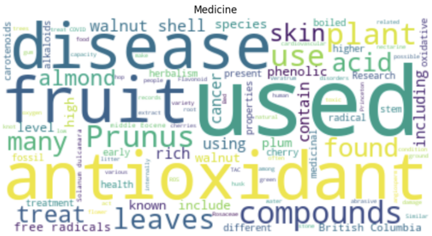

Figure 6 - Word Cloud for the 'Medicine' column