PRI: Building an Information Retrieval System using University Data

Ilina Kirovska
up202301450@fe.up.pt

Gonçalo Almeida
up202308629@fe.up.pt

Žan Žlender
up202302230@fe.up.pt

Abstract

The primary objective of the project is the development of an information retrieval system, prioritizing comprehensive exploration of European higher education. A collection of documents was obtained through data collection and data cleaning processes. Furthermore, in order to better understand the quality and context of the data, a more in-depth analysis with visual representations was performed.

CCS Concepts: • **Information systems → Information retrieval**.

Keywords: data processing, datasets, data preparation, full-text search, conceptual data model, information retrieval system

1 Introduction

University rankings offer extensive assessment of higher education institutions. They are powerful tools, providing insightful analysis of the quality, performance and general credibility of universities worldwide. Furthermore, they help students in finding educational institutions that align with their academic interests.

The Quacquarelli Symonds (QS) World University Rankings [3], debuted in 2004 and has since grown to become the foremost source of comparative data on university performance. As a result, we chose the 'QS World University Rankings 2024' Kaggle dataset [2] to be the base of our data. The dataset was then enriched using Wikipedia [5] and Wikidata [4] APIs, through which we extracted data about the universities and the cities they are located in.

The goal of the project is to build an information retrieval system by utilizing the extensive data made accessible through these APIs and the dataset. This document focuses on the initial stage of the project, the data preparation. In the following sections, we are going to thoroughly describe how we collected and cleaned the data, organised it into documents and explored it using various types of analysis.

2 Data Collection and Preparation

The scope of this project includes all European universities. As such, the first step was to filter out the original University ranking dataset to only get the European ones. For that purpose, we used the Python library Pandas to load and filter the universities by their country of origin ISO code, by comparing them to the European countries' .csv file[1]. After filtering, this dataset included 529 universities located throughout Europe, including ranking information on their campuses and faculties.

At this point, we encountered a problem, which was that some university names were written in their local language instead of English. Since the idea was to use the English Wikipedia API, which searches by the name of the university, the second step was to manually translate all the incorrectly written names into English.

The following step included fetching the WikiData information for that university. With this information, we could identify any problematic entities which cannot be found, as well as get the relevant information about them. The most important being the WikiData identification number for that university. Only one university was not found because it had the symbol & in its name, which was quickly solved by exchanging the symbol with its UTF-8 representation %26.

After that, we could safely start getting the actual information we wanted, which was the plain text of the universities' Wikipedia pages. Using the Wikipedia API and the university names, as well as a Wikitext parser library, we quickly managed to get the text we wanted.

The next step was to include the cities' information where the universities are located. The first step here was to identify how we could achieve that. Since the WikiData is fetched in the third step we can easily query all of that university's information stored in WikiData. Since all the data is structured it includes many useful properties and relations, with the relevant ones being: located in the administrative territorial entity (P131) and location (P276). Using those two properties we could then fetch that city's WikiData information, which would return us the city's name, which could finally be used to get that city's Wikipedia page plain text as well.

Finally, during the analysis, we determined that there were a few columns which were irrelevant or had too many incorrect or missing values, so we removed them.

After the final conversion, the dataset included 529 universities with a file size of around 33MB, as a JSON file.

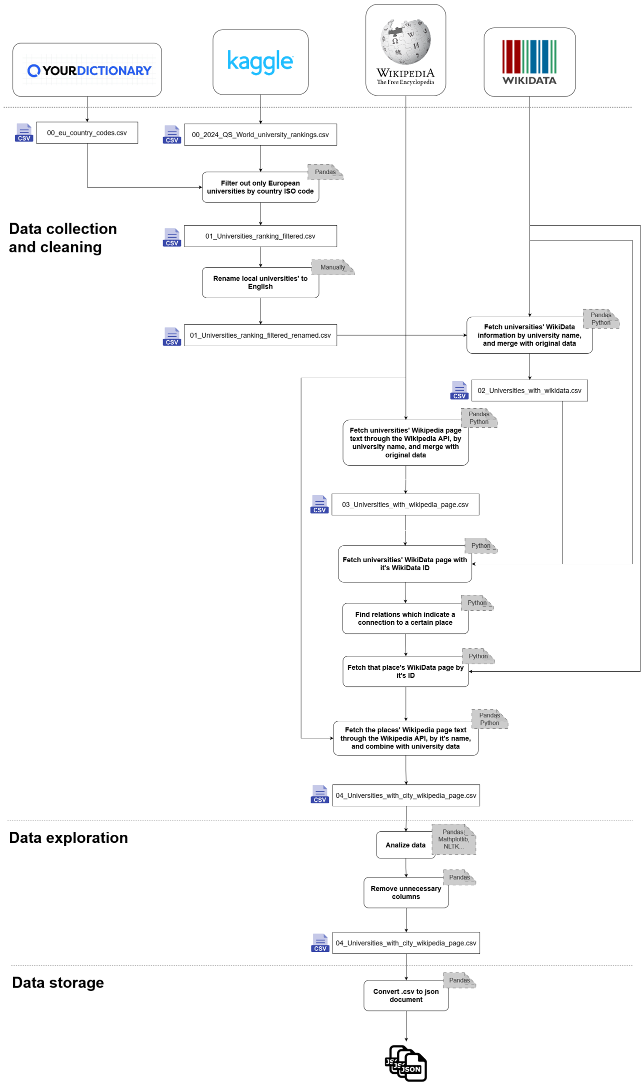
YOURDICTIONARY
kaggle
Wikipedia The Free Encyclopedia
WIKIDATA
Data collection and cleaning
CSV
00_eu_country_codes.csv
00_2024_QS_World_university_rankings.csv
Filter out only European universities by country ISO code
Pandas
01_Universities_ranking_filtered.csv
Rename local universities' to English
Manually
01_Universities_ranking_filtered_renamed.csv
Fetch universities' WikiData information by university name, and merge with original data
Pandas Python
02_Universities_with_wikidata.csv
Fetch universities' Wikipedia page text through the Wikipedia API, by university name, and merge with original data
Pandas Python
03_Universities_with_wikipedia_page.csv
Fetch universities' WikiData page with it's WikiData ID
Python
Find relations which indicate a connection to a certain place
Python
Fetch that place's WikiData page by it's ID
Python
Fetch the places' Wikipedia page text through the Wikipedia API, by it's name, and combine with university data
Pandas Python
04_Universities_with_city_wikipedia_page.csv
Data exploration
Analize data
Pandas, Mathplotlib, NLTK...
Remove unnecessary columns
Pandas
04_Universities_with_city_wikipedia_page.csv
Data storage
Convert .csv to json document
Pandas
JSON

Figure 1. Data collection and preparation pipeline.

3 Conceptual Data Model

The conceptual data model represents the relationship between each entity in the domain. In this case, the model is straightforward since most of the information is centred around the university. The university information includes its rank, as well as past year's ranking. Additionally, there are many metrics which determine the rank, like Employer Reputation Score, Employer Reputation Rank, Faculty Student Score, Faculty Student Rank etc. One of the most important metrics is the **Overall SCORE**. Although we had extracted different information about the university, for example, the Wikipedia text, it's still connected to a specific university. Finally, each university is located in a city, which has its name as well as the City Wikipedia text, extracted from its Wikipedia web page.

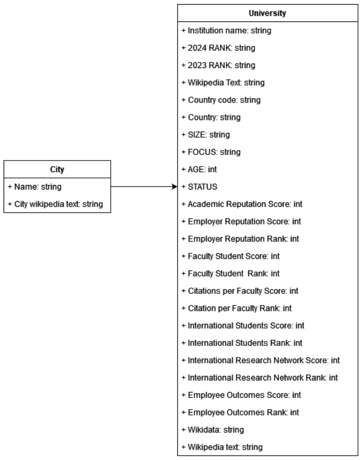
University
+ Institution name: string
+ 2024 RANK: string
+ 2023 RANK: string
+ Wikipedia Text: string
+ Country code: string
+ Country: string
+ SIZE: string
+ FOCUS: string
+ AGE: int
+ STATUS
+ Academic Reputation Score: int
+ Employer Reputation Score: int
+ Employer Reputation Rank: int
+ Faculty Student Score: int
+ Faculty Student Rank: int
+ Citations per Faculty Score: int
+ Citation per Faculty Rank: int
+ International Students Score: int
+ International Students Rank: int
+ International Research Network Score: int
+ International Research Network Rank: int
+ Employee Outcomes Score: int
+ Employee Outcomes Rank: int
+ Wikidata: string
+ Wikipedia text: string
City
+ Name: string
+ City wikipedia text: string

Figure 2. Conceptual data model

4 Data Characterization

Understanding the collected data and identifying its key features is of great importance when building an information retrieval system. This was accomplished by doing additional data analysis, through which we gained more insights about the structure and context of the data. The results, along with proper visualisation, are explained in the following subsections.

4.1 Initial analysis of the acquired dataset

In the beginning stages of the process, a basic statistical analysis of the Kaggle dataset [2] was done. The aim was to identify how many of the universities fit our criteria of being a European university. As seen in Figure 3, 35.34% of all universities were in Europe and hence, kept in the dataset.

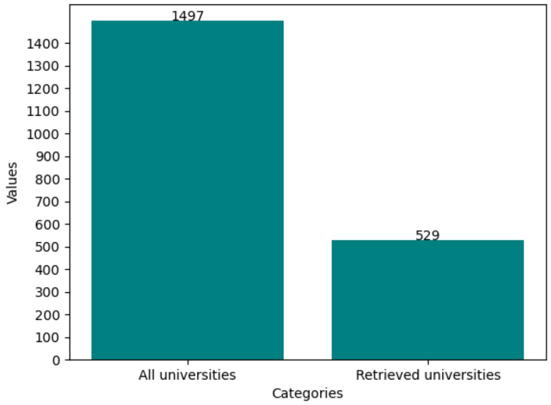

Figure 3. Number of all universities versus the number of the universities that were kept.

4.2 Location analysis

As our focus is on European universities, their distribution over Europe is vital information. Therefore, it is illustrated using a variety of visual representations.
Figure 4 shows that the countries with the most universities are the United Kingdom (90), Germany (49) and Russia (48).

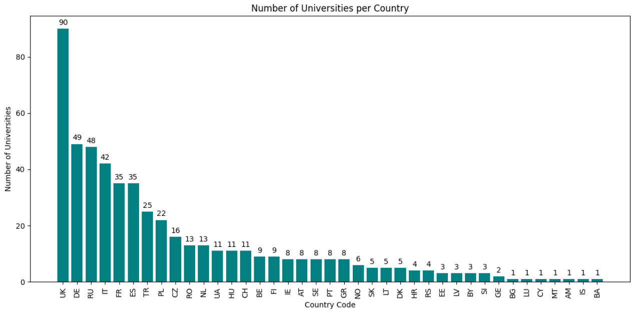

Figure 4. Number of universities per country.

Moving on to geographic visualization, Figure 5 is a map that displays the cities where universities are located. Each city is pinpointed on the map, providing a clear spatial perspective of the university distribution.

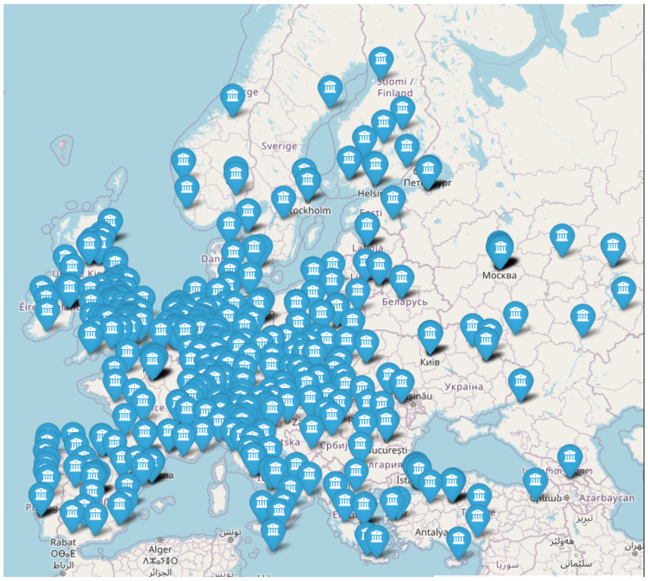

Figure 5. Map of the universities analysed.

Figure 6 is a heat map that visualizes the distribution of universities across Europe, and it helps identify clusters of universities in specific regions.

4.3 Ranking analysis

Figure 7 shows the changes in university rankings over time, by comparing the university ranks in 2024 (2023/2024) and

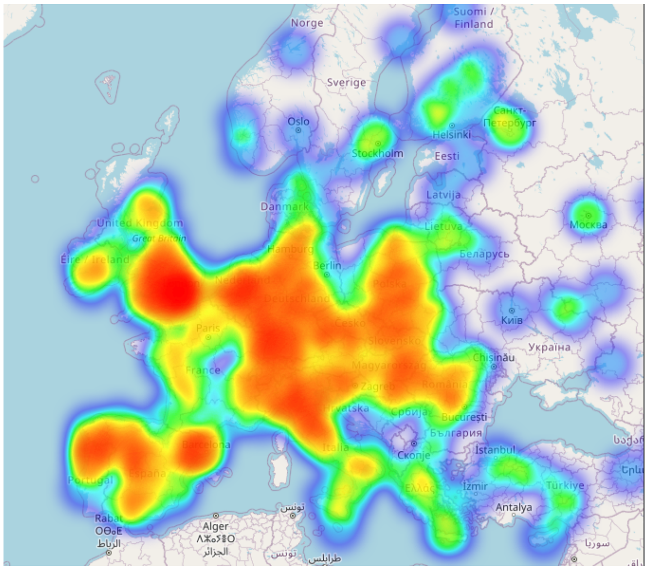

Figure 6. Heat map of the universities analysed

2023 (2022/2023). This dynamic visualization highlights universities that have experienced rank changes, such as improvements or declines, between the two years.

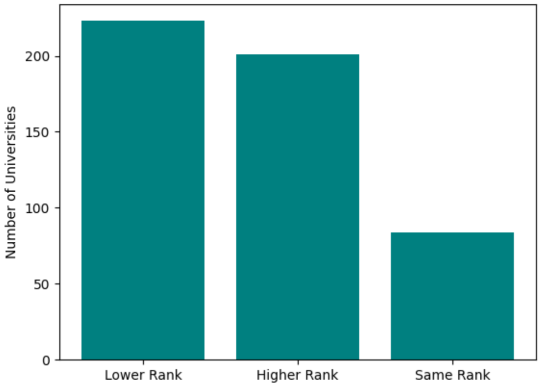

Figure 7. University rank comparison - 2024 vs 2023.

4.4 Size and Age Analysis

Figure 8 illustrates the number of universities falling into different size categories (XL, L, M, S) while considering their age, which ranges from new (Category 1) to older (Category 5). The teal-coloured bars represent the count of universities in each size category, with age categories distinguished by different shades within each bar. This visualization helps us understand how university sizes vary across different age categories.

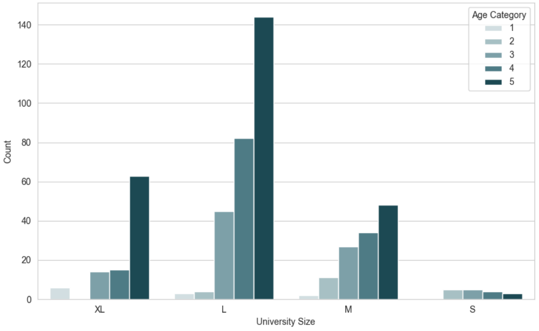

Figure 8. Distribution of university sizes by age category.

4.5 Text analysis

The majority of the data in our system is stored as plain text, hence one of the most efficient ways to analyse it is to identify the most common words. Through them, we can easily find out the tone and content of the text. We used the NLTK library [6] which offers NLP techniques. After the text was preprocessed using NLTK, we calculated the word frequencies and constructed the wordclouds. As shown in Figure 9 some of the most common words in the universities' text data were university (appears 31918 times), student (9584 times), faculty (8344 times) and science (7946 times).

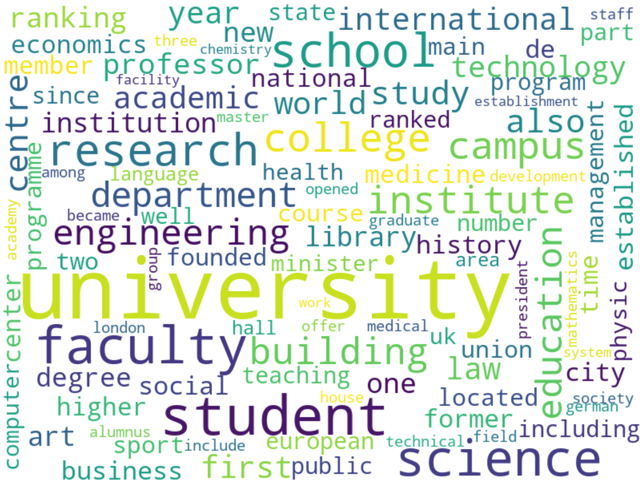

Figure 9. Wordcloud of the most common words in the extracted university's Wikipedia text.

From Figure 10 we can see that in the cities' data, the most common words were city (41723 times), also (10365 times), one (8811 times) and century (8583 times).

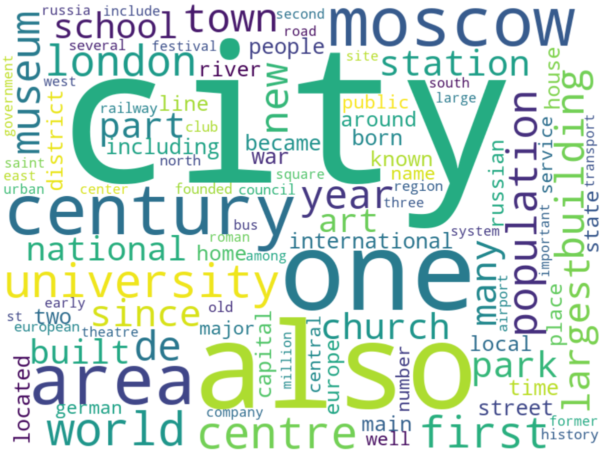

Figure 10. Wordcloud of the most common words in the extracted city's Wikipedia text.

5 Prospective search tasks

The focus of this search engine is to provide information about the available universities. Taking that into consideration, these are some of the possible search tasks:

- Search for universities by country or by city
- Search for universities by rank, size, age
- Search for the best university in a given country/city
- Search for information about a city by university
- Search for universities/cities by using the most common words(keywords) from the Wikipedia text

6 Conclusion

This report details the successful completion of the project's first milestone, which focused primarily on dataset preparation and exploration. The dataset that was used for this research includes all of the basic information about European universities, including rankings, locations, and institutional characteristics. Wikidata and Wikipedia API were used to streamline the data-collecting process and ensure data relevance and accuracy for the study. University rankings and other properties were standardized to produce a reliable dataset for future analysis.

References

[1] [n. d.]. List of European countries. Retrieved October 10, 2023 from https://www.yourdictionary.com/articles/europe-country-codes

[2] [n. d.]. QS World University Rankings 2024 Dataset. Retrieved September 28, 2023 from https://www.kaggle.com/datasets/joebeachcapital/qs-world-university-rankings-2024/data

[3] [n. d.]. Quacquarelli Symonds (QS) World University Rankings. Retrieved October 10, 2023 from https://www.qs.com/

[4] [n. d.]. Wikidata API. Retrieved October 10, 2023 from https://www.wikidata.org/wiki/Wikidata:REST_API

[5] [n. d.]. Wikipedia API. Retrieved October 10, 2023 from https://www.mediawiki.org/wiki/API:Main_page#Endpoint

[6] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.*