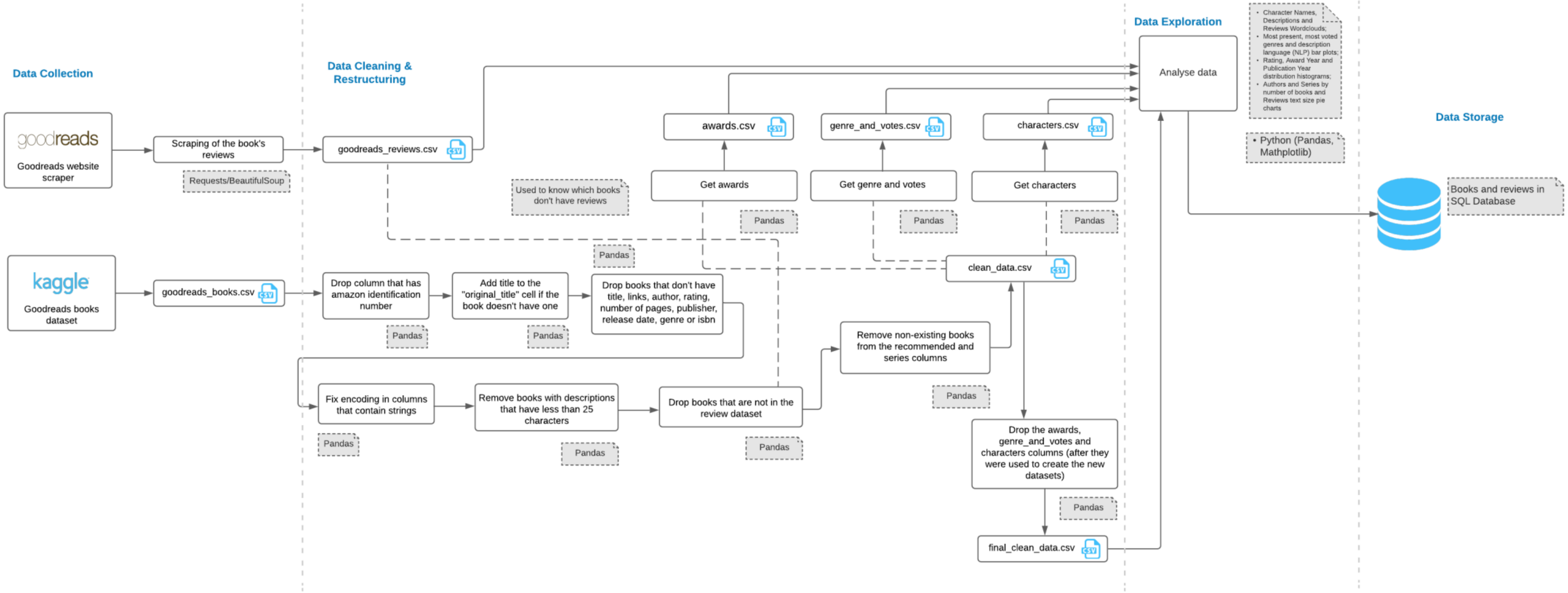
Data Preparation Process

GOODREADS' BOOKS AND REVIEWS

Diogo Almeida (up201806630)
Pedro Queirós (up201806329)

Data Extraction and Preparation



Data Collection

goodreads

Goodreads website scraper

Scraping of the book's reviews

Requests/BeautifulSoup

kaggle

Goodreads books dataset

goodreads_books.csv

Data Cleaning & Restructuring

goodreads_reviews.csv

Used to know which books don't have reviews

Drop column that has amazon identification number

Pandas

Add title to the "original_title" cell if the book doesn't have one

Pandas

Drop books that don't have title, links, author, rating, number of pages, publisher, release date, genre or isbn

Pandas

Fix encoding in columns that contain strings

Pandas

Remove books with descriptions that have less than 25 characters

Pandas

Drop books that are not in the review dataset

Pandas

Remove non-existing books from the recommended and series columns

Pandas

clean_data.csv

Get awards

Pandas

awards.csv

Get genre and votes

Pandas

genre_and_votes.csv

Get characters

Pandas

characters.csv

Drop the awards, genre_and_votes and characters columns (after they were used to create the new datasets)

Pandas

final_clean_data.csv

Data Exploration

Analyse data

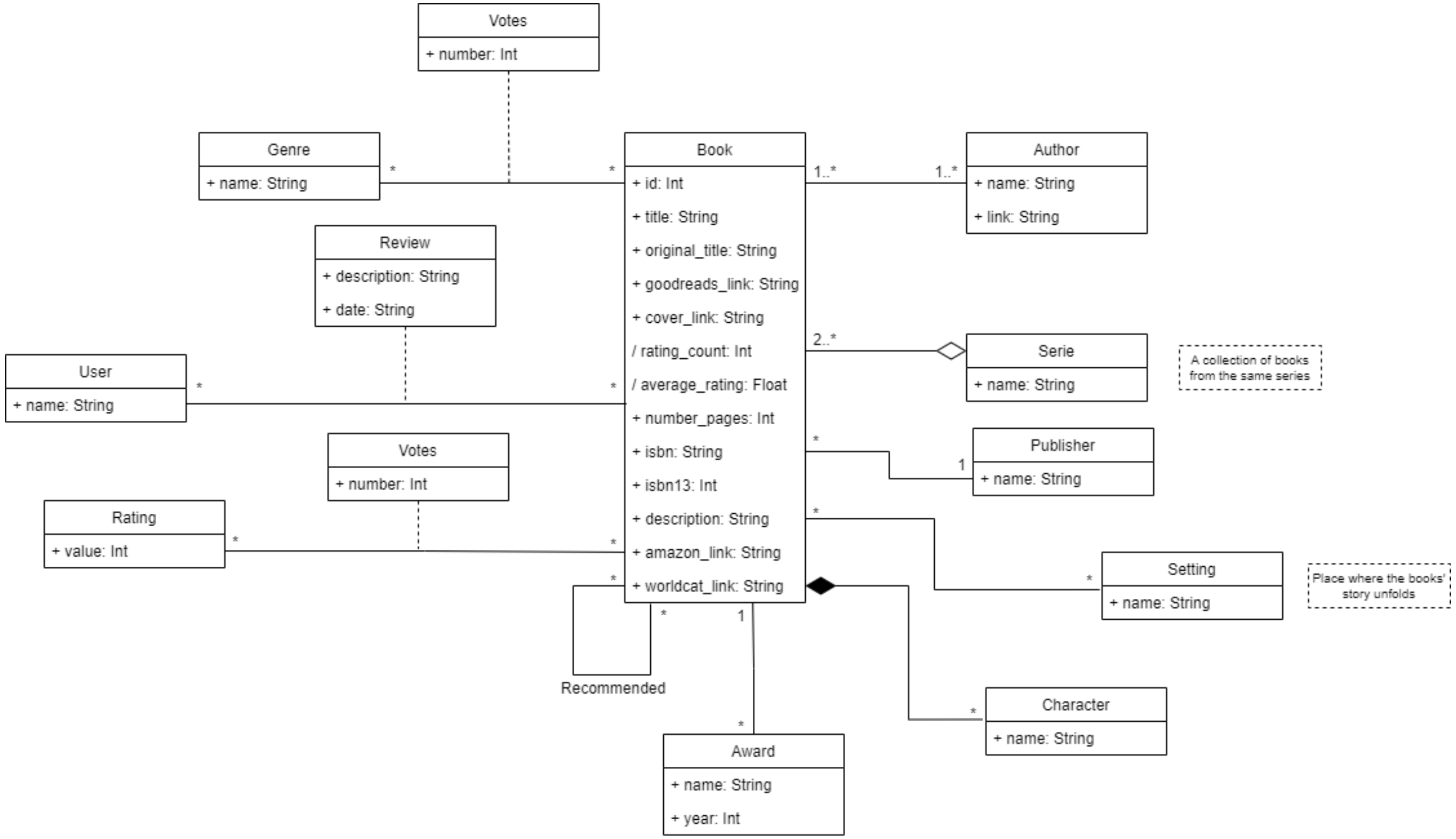
- Character Names, Descriptions and Reviews Wordclouds;
- Most present, most voted genres and description language (NLP) bar plots;
- Rating, Award Year and Publication Year distribution histograms;
- Authors and Series by number of books and Reviews text size pie charts

- Python (Pandas, Mathplotlib)

Data Storage

Books and reviews in SQL Database

Domain Conceptual Model



Votes
+ number: Int

Genre
+ name: String

Book
+ id: Int
+ title: String
+ original_title: String
+ goodreads_link: String
+ cover_link: String
/ rating_count: Int
/ average_rating: Float
+ number_pages: Int
+ isbn: String
+ isbn13: Int
+ description: String
+ amazon_link: String
+ worldcat_link: String

Author
+ name: String
+ link: String

Review
+ description: String
+ date: String

User
+ name: String

Serie
+ name: String

A collection of books from the same series

Votes
+ number: Int

Publisher
+ name: String

Rating
+ value: Int

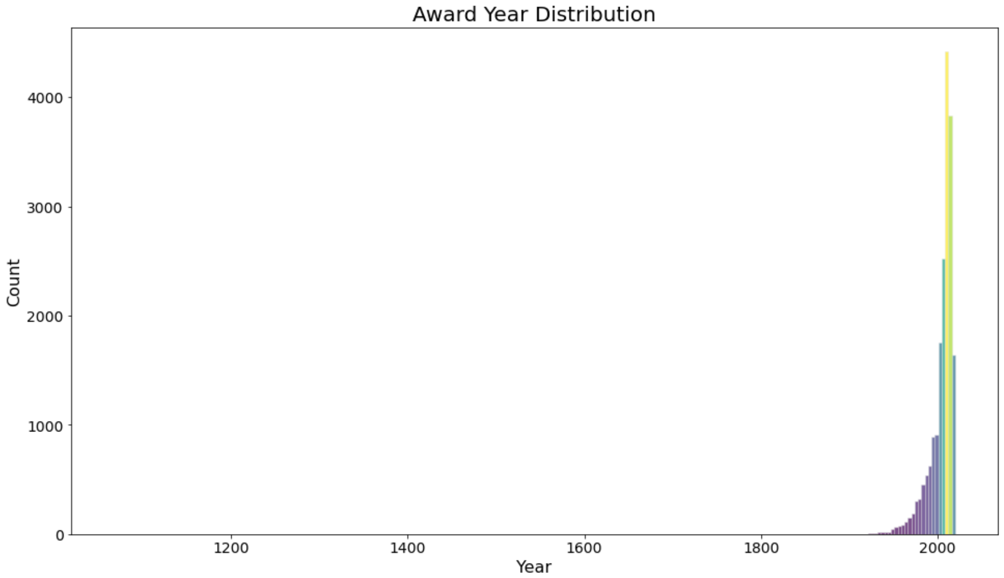
Setting
+ name: String

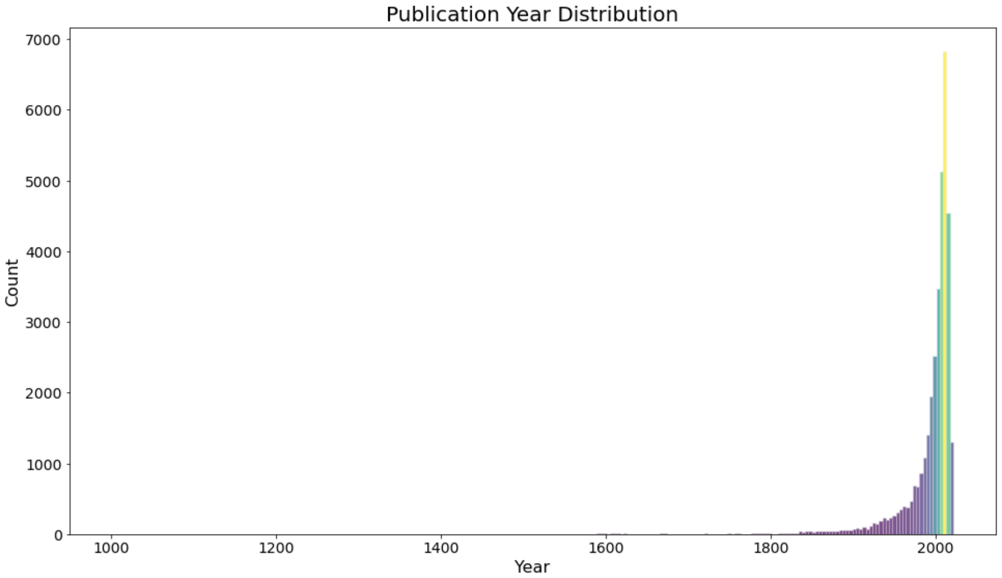
Place where the books' story unfolds

Character
+ name: String

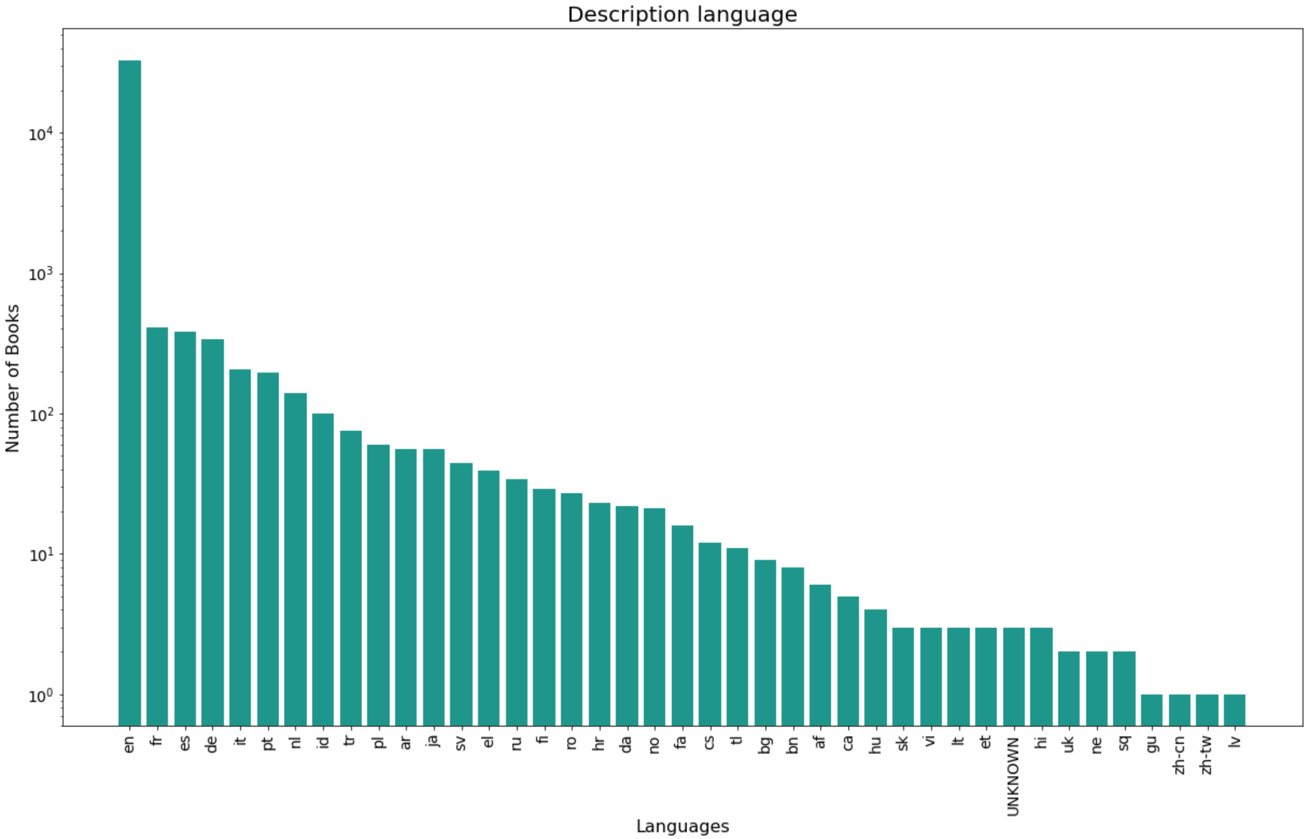
Award
+ name: String
+ year: Int

Data Characterization

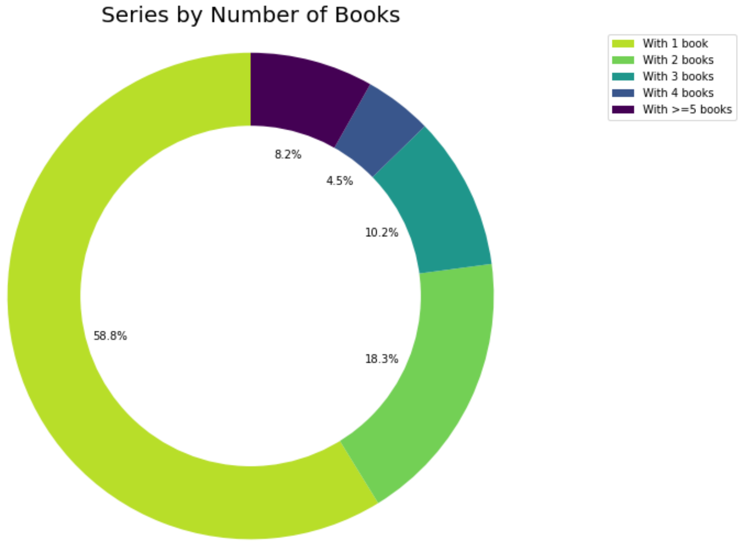


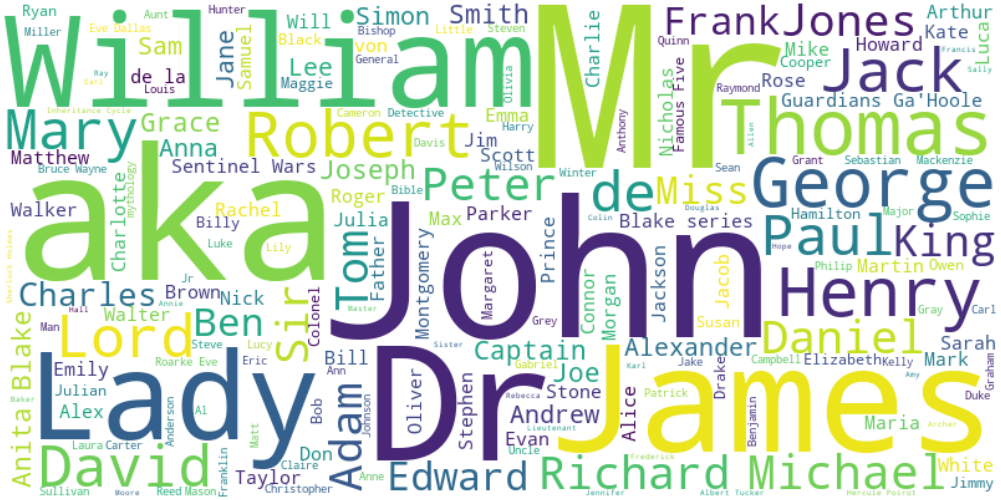


Data Characterization



Data Characterization





Conclusions and Future Work

For the extraction and preparation of the information the following steps were made:

- Gathering data and enrich it with data from diferentes sources
- Data cleaning
- Analyse the final data

Regarding future work, some of the implemented search tasks will be:

- Search books by author
- Search books based on their language
- Search books by genre, awards or even by their reviews or rating

References

❖“Goodreads Books – 31 features” dataset:

- https://www.kaggle.com/austinreese/goodreads-books

❖Goodreads website:

- https://www.goodreads.com

❖Python libraries: pandas, requests, beautifulsoup and spacy:

- https://pandas.pydata.org/
- https://docs.python-requests.org/en/latest
- https://beautiful-soup-4.readthedocs.io/en/latest
- https://spacy.io/universe/project/spacy-langdetect