Pitchfork Music Reviews

Fabio Huang
Faculty of Engineering of the
University of Porto
Porto, Portugal
up201806829@up.pt

Lisa Sonck
Faculty of Engineering of the
University of Porto
Porto, Portugal
up202202272@up.pt

Temaco
Faculty of Engineering of the
University of Porto
Porto, Portugal
up202200606@up.pt

Abstract

Information communication technologies are ubiquitous in modern societies. An ever-growing number of activities depend on the ability to extract value from information. The goal of this first milestone is to prepare, explore and characterize a selected data-set. As a result, a reproducible pipeline of data processing is achieved. In our case, a data-set about music reviews on Pitchfork (site best known for its daily output of music reviews) was retrieved from Kaggle. The data is analysed and explored using OpenRefine, which was then cleaned and refined using Python's Pandas Library [2]. To characterize the data, graphs about the data were formed and analysis was conducted.

Keywords: Information Processing and Retrieval, Datasets, Pitchfork Music Reviews

1 Introduction

The amount of available data is growing every year [1] and with it, the importance of its correct management. Being able to efficiently use the data and extract interesting results is essential for further development of our current society. The course Information Processing and Retrieval at FEUP aims to enlighten us about information search systems and their use for large amounts of data. To put this knowledge into practice, we were tasked to develop a search system, with the corresponding tasks of data collection and preparation, and the information processing and retrieval. This project is divided into three milestones: 1. Data Preparation, 2. Information Retrieval and 3. Building the final Search System.
This report elaborates on the results and actions of the first milestone, which include:

- Choosing a dataset.
- Assessing the data quality and the authority of the source.
- Building a data processing pipeline.
- Cleaning of the data.
- Making a conceptual model of the data.
- Performing an exploratory data analysis.
- Characterizing the data.

2 Data Collection

Developing a information search system requires a good and extensive dataset. The dataset should contain a lot of rows and columns, where at least one has an exhaustive, diverse body. This way text search can be implemented and interesting results can be obtained.

2.1 Dataset Choice

The choice of a dataset is important, since it determines the course of the project. At first the theme 'Music' was chosen, since a lot of large and comprehensive datasets exists for this subject and we have a big interest in music and its characteristics. We first found datasets were about artists, albums and songs, and the lyrics associated with each song, however, due to copyright reasons we were not able to get the complete content of the lyrics, and faced the problem of lacking textual data to work with. Luckily we managed to find a dataset that contained music reviews, which consisted of interesting comments and rich text that could be used for this project. [4]

2.2 Dataset Content

The selected dataset consists of a .sqlite file that contained the following tables:

- Artists table: contained the name of the artists (18.8k rows)
- Content table: contained the review's content (18.4 rows)
- Genres table: contained the song's genre (22.7k rows)
- Labels table: contained the album's label (20.2k rows)
- Reviews table: contained the title of the review, name of the artist, url link for the pitchfork's review page, rating score of the review, best new music, review's author name, author type, review's publication date and week day (18.4k rows)
- Years table: contained the year that the song was released (19.1k rows)

2.3 Data Quality and Source

Our dataset was retrieved from Kaggle [4], which is a popular community for data researchers and analysts and where the vast majority of the content is reliable. All of the content of this dataset was scraped from Pitchfork and since this dataset has over 10k downloads and over 8000 upvotes, we can infer that this source is reliable. The quality of the data meets our usage standard.

3 Pipeline

To build our Pipeline, we created python scripts using libraries such as Pandas [2], Matplotlib [5] and Seaborn [6] to manipulate the data and to generate graphs. OpenRefine [7] was also used to check for data facets. Our pipeline begins with a .sqlite file extracted from Kaggle, which contains multiple tables, as mentioned before. These tables are merged and a .json file is produced. This file is then processed and cleaned. At the end of the pipeline, the final dataset is used for data characterisation.

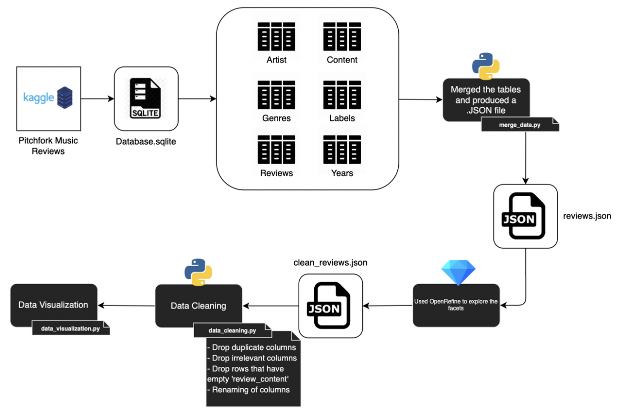

Figure 1. Pipeline Diagram

4 Data Cleaning

Before cleaning the data, we used OpenRefine [7] to explore the text facets on each column. After having an overall idea of the values of each column, duplicated and irrelevant columns were removed. Also entries with an empty review content were removed. At last some columns were renamed for an easier interpretation of the values.

5 Conceptual Model

The conceptual model is based on the initial database, extracted from Kaggle. However, some of the attributes are removed after the data was refined and cleaned. The final model can be seen in figure 2.

6 Dataset Characterisation

To better understand and grasp the data at hands, analysis and characterisation of the dataset was performed. To execute and visualize the characteristics of the dataset, the language Python was used with the packages Pandas [2] and Seaborn [6]. Pandas is very useful to handle big amounts of structured data and with Seaborn, graphs can be made easily. At first all the fields in the data set were investigated in a general way; Look at the possible values, what is the average value, are their outliers, etc. This initial research showed that the average score, given in a review, is 7.0 and the median is 7.2.

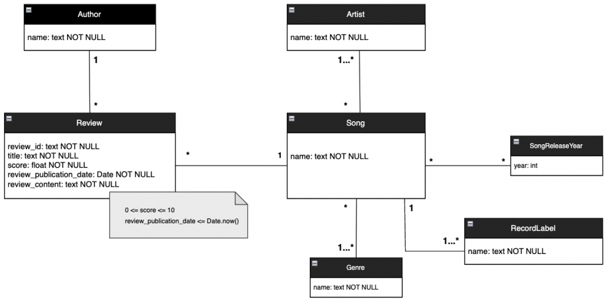

Figure 2. Conceptual Model

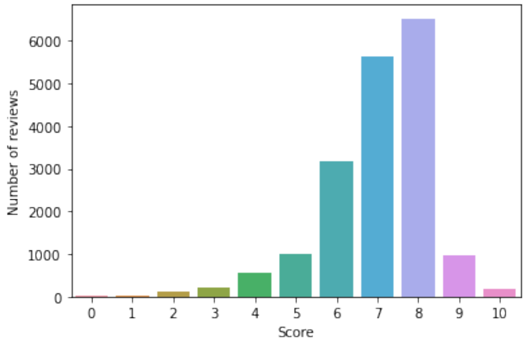

Figure 3. Number of reviews per rounded score

In figure 3 the number of reviews per score are represented. For every review the score is rounded to the nearest integer and this shows that the score eight is given the most, followed by seven.

After investigation of the genre of the songs of the different reviews, we concluded that most of the reviews are written of rock songs. This is seen in figure 4.

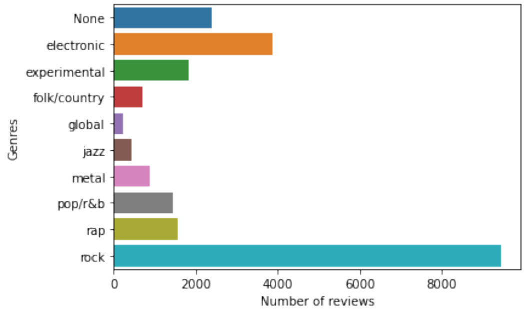

Figure 4. Number of reviews per song-genre

This first analysis about the data gave a general overview about its characteristics and how our data set was built. To see if certain fields of the data had a correlation, cross analysis over the different fields was performed. Figure 5 shows that the average score is approximately the same for every genre and equal to 7.0, which also establishes the analysis of the score over all the reviews.

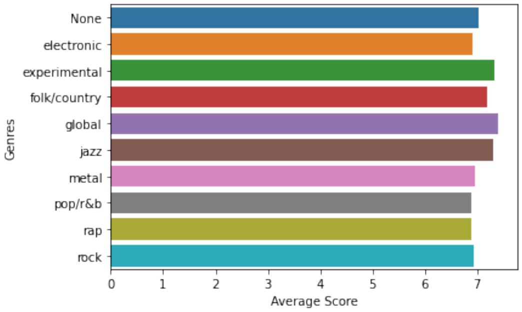

Figure 5. Average score per genre.

To have an overview about the length of a review, an analysis about the amount of words within one review was executed. In figure 6 the density of the amount of words in a review is plotted. This shows that most of the reviews consist of approximately 700 words. This is also concluded from the average word count and median, which are respectively equal to 703 and 649. To be able to make a representative further analysis, all the reviews who have a word count higher than 1500 are considered together.

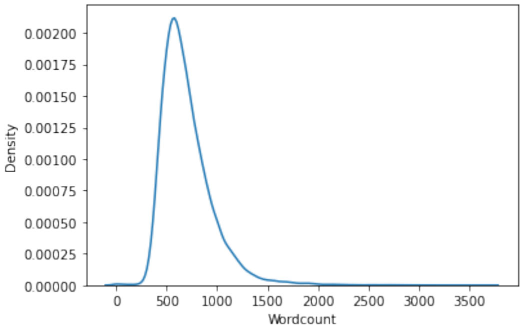

Figure 6. Density of the amount of words in a review.

As seen in figure 7 reviews for every genre have approximately the same amount of words. However, figure 8 shows an interesting result. It shows that the higher the word count of a review, the higher the average score is, with a difference of two points, on a scale of 10, between the shortest and longest reviews.

Besides this, text analysis of the reviews has been executed, where the most common words of reviews with a score above the average are found. Also for reviews with a score above the average the most common words are searched for. These findings can be further explored during the next milestone.

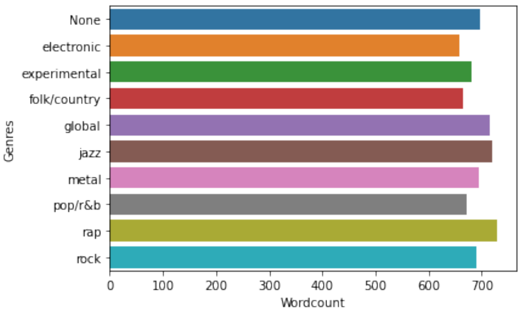

Figure 7. Average amount of words of a review per genre.

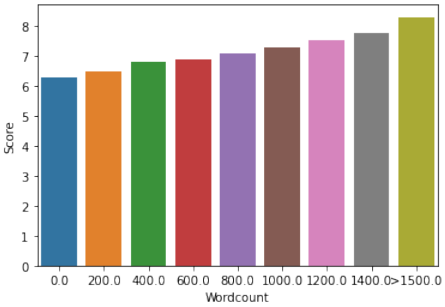

Figure 8. Average score of a review based on its word count.

7 Search tasks

After retrieving, cleaning and characterizing the data, insights about the properties of the data were acquired. With these insights, it was possible to devise useful queries for our search system.

- Search for a song by genre, artist, review-score, name.
  - Returns a list of songs that meet the presented requirements.
- Search for a review based on the authors name, the genre, the score or the content of the review.
  - Returns a list of reviews that meet the presented requirements.
- Check which genres are more reviewed by Pitchfork.
  - Returns the number of reviews per genre.
- Check which authors write good reviews per genre.
  - Returns a list of authors who have written more than 10 reviews for one genre of music, where all these reviews have a total average score of ≥ eight.

8 Conclusion

The first milestone was successfully completed, since a relevant dataset for the project has been selected, prepared and characterised. The quality of the data and authority of the source have been assessed and approved. The data has been analysed and characteristics have been composed. A data processing pipeline was developed and a conceptual model for the data has been established.

An encountered set back, was the switching of OpenRefine to python for the data cleaning in the middle of the milestone. This choice was made, so the cleaning process could be automated in the pipeline.

During the next milestone, we will choose an information retrieval tool and build useful indexes. With these an indexing and retrieval process will be generated.

References

[1] Idera. Data Growth. https://www.idera.com/glossary/data-growth/

[2] [n.d.]. Pandas. https://pandas.pydata.org/

[3] [n.d.]. Beatiful Soap. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[4] Nolan Conaway. 18,393 Pitchfork Reviews. https://www.kaggle.com/datasets/nolanbconaway/pitchfork-data

[5] [n.d.]. Matplotlib. https://matplotlib.org/

[6] [n.d.]. Seaborn. https://seaborn.pydata.org/

[7] [n.d.]. OpenRefine. https://openrefine.org/