PRI · Information Processing and Retrieval, 2021/22

ANTÓNIO BEZERRA, GONÇALO ALVES, and PEDRO SEIXAS

Fig. 1. Representation of Information Processing and Retrieval

As part of the curricular unit of "PRI", this report compiles the several milestones of our group project. As such, this document will explore the following topics: Data Preparation, Information Retrieval and Search System

Additional Key Words and Phrases: datasets, information processing, information retrieval, statistical analysis

ACM Reference Format:
António Bezerra, Gonçalo Alves, and Pedro Seixas. 2021. PRI · Information Processing and Retrieval, 2021/22. 1, 1 (November 2021), 3 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

As part of the course **PRI**, a project, to develop during the whole semester, was requested. The end product of this project is an information search system, that includes work on data collection and preparation, information querying and retrieval, and retrieval evaluation.

As per the topic of our group, we firstly thought of books, music and even Github commit messages, before chosing to work with a dataset related to food. This decision was based on two fundamental properties. Firstly, we wanted a dataset that had a large amount of data, as in a minimum of 10 columns and thousands of rows, so that we cpuld work with a large sample size. Secondly, we wanted a dataset that had both textual and numerical data that we could work on, to create a complete search system.

2 DATA PREPARATION

This first milestone consists of the preparation and characterisation of the datasets chosen. In our project, these datasets are *reviews.csv* and *recipes.csv*, obtained through Kaggle.

After a brief analysis of the datasets we concluded that both these datasets would result in a challenging but interesting project, not only by its size, but by the information it contained (ingredients, nutritional information, ...).

2.1 Data Pipeline

After this first analysis, a concept for a pipeline, represented in the next figure, was designed.

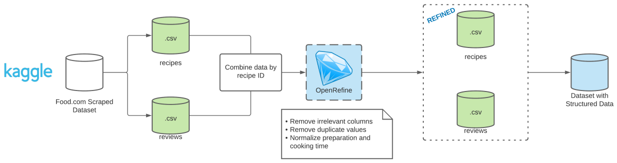
kaggle
Food.com Scraped Dataset
.csv
recipes
.csv
reviews
Combine data by recipe ID
OpenRefine
• Remove irrelevant columns
• Remove duplicate values
• Normalize preparation and cooking time
REFINED
.csv
recipes
.csv
reviews
Dataset with Structured Data

Fig. 2. Initial Pipeline

As we were developing the **Makefile**, required for this milestone, and a second, more detailed, analysis of the datasets was made, a new version of the pipeline was developed:

As we can see in this pipeline, we extracted the authors' names from both the *reviews.csv* and *recipes.csv* datasets, to a new dataset called *users.csv*. Besides being good practice, from a database standpoint, we wouldn't have to deal with redundancy from both datasets.

Next, we eliminated every duplicate record we could find in these datasets.

Authors' address: António Bezerra, up201806854@up.pt; Gonçalo Alves, up201806451@up.pt; Pedro Seixas, up201806227@up.pt.

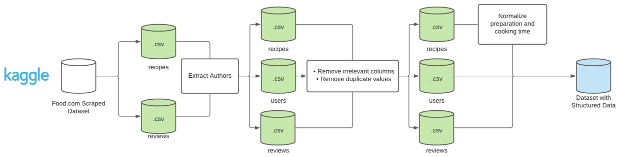

Fig. 3. Second Pipeline

And finally, we normalized the cooking, preparation and total time of the *recipes.csv* dataset, passing from **ISO 8601** to **seconds**.

On further inspection, we discovered that the **ingredients** and **ingredient quantities** columns of the *recipes.csv* dataset had rows that didn't match up, for example: a recipe with 2 ingredients had 5 ingrdient quantities. This is because only ingredients with a dedicated page within **Food.com** were scrapped. As such, we thought best to add a new column to our *recipes.csv* dataset, called "URL", that contains the hyperlink to the recipe page. We also only modified the date columns to only keep the day when a recipe/review was posted. In the case of a **review**, we removed the **DateSubmitted** column and kept the **DateModified** column.

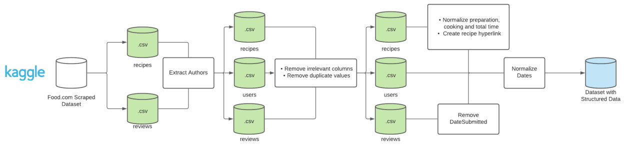

Fig. 4. Third Pipeline

Finally, we wanted to transform our datasets into an SQL databse, so that our future work would be simplified. We also: removed some outliers from our *recipes.csv* dataset, based on the number of calories and cook time; normalized the number of images per recipe and a recipe's instructions.

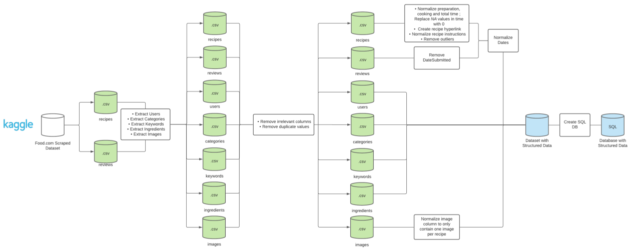

Fig. 5. Fourth Pipeline

2.2 Dataset Characterization

Our dataset can be mainly divided into two parts, the **recipes** and their **reviews**.

As for the **recipes**, they are identified by their unique ID and their name. In this dataset, there is also the author's ID, which can be used to retrieve the author's name in the **users** table. There are three columns that represent the time it takes to follow the recipe. The first one, **CookTime**, as the names suggests, represents the time it takes, after all the preparation, to get the recipe ready. The second one, **PrepTime** represents the time it takes to make everything ready to bake. The sum of this two times make up the **TotalTime**, which is the time it takes to make this recipe, preparation and baking included.

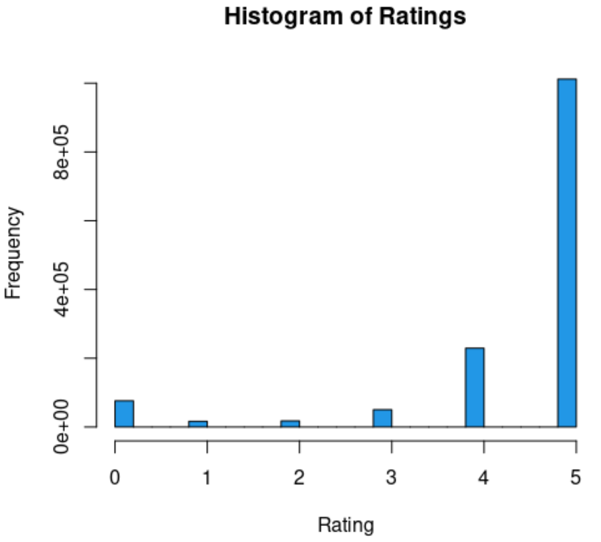

Fig. 6. Rating Histogram

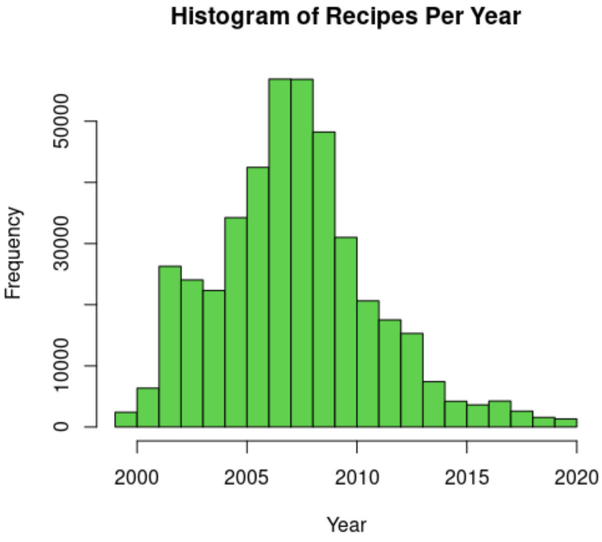

Fig. 7. Recipes Histogram

2.3 Data Domain Conceptual Model

After the data preparation phase, we ended up with the follwing relations:

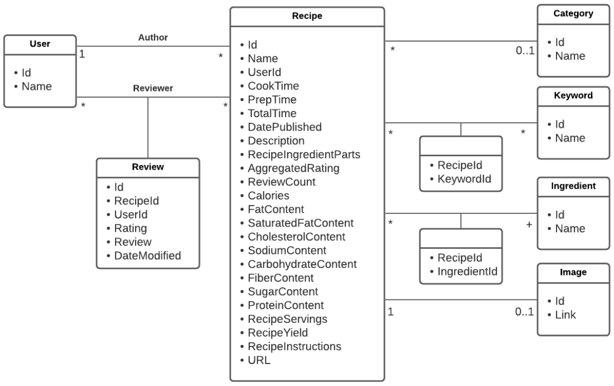

Fig. 8. Data Conceptual Model

search repositories for datasets; done select convenient data subsets; done assess the authority of the data source and data quality; perform exploratory data analysis; prepare and document a data processing pipeline; done characterize the datasets, identifying and describing some of their properties; identify the conceptual model for the data domain; done identify follow-up information needs in the data domain.

3 CITATIONS AND BIBLIOGRAPHIES

The following section serves to reference our citations and to list the materials used in this project.

Online citations: [1].

REFERENCES

[1] Sérgio Nunes. 2021. *PRI 2021/2022.* Retrieved November 14, 2021 from https://web.fe.up.pt/~ssn/wiki/teach/pri/202122/index

4 ACKNOWLEDGMENTS

ACKNOWLEDGMENTS

To Sara, for putting up with us.