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), 4 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.

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

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.

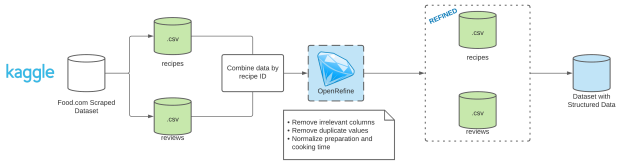

Fig. 2. Initial Pipeline

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.

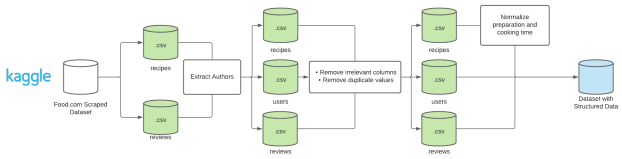

Fig. 3. Second Pipeline

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.

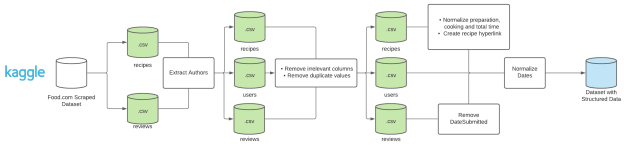

Fig. 4. Third Pipeline

Finally, we removed some outliers from our *recipes.csv* dataset, based on the number of calories and cook time and normalized the number of images per recipe and a recipe's instructions.

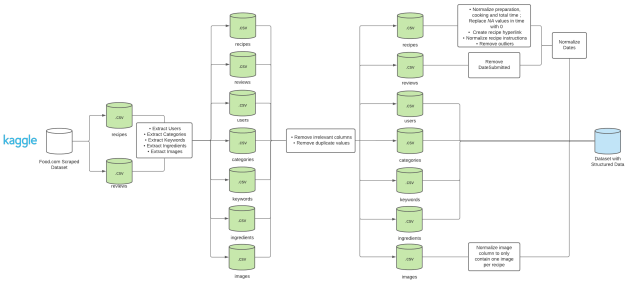

Fig. 5. Fourth Pipeline

2.2 Dataset Characterization

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

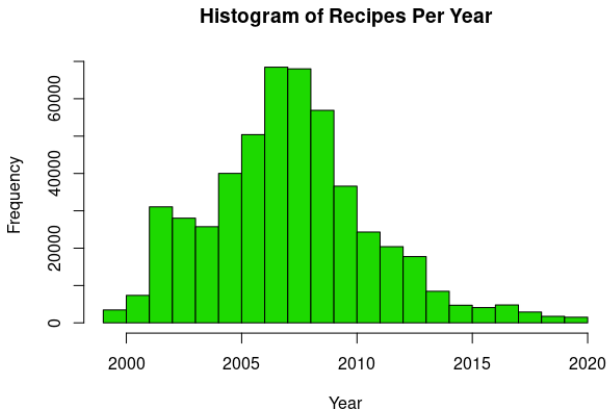

Fig. 6. Recipes Histogram

As for the **recipes**, our dataset has more than 500 thousand entries, published in more than 20 years. Looking at 6 we can see that the amount of recipes uploaded to **Food.com** peaked 14 years ago, in 2007, with more than 60 thousand recipes being uploaded in a single year. This was due to the website's popularity in that time. After that we can see a significant decrese in the number of recipes uploaded to the website, being currently on an all time low, according to this dataset. This doesn't mean that the website has no current active users, since there can be many recipes uploaded to the website that are not being scrapped by the creator of the dataset.

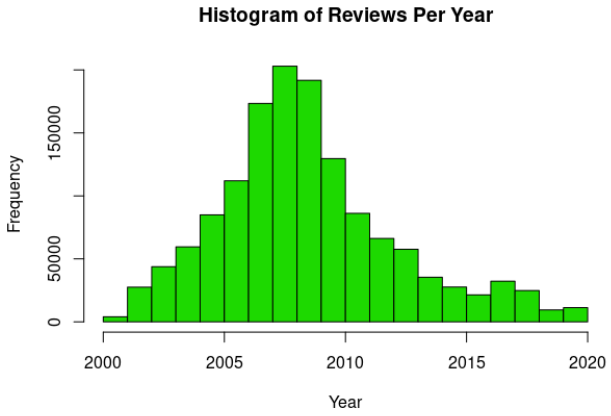

Fig. 7. Review Histogram

When talking about **reviews**, this dataset contains over 1.4 million reviews. The graph (7) looks very similar to the **recipes**' one, even though the frequency values are a lot higher which is expected since the reviews are scrapped from the recipe's page and the recipes can have tens or even thousands of reviews.

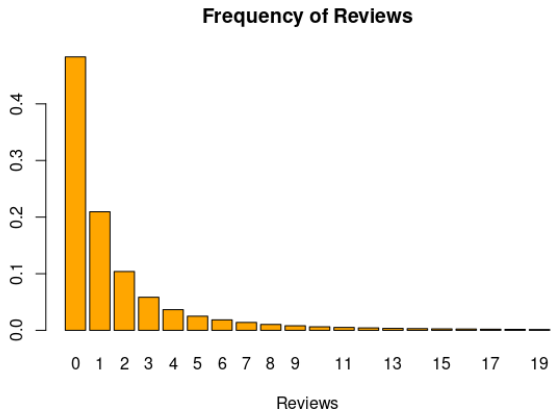

Fig. 8. Reviews per Recipe Histogram With 0

In fact, as we can see from the Figure 8, more than 50% of the recipes don't have any reviews. To get this graph, we needed to filter the recipes, as some had more than a thousand reviews, being undeniable outliers. Without those recipes, the resulting graph shows a clear exponential decrese in the percentage of recipes, as their number of reviews increase. To get a better understanding of how was the exponential pattern, the recipes that did not have any reviews were removed, temporarily, to create graph in Figure 9.

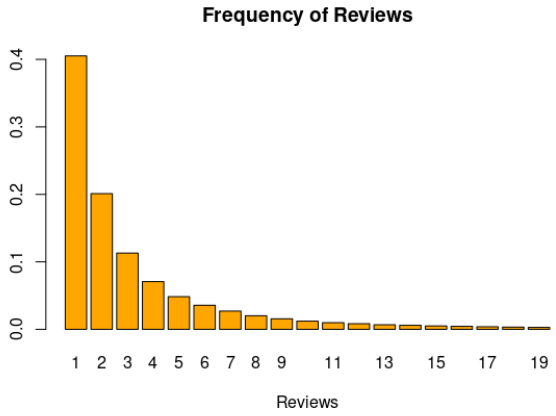

Fig. 9. Reviews per Recipe Histogram Without 0

Even without the recipes that did not have any review, we could clearly see the exponential decrease. This was also verified if we filtered out the recipes with 1, 2, 3, 4 and so on, which resulted in the expected way.

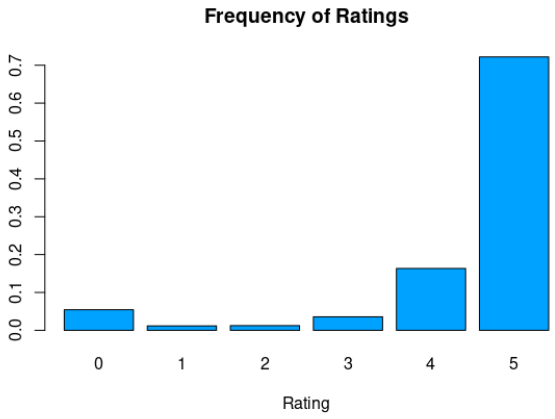

Fig. 10. Rating Histogram

After analysing the number of reviews, we thought it was interesting to find out what was the rating of this reviews to see if they are mostly positive or negative. With the graph in the Figure 10 we can clearly see that the number of positive reviews is significantly higher than the negative ones. One curious statistic is the fact that there are almost no reviews with 1, 2 or 3 stars, inidicating that the users tend to use the extremes when reviewing, either positively or negatively.

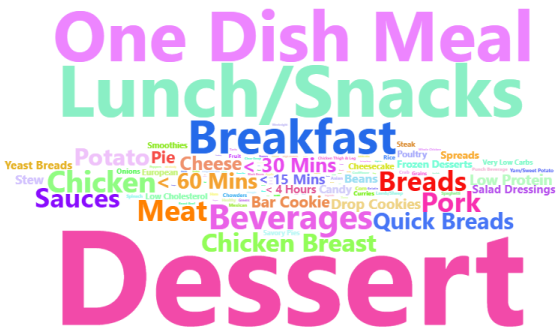

Fig. 11. Categories Word Cloud

After all the analysis in the reviews dataset, we found out that it would be interesting to know, since most of the reviews are positive, which categories of recipes were the most popular. For that, we made a word cloud (a great type of graph to measure the frequencies of certain words). In this word cloud, 11, we clearly have some categories that are more frequent than the others, such as "Dessert" or "Lunch/Snacks". This statistic has just confirmed the thoughts at the start of this project, that there would be a lot of recipes about cakes and other traditional dishes. There are over 300 categories in this dataset, so this word cloud can not demonstrate the full list of categories.

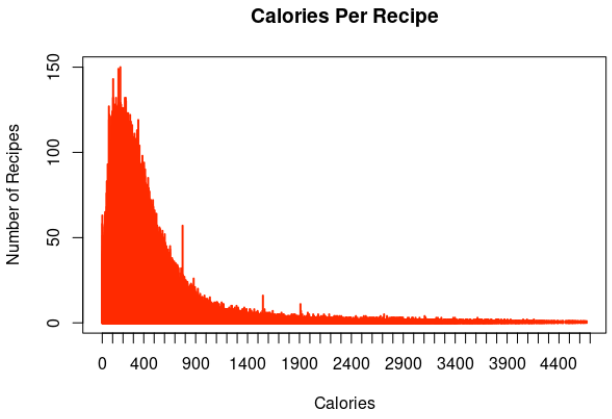

Fig. 12. Calories Per Recipe

With the categories in mind, another statistic that we could study would be the total calories of the recipes. As two of the main categories were "One Dish Meal" and "Lunch/Snacks" we already expected to have the highest frequency at around 300 calories, which is the normal calorie count for a complete meal and the graph in Figure 12 is demonstrating just that. The higher calorie recipes are mainly desserts or recipies with a high number of portions.

2.3 Data Domain Conceptual Model

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

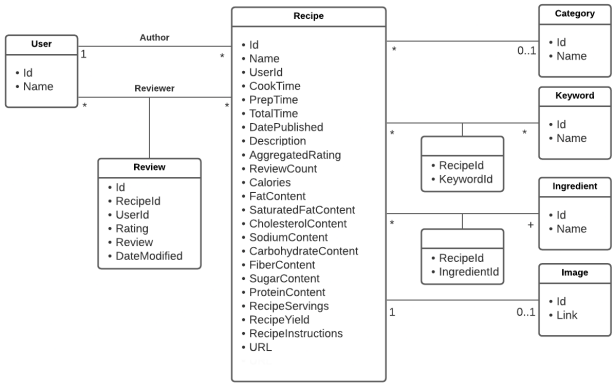

Fig. 13. Data Conceptual Model

Our main class is the **recipe** that has the following attributes:

- **Id** : Unique recipe ID;
- **Name** : Recipe name;
- **UserId** : Recipe's author ID;
- **CookTime** : Time it takes to bake the recipe, in seconds;
- **PrepTime** : Time it takes to prepare the recipe, in seconds;
- **TotalTime** : Time it takes to prepare and bake the recipe, in seconds;
- **DatePublished** : Date that the recipe was published in the website;
- **Description** : Description of the recipe;
- **AggregatedRating** : Review ratings mean;
- **ReviewCount** : Number of reviews;
- **Calories** : Number of calories;
- **FatContent** : Fat in the recipe, in grams;
- **SaturatedFatContent** : Saturated Fat in the recipe, in grams;
- **CholesterolContent** : Cholesterol in the recipe, in miligrams;
- **SodiumContent** : Sodium in the recipe, in miligrams;
- **CarbohydrateContent** : Carbohydrate in the recipe, in grams;
- **FiberContent** : Fiber in the recipe, in grams;
- **SugarContent** : Sugar in the recipe, in grams;
- **ProteinContent** : Protein in the recipe, in grams;
- **RecipeServings** : Number of servings in the recipe;
- **RecipeYield** : Recipe yield;
- **RecipeInstructions** : Instructions to follow the recipe;
- **URL** : Food.com url;

Each recipe has a author, which is an instance of the **user** class which has only its **Id** and its **Name**. Each **review** is an assotiation class between an author and a recipe and has the following attributes:

- **Id** : Unique review ID;
- **RecipeId** : Recipe's ID;
- **UserId** : User's author ID;
- **Rating** : Rating given in the review, between 0 and 5;
- **Review** : The actual review text;
- **DateModified** : Date that the review was made, or last modified;

Each recipe also has a **category** associated. It can also have an **image**, multiple **keywords** and multiple **ingredients** associated.

3 CITATIONS AND BIBLIOGRAPHIES

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

Online citations: [2], [1].

REFERENCES

[1] irkaal. 2020. *Food.com - Recipes and Reviews.* Retrieved November 14, 2021 from https://www.kaggle.com/irkaal/foodcom-recipes-and-reviews?select=reviews.csv

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