Food.com Recipes - Information Processing and Retrieval

Beatriz Mendes (up201806551@edu.fe.up.pt)
Henrique Pereira (up201806538@edu.fe.up.pt)

Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

November 17, 2021

Abstract

On a daily basis, we are confronted with different dishes that might spark our interest, due to their exquisite nature or taste, and sometimes we are motivated to recreate them on our own, but having no idea how to do so is not really helpful. Thankfully, nowadays, recipe books are common and easy to access. Nevertheless, looking for a specific kind of recipe with certain requirements in mind can be very time-consuming and exhausting. In order to make this process easier and quicker, we idealised the concept of a database in which recipes can be filtered by their details.

1 Introduction

Recipes are an originally verbal way of sharing knowledge regarding the preparation of a certain dish. More recently, these have been compiled into cooking books, where people can have access to multiple recipes in one go.

However, in the past few years, more and more recipes are being shifted into the digital world, where anyone can create their own recipes and publish them, which gives people an easier access and an even larger amount of free information.

Because of this, the amount of information present in databases such as Food.com ("Food.com" 2021) is quite overwhelming, and thus, we felt the necessity to create a new way of recipe searching.

Our tool will enable filtering the Food.com database by the details that most people use to identify certain recipes, such as:

- Ingredients used in the recipe;
- Time of preparation;
- Number of portions;
- etc.

2 Data Preparation

For this project, we found the dataset created by Shuyang Li in 2021 on Kaggle under the name "Food.com Recipes with Search Terms and Tags" (Li, 2021). Kaggle is an online community of data scientists and machine learning practitioners where users are allowed to publish and find datasets that may vary on the subject and quantity of data. (Wikipedia, 2021)

According to the information available, the original dataset consists of five hundred thousand recipes covering eighteen years of user-submitted uploads on said website (including when under the former name, GeniusKitchen), and was scrapped via Python Requests/BeautifulSoup.

2.1 Data Analysis

According to the dataset extracted, a recipe corresponds to an entry and each recipe is characterised by nine different parameters (each one corresponding to a column).

- The first column of the dataset corresponds to the identification of the recipe: its *id*. This parameter is useful since two recipes can have the same name but have different steps, ingredients or number of servings, so one easy way of distinguishing them is by using their ID.

- The next six columns are the more common parameters used to describe a recipe: its *name*, a *description*, the ingredients used with their quantities associated (*ingredients_raw_str*), the serving size (*serving_size*), how many *servings* are being made and the *steps* associated.
- Finally, the last two columns in the dataset retrieved are useful for filtering and searching purposes, respectively: the *tags* and the search terms (*search_terms*).

Keeping this in mind, we can attest that the dataset chosen is rich in information, since it contains a large number of recipes with a huge amount of unique descriptions, ingredients, servings, serving sizes, etc.

2.2 Data Processing Pipeline

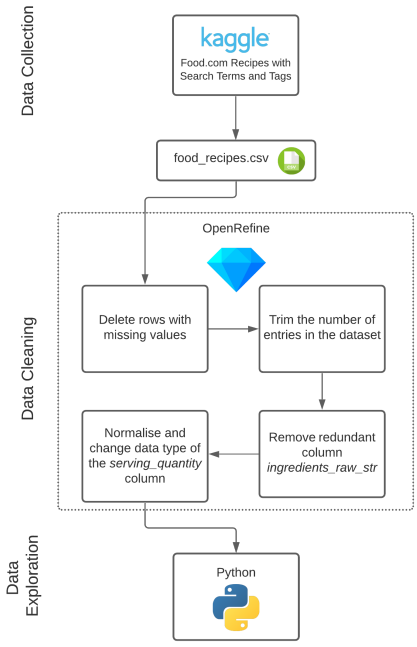

Figure 1: *Data Processing Pipeline*

This dataset was found in the form of a CSV file, so no further exploration was needed apart from the single dataset. Since five hundred thousand recipes can be quite heavy to process, we opted for trimming the dataset, as described on the data processing pipeline (Figure 1), using OpenRefine ("OpenRefine" 2021), a tool designed to deal with messy data and clean it. The cleaning process was performed in four different steps:

- First, we reduced the total amount of recipes to sixty thousand. We chose this value because we considered that, for this project, we did not need a number of entries as large as the one in the original dataset and we knew that the next steps of the data cleaning would remove some of the rows, so choosing a lower number of recipes would be imprudent;
- Then, we removed the rows with missing values (all of them were in the *description* column), to make sure that all rows we were going to use would be complete, creating, therefore, a dataset with a total number of entries of fifty eight thousand and seventy two;
- Afterwards, we removed the *ingredients_raw_str* column, due to it being redundant with the column *ingredients*, which already contains all the ingredients and their quantities;
- Lastly, we transformed the values in the column *serving_size* to only their weight, since all of them had the information that it was equivalent to one serving (redundant), and their measuring (all of them were weighted in grams), making the data cleaner.

After the data cleaning, the dataset was ready to be used and explored, using Python ("Python" 2021), with the help of the Pandas ("Pandas Documentation" 2021) library, the Matplotlib library ("Matplotlib" 2021) and the WordCloud library ("WordCloud" 2021).

2.3 Data Exploration

For better exploration of our dataset, we decided it could be interesting to perform several observations.

Firstly, we chose to explore the most common numbers of servings out of all the recipes in the dataset, keeping the range of servings from zero to sixty.

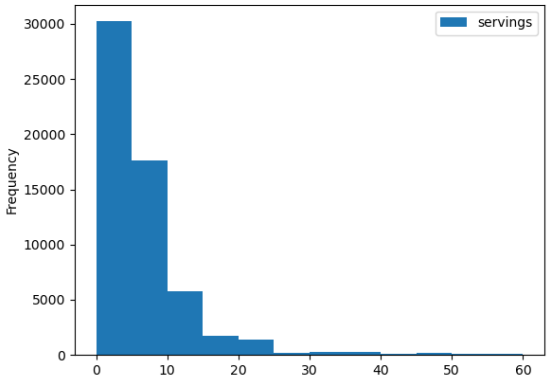

Figure 2: *Histogram with the frequency of the number of servings per recipe*

As shown in the Figure 2 histogram, the most common number of servings is between zero and five, with more than half of the total recipes resulting in a number of servings in this range. This information helps us understand that most of the recipes are made for a small amount of servings.

Secondly, we also decided to further explore the relation between the number of servings and the size of each one of those servings.

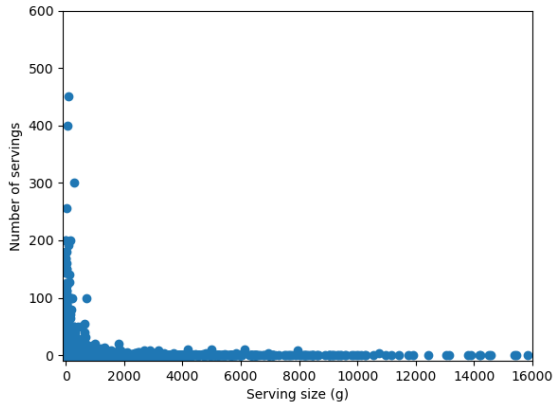

Figure 3: *Scatter plot with the relation between the number of servings and the size of each serving*

As shown in the Figure 3 scatter plot, there is an inversely proportional relation between these two variables, given the fact that the highest values for the serving size are registered for a lower number of servings, and vice-versa.

Lastly, we chose to determine what are the most used search terms and tags for all the recipes. For this, we decided to make a word cloud for each of those two parameters.

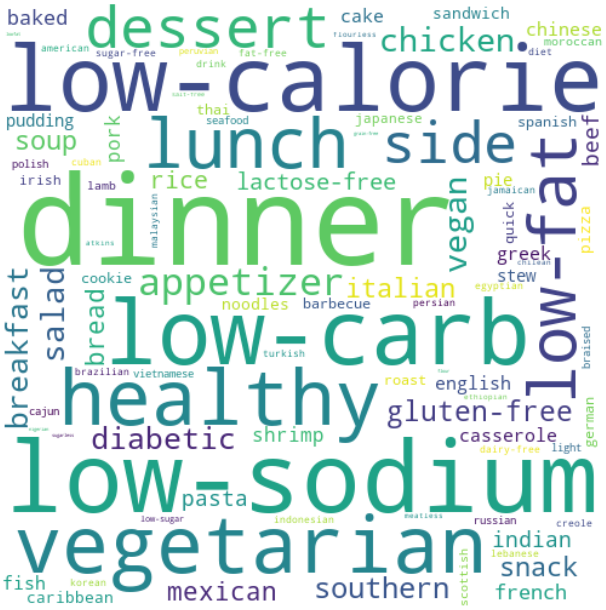

Figure 4: *Word Cloud of the search_terms*

In Figure 4 we can observe that the most common search terms are: low-carb, low-calorie, low-sodium, healthy, etc., which let us determine that the most common recipes and probably the recipes that will be more searched by the user are overall healthy, which will help us to have in mind the target audience of our project.

As for the tags, in Figure 5, we can see that the most used ones specify certain requirements of the

Figure 5: *Word Cloud of the tags*

recipes, for example the time to prepare, whether or not the recipe needs prior preparation or even specific ingredients present in the recipe.

After this exploration of our dataset, we will be able to work with the data having in mind what is more frequent and what will probably be more searched by the users, as well as a more broad perspective of the subset's potential.

2.4 Data Model

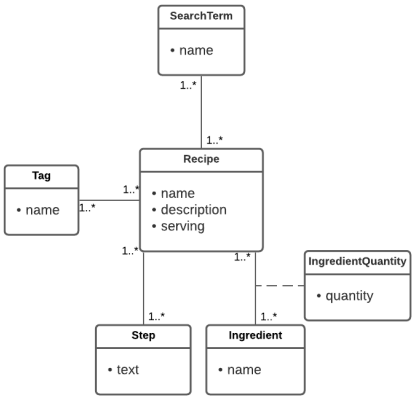

Figure 6: *Dataset's Data Model*

For this dataset we have put together a conceptual model (Figure 6), which describes the relations between every column in the dataset. First, we start by having a class *Recipe*, which contains the attributes *name*, *description* and *serving*, this last one indicates the number of servings of the final recipe.

Then, associated to this class, we have the class *Ingredient*, which contains the attribute *name*. Both classes

are associated with a multiplicity of one to many in both directions, meaning each recipe has at least one ingredient associated to it, and vice-versa. Furthermore, we have an association class called *IngredientQuantity*, which provides additional information about the association relationship between the class *Recipe* and the class *Ingredient*. The attribute of this association class is the attribute *quantity*, which indicates the quantity of one ingredient used in one specific recipe.

In addition, we have a class called *Step* associated with the class *Recipe*, containing the attribute *text*, that represent each individual step of a given recipe. Both classes are associated with a multiplicity of one to many.

Finally, associated to the class *Recipe* we also have the classes *Tag* and *SearchTerm*, which both contain the attribute name. The class *SearchTerm* represents, as the name suggests, the terms used when searching for a specific recipe, meaning, when a term is used for searching, all the recipes associated to that term will be shown. Similarly, the class *Tag* represents the words by which a recipe can be filtered. The relation between these two classes and the class *Recipe* has a multiplicity of one to many as well in both direction.

2.5 Follow-up Information Needs

The dataset is intended to be used by anyone who needs to find a recipe. The information needs that may lead to this can be, for example: to find recipes containing a specific ingredient, with a certain cooking time, for a certain meal/part of a meal (dinner, breakfast, dessert, etc.), following food restrictions (vegan, non-dairy, etc.), for a given number of people, kitchen appliances needed (oven, freezer, stove, etc.), among many others.

3 Conclusion

The goals of this first milestone were to search for and choose a dataset that suits the requirements (by being reliable, having a reasonable size, etc.), determine what tools we should use in order to process it and, finally, perform a data analysis on the subset obtained.

During the completion of these tasks, we comprehended that the dataset, obtained from reliable sources, was too large to be worked with and had redundant and missing data, which allowed us to experience the usage of data cleaning tools, such as OpenRefine, which we believed was the most efficient alternative.

Furthermore, we performed data exploration using Python in order to achieve a better understanding of our subset, which helped us notice a few changes needed to the subset, such as column data types which were different from the ones we needed for the exploration stage.

We believe that the fulfilment of the goals enumerated will help us achieve the goal for the subsequent milestones and make it easier to implement an information retrieval tool and a search system on top of the subset produced.

Bibliography

"Food.com" (2021). In: URL: https://www.food.com.

Li, Shuyang (2021). "Food.com Recipes with Search Terms and Tags". In: URL: https://www.kaggle.com/shuyangli94/foodcom-recipes-with-search-terms-and-tags.

"Matplotlib" (2021). In: URL: https://matplotlib.org.

"OpenRefine" (2021). In: URL: https://openrefine.org.

"Pandas Documentation" (2021). Version 1.3.4. In: URL: https://pandas.pydata.org/docs/.

"Python" (2021). In: URL: https://www.python.org.

Wikipedia (2021). "Kaggle". In: URL: https://en.wikipedia.org/wiki/Kaggle.

"WordCloud" (2021). In: URL: https://github.com/amueller/word_cloud.