Network-based Food Exploration

Team 10: Andrei Furtuna, Paul Gafton, Sorin Mircea, Alexandru Mocanu

December 2019

1 Introduction

Our way of eating drastically changed over the years: people tend to pay more and more attention to the way they eat (« healthy » way of eating, ethical and bio products, price, etc.), while having less time to spend on cooking (studies, working schedule). In recent years, the development of catering services and food-delivery services helped people dealing with the time issue, but those solutions are often costly and not really the healthiest. Online cooking recipes seem to be a good compromise. Indeed, one can search online for recipes corresponding to several criteria : the time they want to spend in the kitchen, the expected cost, the calories, etc. Thus, we wanted to focus in this project on the Recipes 1M dataset.

2 Datasets

The main dataset that we will use is the Recipes 1M dataset. This will be used for its data about recipes and their ingredients.
In order to find more data about the recipes (ex. the country/region of origin), we will also crawl Wikipedia and explore other datasets, such as Open Food Facts. We will not be able to retrieve such information for all the recipes, so we will then perform some semi-supervised learning to label the other recipes as well.

3 Goals

Depending on the time that we will have to explore the various aspects of the data, the things that we would like to cover are:

1. determine what ingredients fit with each other based on how they occur in the same recipes
2. deduce the country/region of origin for the recipes, based the labeling that we can get for some of the recipes
3. deduce the type of dish (starters, main course, dessert, drink)
4. deduce time to cook a dish and estimate cost, as this could be an information of interest, especially for students
5. estimate the number of calories for the recipes and determine if there is a connection to the region of the recipes
6. cluster the recipes based on dietary styles (raw veganism, paleo) and find similarities (in terms of calories, fats, most common products used) between these groups of recipes

4 Plan

The dataset may be too large to work on directly if we do not find a cluster to host it, so we may need to downsample it first.
The second step will be to determine the list of ingredients for each of the recipes. Currently each recipe has a list of ingredients which are however mostly given in the format QUANTITY_INGREDIENT. We will need to parse the ingredients out of these items.
Thirdly, we will extract additional information about several recipes, such as the country/region of origin and the number of calories.
Finally, we will build two graphs for our analysis: a graph where the nodes are ingredients and the edges are linking ingredients occurring in the same recipes, a graph where the nodes are recipes and the edges are linking recipes with common ingredients. The first type of graph would be useful for task 1. The second type of graph would serve for the other tasks.