NTDS Project Summary

Team 36
Maria Katergi, Davit Martirosyan, Carla Ohanesian, Iuliana Voinea

Dataset: https://www.kaggle.com/hugodarwood/epirecipes#full_format_recipes.json

Motivation

In an ideal world delicious and healthy would be synonyms, but this is not always the case in reality. However, in the last few years, many people started to be more conscious about what they are eating, and are trying to adopt a healthier lifestyle. Thus, it is interesting to explore whether healthy or unhealthy food tends to have higher ratings. Moreover, we are eager to identify which ingredients or nutritional aspects in a dish influence people's preferences and ratings.

Data description

The dataset found on Kaggle (in json format) provides information about more than 20,000 dishes: ingredients, recipe description, nutritional aspects (calories, proteins, sodium, etc.), average rating given by people and more. However, the information in the dataset is not complete for all the rows, hence we will need to handle the missing values.

Research questions

- What are the main nutritional factors affecting people's culinary preferences?
- Are certain ingredient combinations correlated with higher scores?
- Do similar recipes tend to have similar ratings?
- Can we predict the ratings for a dish based on similar dishes' ratings and other features?
- Do vegetarian/vegan dishes tend to have lower ratings?

Data exploitation

To find answers to our questions we are going to construct a graph, with nodes represented by the recipes (dishes) and connect them based on the common ingredients they share. To do so, we will first need to apply natural language processing techniques to extract the key words for each recipe from the ingredients column and then use Jaccard similarity index to define the similarity between recipes.

Subsequently, we will look at some of the features as graph signals. For instance, we will visualize the ratings, amount of fat or sodium on the graph and look at the variation and smoothness of the resulting signals.

We also want to use graph filters to construct features to train a logistic regression model that predicts ratings based on the information we have about the dishes (ingredients, calories, fat content…). In addition, we may try to build a simple recommender system for recipes based on k-nearest neighbors (KNN).