Andrew Hall - amh284

DATA301 - Project Proposal

What are the most common ingredients for food recipes that are reviewed highly and reviewed poorly?

Summary:

This project will analyse recipes from Food.com, find ingredients most commonly used on highly reviewed recipes, and ingredients in poorly reviewed recipes. This project will use the Food.com Recipes and Interactions dataset, which contains both recipe information, and user reviews.
The algorithm TF-IDF will be used to deprioritise ingredients like salt and flour, which most recipes contain.
The intended result is to find if there is correlation between certain ingredients and how much people enjoy a recipe, and if so, what ingredients make recipes better or worse. This information could be used to help home cooks create nicer meals.

Motivation:

This question was chosen out of personal curiosity. I am someone who enjoys cooking & baking, and would like to know what ingredients are considered “good” and “bad” in public reviews, if such correlation even exists, possibly finding interesting ingredients to help people to create better meals.

Background:

The data contains two datasets: One containing information on the recipe, the relevant components being the recipe ID and the list of ingredients. Another contains a list of reviews, the relevant components being the score (integer from [0, 5]) and the referenced recipe ID.
The TF-IDF algorithm is used to see if data appears more in a certain category than any other. It does this by checking how frequently it occurs in a certain category compared to the rest of the data.

Research Question or Hypothesis:

The question is “What are the most common ingredients for food recipes that are reviewed highly and reviewed poorly?”. The Food.com Recipes and Interactions dataset contains information about over 180k+ recipes, containing information including their ingredient lists, and 700k+ reviews for said recipes, giving a numerical 0-5 score for each recipe.
The TF-IDF algorithm is used to check how frequently items occur in specific categories compared to everything else. In our case, we will use this to see what ingredients occur most often in high/low scoring recipes, while reducing the importance of ingredients which naturally occur frequently (e.g. cooking oil, salt, flour, water, etc.)

Design and Methods:

Reviews have an integer rating of either [0, 1, 2, 3, 4, 5]. A recipe's 'review score' will be its average rating. Recipes without too few reviews will be ignored.
The dataset may contain different names for the same ingredient, so we will have to give them the same name. (see below)
Each recipe will be categorised as either low, average, or high scoring, depending on its review score.
Afterwards, a TF-IDF will be run against the low and high scoring recipes, where 'term frequency' occurs within review score categories, and 'document frequency' is its occurrence across all recipes.

A foreseeable difficulty is the different namings of food. There could be different string names for the same ingredient (e.g. 'chicken', could mean 'chicken breast', 'chicken thigh', etc), and it will be difficult to find a way to categorise them all as the same. However, some similar things should not be lumped together, for example, oils; 'canola oil' can be defined as a 'cooking oil', 'sesame oil' and 'chilli oil' are a 'flavouring oil', but 'olive oil' is used for both.

References:

Generating Personalized Recipes from Historical User Preferences
Bodhisattwa Prasad Majumder*, Shuyang Li*, Jianmo Ni, Julian McAuley
EMNLP, 2019
Link to pdf