Food2Fork Clustering Project Proposal

Blue Team 3: Andrea Ruiz Rabadan (Team Lead), Catherine Barnes, Will Burton, Michael Byman, DT Oliver

Project Description

We will collect and analyze recipes collected from Food2Fork, a website that aggregates more than 200,000 recipes from many other food-related sites onto its platform. Specifically, we will look at ingredients, ingredient combinations, and possibly even ingredient ratios combined with recipe rankings (proprietary Food2Fork ratings based on social media attention) and apply text clustering and classification to provide a summary of the most frequent and popular ingredient combinations.

Data Source

We will use the Food2Fork API key to collect ingredients and rankings from recipes on the Food2Fork website. According to the website, the Food2Fork API allows 500 calls per day, with 30 results per call. Using 5 API keys (one from each team member), we should be able to collect all their recipes within 3 days.

Analysis

We will perform text clustering on the recipe ingredients to identify frequent ingredient combinations and apply text classification in an attempt to predict the most popular ingredient or ingredient combinations.

Challenges

The following are some potential problems we could encounter:

1. The ranks may not be a perfect representation of popularity. First, it is somewhat of a “black box” calculation, since Food2Fork doesn’t explicitly give the formula. Second, since the ranking is a function of social media attention, the rankings can also be a function of release date and not just pure popularity.
2. Some individual ingredients are more than one word in length (e.g. - green pepper). This will hopefully be resolved through the structure of the data - most ingredients (and the amount used) in a recipe are on their own line.

Data Source Justification

The Food2Fork website has a very large, and wide variety of recipes (200,000+). It has a consistent structure for listing ingredients and the recipe ranking on each recipe page. It also has a free and properly functioning API.

Deliverables

We will provide the following deliverables at the end of the project:

1. A dataset containing recipes/ingredients and the associated rank.
2. A summary of derived clusters, with appropriate labels.
3. A set of ingredients or ingredient combinations that are the most popular.
4. A short in-class presentation of our findings, discussions of their meaning, and general "lessons learned" from our project.