DS5230 - Project Abstract
Applying Recommendation System to Food Recipes
07.21.2020

Akshit Jain
Naga Santhosh Kartheek Karnati
Samar Dikshit

Normally, people have three meals a day. For each meal, there are two important questions to answer, "*what do we make?*" and "*what ingredients do we need to make it?*". Although food is essential to our life, answering the same questions repeatedly can be tiring, and it would be helpful to have a system that can suggest recipes and ingredients on the fly.

In order to build a recommendation system, we obtained data from the web platform Yummly. This dataset contains over 4,800 unique recipes to prepare 101 different dishes. On average, each dish has 50 alternative recipes. The unique recipes contain more than 3,200 ingredients (10 per recipe on average), providing multiple ways of preparing a single dish. Hence, it captures at the same time the intra and inter-class variability of cooking recipes.

To demonstrate our understanding of the dataset, we will perform initial exploratory data analysis. As a part of this, we will do the following: (i) clean raw ingredients in the recipes to obtain a simplified list of ingredients (eg. *yellow onion, flour* is changed to *onion, flour*), (ii) find the most common ingredients across all recipes, (iii) find the most unique ingredients across all recipes, and (iv) perform dimensionality reduction using techniques such as Principal Component Analysis (PCA) and Stochastic Neighbour Embedding (SNE) on document-term matrices with unique recipes as documents (rows) and unique ingredients as terms (columns).

The main objective of our project is to build a recommendation system that provides similar recipes and other ingredients required for those recipes. We plan on a two-pronged approach: (i) the user lets us know what ingredients they have, and we recommend what recipes they can make with those ingredients, and what other ingredients they'd need for that recipe (ii) the user lets us know what dish they want to prepare, and our system recommends alternative recipes for that dish, and similar recipes for other dishes if they exist. The first approach is based on ingredients' similarity (using term vectors), and we expect it to be a low-risk solution. The second approach would use various string matching techniques, along with clustering and similarity scores, and is a more high-risk solution to the problem.

We plan on creating multiple clustering models using k-means clustering and hierarchical clustering and verify the clusters with the PCA and t-SNE solutions. We will calculate similarity scores using metrics such as cosine similarity and Jaccard similarity, and output the top-n recommendations.