Team Hubble

Cuisine Type Classification – Executive Summary

Our objective for this coding project was to develop an algorithm that is capable of taking a list of user inputted recipe ingredients and classifying it into one of twenty cuisine types. The dataset consisted of 39,744 unique recipes across all 20 cuisine types, with 6,714 unique ingredients.

Data cleaning was performed first to eliminate punctuation, digits, brackets, and brand names that could otherwise interfere with accurate results. Next, *Term Frequency – Inverse Document Frequency (TF-IDF) Vectorization* was performed, which is a fancy way of saying each ingredient was assigned a weighted value for each recipe. These weights were used to infer the importance of each ingredient relative to the twenty different cuisine types. Finally, we attempted three different methods for predicting cuisine types, with the goal being maximum accuracy of classification. Maximum accuracy was achieved with the Support Vector Machine (SVM) method, which classified recipes to their correct cuisine type 80.87% of the time. At a high level, the SVM method works by generating a hyperplane that divides the data points we wish to classify. As a simple example, if we only had two types of data to classify then the hyperplane would be generated in such a way that it separates the data set into those two distinct groups, perhaps by a simple line through the data set that allows for as much buffer space between itself and the nearest data points from each group on either side of it as possible.

This solution will allow for easy and efficient classification of all recipes submitted to the website, which will allow users to filter and search for the various cuisine types they are interested in. A user-friendly experience that enables users to quickly identify what they are looking for is vital to ensuring they return to use the website again and also recommend it to others within their social circles, further driving increased traffic to the website. We are confident that our methods will produce positive results in this regard, and we appreciate the opportunity to demonstrate that.