**No Free Lunch**

**A Palatable Project Proposal by**

**Leon Chen, Daniel Lai, Andrew Reuben, Bejan Sadeghian, Samruddhi Somani**

**Problem:**

Provided a list of ingredients, can we determine what cuisine a particular recipe represents? Knowing which ingredients are predictive of which cuisines plus knowing which cuisines users prefer will allow companies like Yummly to recommend new cuisines to users as well as recipes within their preferred cuisines.

**Preface:**

*Yummly* is a website that aggregates recipes published to blogs around the web. The website allows users to filter recipes by several categories, including ingredients and cuisine types. It also asks users for their cuisine preferences and dietary restrictions as well as their skill level and household size. These characteristics are used to curate recipes for particular users, and the recommendations are refined as users engage more with the website.

**Data:**

*Yummly* has released a training dataset on Kaggle that includes a recipe identification number, cuisine type (our target variable) and list of ingredients. The training data contains approximately 40,000 recipes, averaging 11 ingredients each. We will use this data to create a document term matrix with ingredients as terms and individual recipes as documents.

**Approach:**

*Data Preprocessing:*

Like many data science situations, a lot of our work will be dedicated to preprocessing the data to achieve the highest prediction accuracy. We anticipate that many ingredients will have multiple names (for example, diced tomatoes and cubed tomatoes as well as cinnamon and cinnamon sticks). Some of these will be explanatory--powdered cinnamon is probably used in different cuisines than cinnamon sticks. Others, like the tomato distinction, are probably noise and will be merged. In addition, we expect to have some ingredients that are used in many recipes; we plan to use an IDF transformation to minimize the impact of common words in our model, allowing us to reduce noise.

To address our question, we will need a method that will allow us to classify new observations based on a training dataset and one that will be able to handle a large number of features in a mostly sparse matrix. We will be testing the following supervised learning methods: K-Nearest Neighbors, Multinomial Naive Bayes, Logistic Regression, and Ensemble Methods.

*K Nearest Neighbors:*

Because we are looking to classify test data we can use K-NN classification to determine the test observation’s most similar neighbors in the training set. Since we are working with highly sparse data we will utilize cosine similarity for our nearest neighbors comparison instead of euclidean distance. Because our dataset is not extremely large KNN is feasible, memory-wise.

*Multinomial Naive Bayes:*

Because we will be using an IDF transformation of our features, we will also utilize Multinomial Naive Bayes. From what we’ve learned about text mining, it is normally safe to assume feature independence when using a bag of words (ingredients in this case) model. This model will calculate the probabilities that a set of ingredients are in a specific class (type of cuisine) and use that in conjunction with the class prior to determine which cuisine class is most probable.

*Logistic Regression:*

We will use linear discriminant analysis to determine the vectors that best distinguish between classes and use these as features. This approach will allow us to minimize variance and consequently test error while minimizing information loss. We plan to use these features in a logistic regression and compare results with the Naive Bayes classifier to determine whether the independence assumption holds.

*Ensemble Methods:*

Due to the nature of our project--categorizing observations with a large set of features--we will test the ensemble methods Random Forest and Boosted Trees. Random Forest will allow us to use the full feature set for prediction.

**Conclusion:**

Looking at the leaderboards for the competition on Kaggle, most models have an accuracy above 70%. We hope to meet this benchmark, and through further refinement of the model and the data achieve an accuracy within the top 10%.

**Citations:**

Yummly & Kaggle. (n.d.). What's Cooking? Retrieved September 30, 2015.

Yummly. (n.d.). How It Works: Introducing the Ultimate Cooking Tool. Retrieved September 30, 2015.