Milestone Report (DS 5230) - Classifying Food Items into Appropriate Categories
Pranshu Diwan (001347331) and Shagun Jindal (001596493)

Introduction

The aim of our project is to classify a set of food items into their appropriate categories using various unsupervised Machine Learning Algorithms. The dataset contains a list of more than 13000 food items and their item id's. This project will help restaurants gain insight into consumers and their choices, which will help curate personalized offers based on needs helping businesses thrive and gain profits.

Ideas considered in this phase of work, with expected success and failure of each idea:

1. **Defining categories to segment food items:**
   After careful evaluation of the food items in our dataset we consider categorizing on different types:
   - Based on food categories like veg/non-veg, alcoholic/non-alcoholic beverages, and desserts. This is a baseline bracket of categories we are considering.
   - Based on cuisine - The data is from many Indian restaurants so there is a bias towards Indian cuisine, hence it is a failure and we don't consider it.
   - Combination of Cuisine and food categories - Some food items in a category can lie in multiple cuisines which will be a repetition and hence we don't consider it.

2. **Text cleaning and vectorization:**
   We have successfully cleaned the dataset by dropping unnecessary columns, converting data to lowercase, and removing digits, punctuation, or specific words that have no significance. Along with that we have vectorized food items using Word2Vec and plotted them using PCA. We found distinct clusters depicting desserts and spicy foods (displayed in visualization #2&3)

3. **Implementation of K-means:**
   This algorithm is an excellent failure. We began by creating a Word2Vec model. Once the food items were vectorized, we fitted the K-means model on the vectors. The k-means model suggested that we should have more than 5 categories, based on the elbow method (graph shown in visualization #4). However, creating more categories leads to results being very specific so we don't believe it will obtain better results. Also, some items are placed in incorrect categories. For example, Milk chocolate tub is being classified as a non-alcoholic beverage.

In general, classifying these food items correctly based on the number of clusters we want will be a difficult task, since we can just as easily change our approach to define categories. For example, if we consider solids and liquids as categories, we expect 2 clusters but K-means won't accurately classify them as solids/liquids. To conclude, our thinking of the categories and working of the K-means algorithm won't always match so we cannot depend on it.

New ideas not mentioned in the abstract:

1. **Trying a supervised ML approach based on clustering results**
   We wish to manually take a small part of the dataset, label it based on clustering results and try classification algorithms on the same dataset. We can manually label some food items

2. **Trying a cosine similarity approach:**
   We could define words that strongly represent each category like for dessert we could include chocolate, tiramisu. We could vectorize these "representative" words and find the cosine similarity of each word vector with these representations. Every food item will thus get categorized based on the highest similarity with the representative words.

The proposed scope of work for the final report:

1. **Trying different clustering algorithms**
   We wish to try different clustering algorithms like Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Expectation-Maximization (EM) Clustering using Gaussian Mixture Models (GMMs), and Agglomerative Hierarchical Clustering to see if we can get notable improvements in the results over K-means algorithms.

2. **Use the cosine similarity approach to obtain tangible results:**
   We have a few keywords in mind for each category. We wish to try the above-mentioned approach and analyze how well this approach works. We will analyze the accuracy based on the dataset created for our supervised ML approach

3. **Optimize the models by performing Hyperparameter tuning:**
   Perform techniques such as Grid Search and Randomized Search to find and use the optimized hyperparameter for the models.

Visualizations: (Graphs are a little small, please zoom in to see clearly!)

1. Wordcloud of all food items

Figure 1

2. Vectorized food items using Word2Vec plotted using PCA - observe the clusters

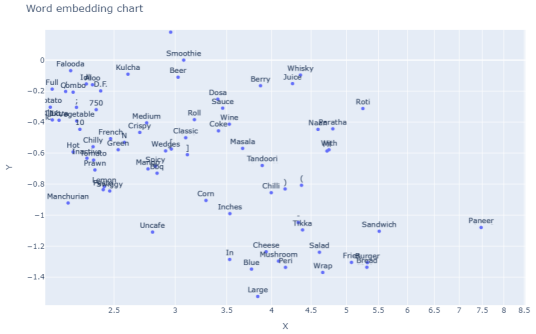

Figure 2

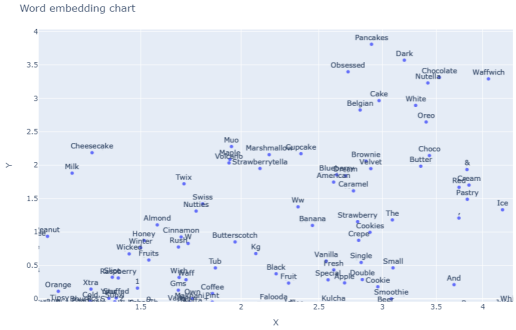

Figure 3

3. K-means elbow method graph - graph suggests we should have more clusters

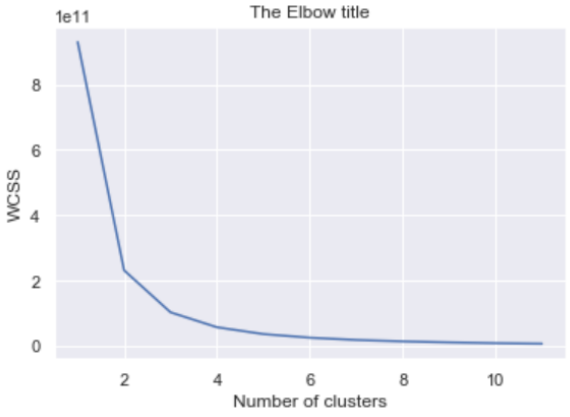

Figure 4