Data Mining Capstone Task 2

February 13, 2020

Asad Bin Imtiaz CS598
Data Mining Capstone
MCS-DS UIUC
Email: aimtiaz2@illinois.edu

1. Introduction:

In this task, cuisines from Restaurants in the Yelp Review Academic Challenge Dataset are extracted and their respective views were analyzed for similarity. The analysis and findings are part of Task-2 for Data Mining Capstone course, wherein the Yelp restaurant reviews were explored to uncover interesting patterns and knowledge. The goal of this task was to obtain an enriched representation of different cuisines and understand their closeness and similarity.

In the following, the implementation of the sub tasks for this task are explained

1.2 Tools and Libraries:

The exploration and analysis of Yelp restaurant dataset was performed with Python 3 using Jupyter notebook. Several python libraries such as **Gensim**, **scipy, numpy** and **Scikit-learn** were used to analyze the review texts and perform Natural Language Processing (NLP). For visualization, **D3.js** and python Matplotlib were used.

1.3 Data preparation:

The first step was to gather all the reviews for all the qualifying categories. To accomplish this, all reviews and business of category 'Restaurants' from the Yelp Dataset were read in Pandas data-frames. For each qualifying restaurant, all categories (except 'Restaurant' category) were extracted and mapping for all these categories to all reviews associated with these categories was constructed. The result of this step were a list of all qualifying categories and a dictionary of each of these category & list of its respective review text.

For simplicity and efficiency of processing, only 50 categories were selected for further analysis. These categories were a mix of regional cuisines (such as Italian, Thai etc.), cuisines types (such as Pizza, Bar, Buffet etc.) and cuisine styles (such as Vegan, Kosher, Halal etc.).

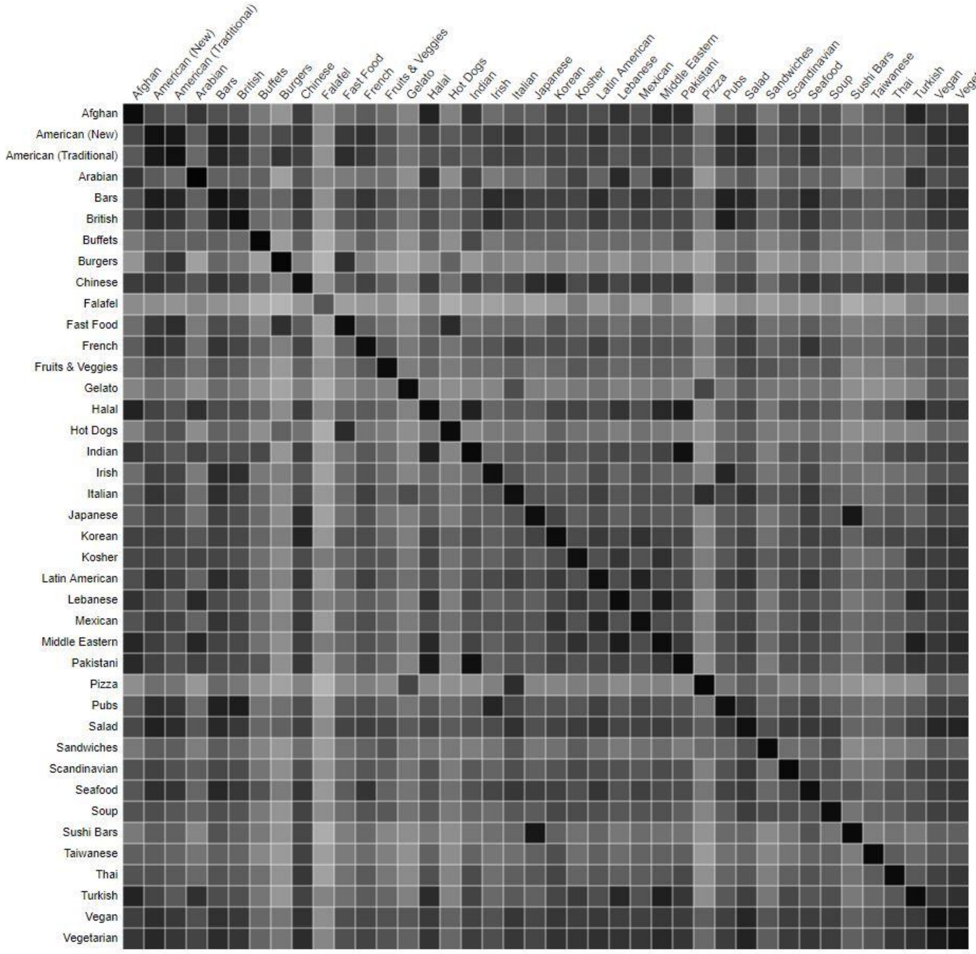
To further do the processing efficiently, a number of 5000 reviews for each category were retained randomly using stratified sampling within all restaurants in each category. This resulted in equal number of reviews getting selected from all restaurants for each category, with a total number of reviews in that category being 5000.

The Idea was to gain an understanding the similarity between regional cuisines among one another and to specific cuisine styles and types.

2. Task 2.1: Visualization of the Cuisine Map

In this sub-task, the raw text of the reviews for the selected 50 cuisines (as described in chapter 1) were processed for similarity analysis. To gain performance, all qualifying reviews per category were concatenated in groups of 200. A maximum of 10,000 features were used for similarity vectors computation. The **Figure 1** shows the similarity matrix of the result.

The opacity of each cell represents the similarity of the corresponding cuisines; the darker the cell, the similar they are. The visualization does show some similarity between cuisines; however, it is hard to tell which cuisines are similar to what category.



Max Features	10000	Max df	1.0
Norm	L2	Min df	1
IDF used	FALSE	# Clusters	None

Figure 1 Similarity Matrix Raw text

To visualize better, a force directed cluster graph was generated as shown in **Figure 2**. Here each node represents one of the 50 selected categories and the distance between the nodes depicts the similarity, the further the nodes, the dissimilar they are.

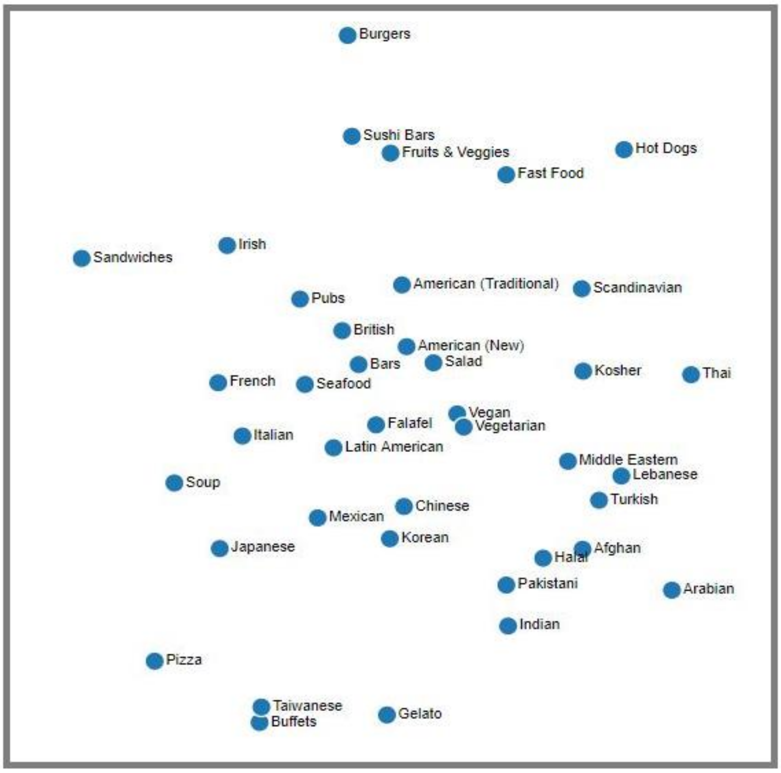


Figure 2 Cluster Graph for raw review text

As one can see, this similarity comparison was not very useful and needed to be optimized. To improve similarity analysis, instead of raw text, a processed review text was taken where all the stop-words were removed, the individual words were lemmatized and only part-of-speech of type Proper Nouns, Nouns, Verbs, Adverbs, and Adjectives were retained. **Figure 3** shows the result in a similarity matrix.

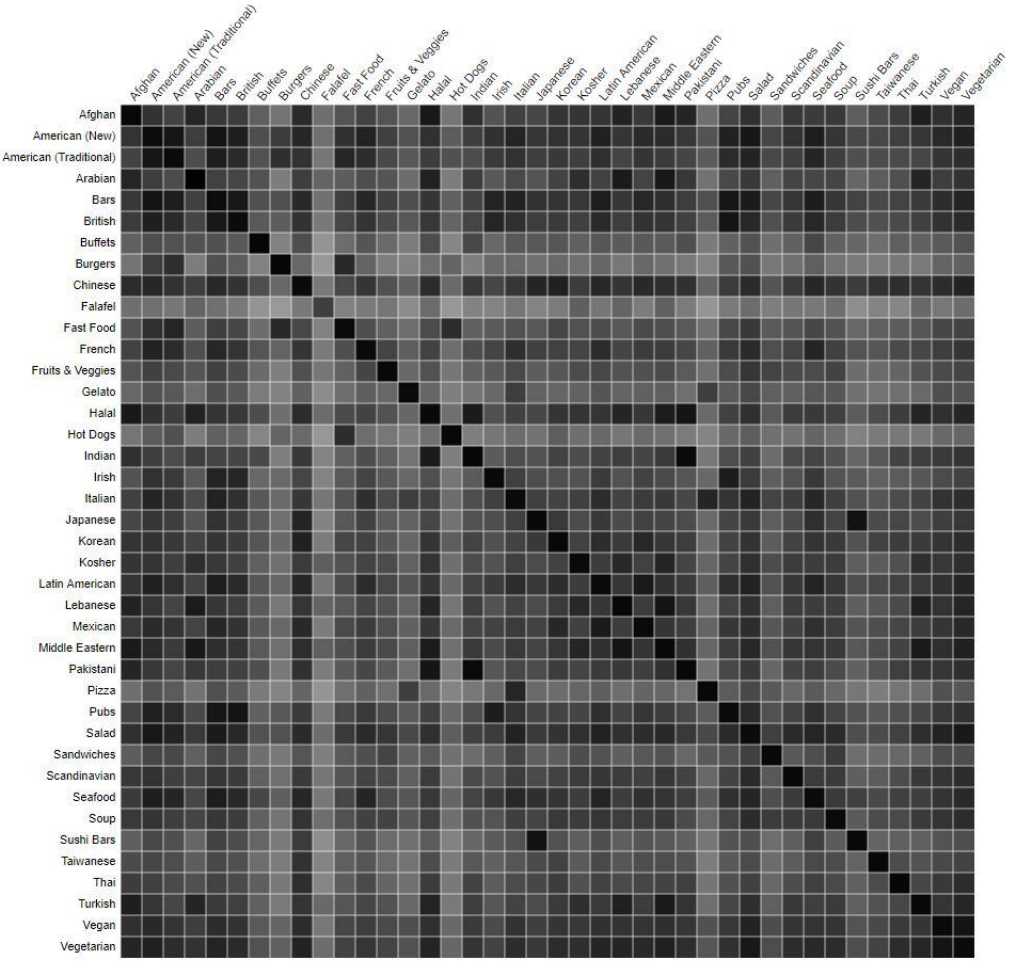


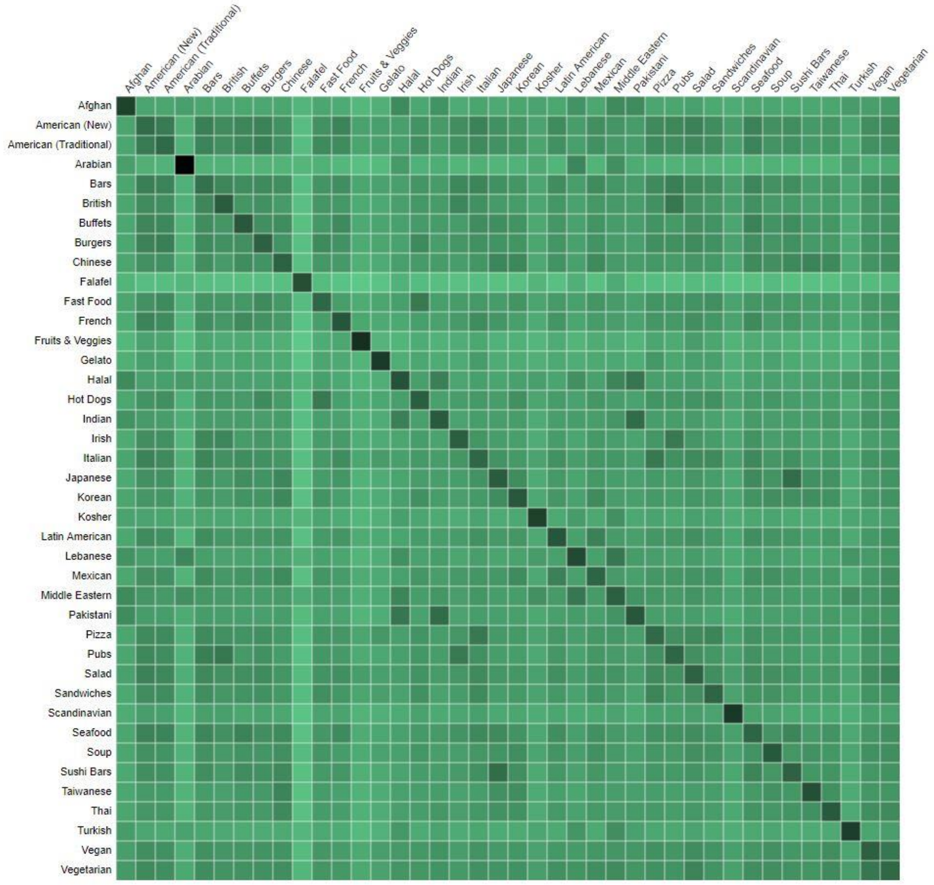
Figure 3 Similarity Matrix Processed Text

In the following sub tasks, improvements in similarity analysis was done.

3. Task 2.2: Improving Clustering Results

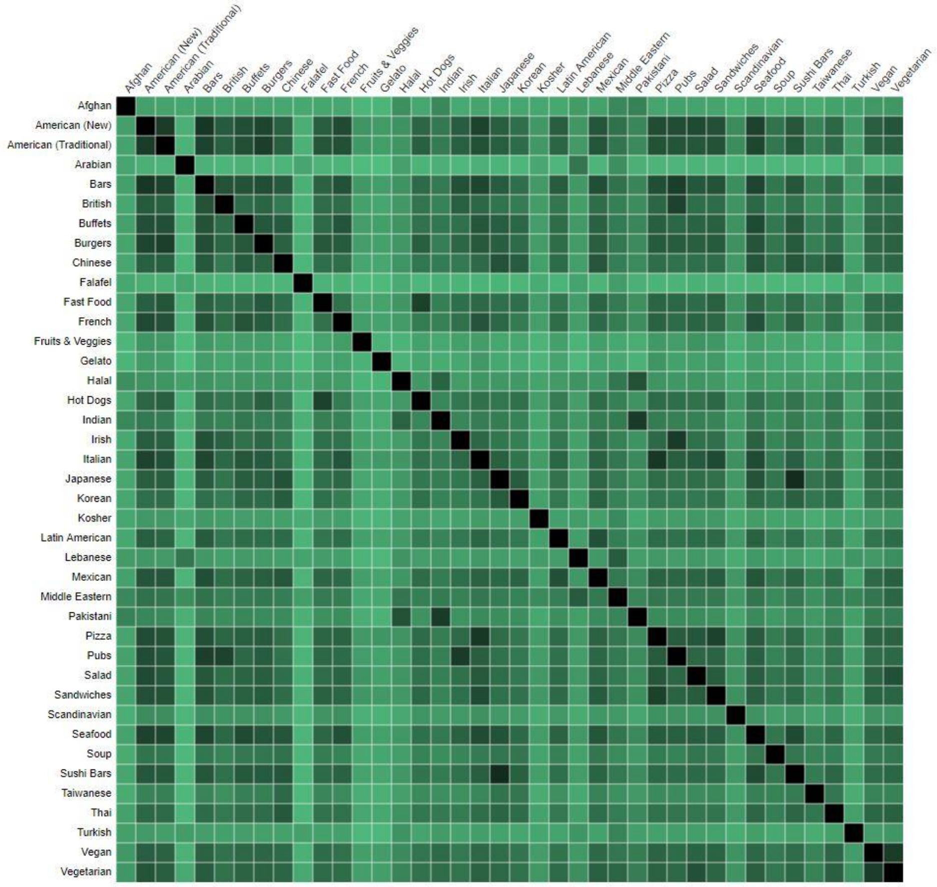
To improve the similarity computation, a TFIDF vectorizer was instantiated to give vectors with Inverse Document frequency (IDF). Enabling IDF removed high frequency words present in all the reviews in all categories. Thus, the similarity computation was done on distinctive features and common words in all the category-reviews were penalized. All reviews in a category were concatenated any analyzed with a setting of max features of 10,000, using sublinear term frequency & IDF, and with min DF of 80% & max DF of 1 document , a new similarity matrix was generated as shown in the **Figure 4**.

Here one can distinguish some cuisine groupings appearing more similar (darker cells) to one another, such as 'American Traditional' coming up a lot similar to 'American New', and 'Arabian' coming up somewhat similar to 'Falafel', 'Lebanese' and 'Turkish', giving some sort of indication of resemblance or correlation between these cuisines, which also seem logical and based on common sense. However, this chart (**Figure 4**) is also far from usable and further optimization was needed. To further improve, reviews were grouped as documents in group of 50 and **Figure 5** was generated, keeping the same settings as above. The **Figure 5** showed some improvements over **Figure 4**, and the cuisine clusters were a little more obvious.



Max Features	10000	Max df	0.8
Norm	L2	# Topics	3
IDF used	TRUE	# Clusters	None

Figure 4 Similarity Matrix, using IDF & Merged Reviews



Max Features	10000	Max df	0.8
Norm	L2	# Topics	3
IDF used	TRUE	# Clusters	None

Figure 5 Similarity Matrix with TDIDF & IDF = 0.8, groups of 50 reviews

4. Task 2.3: Incorporating Clustering in Cuisine Map

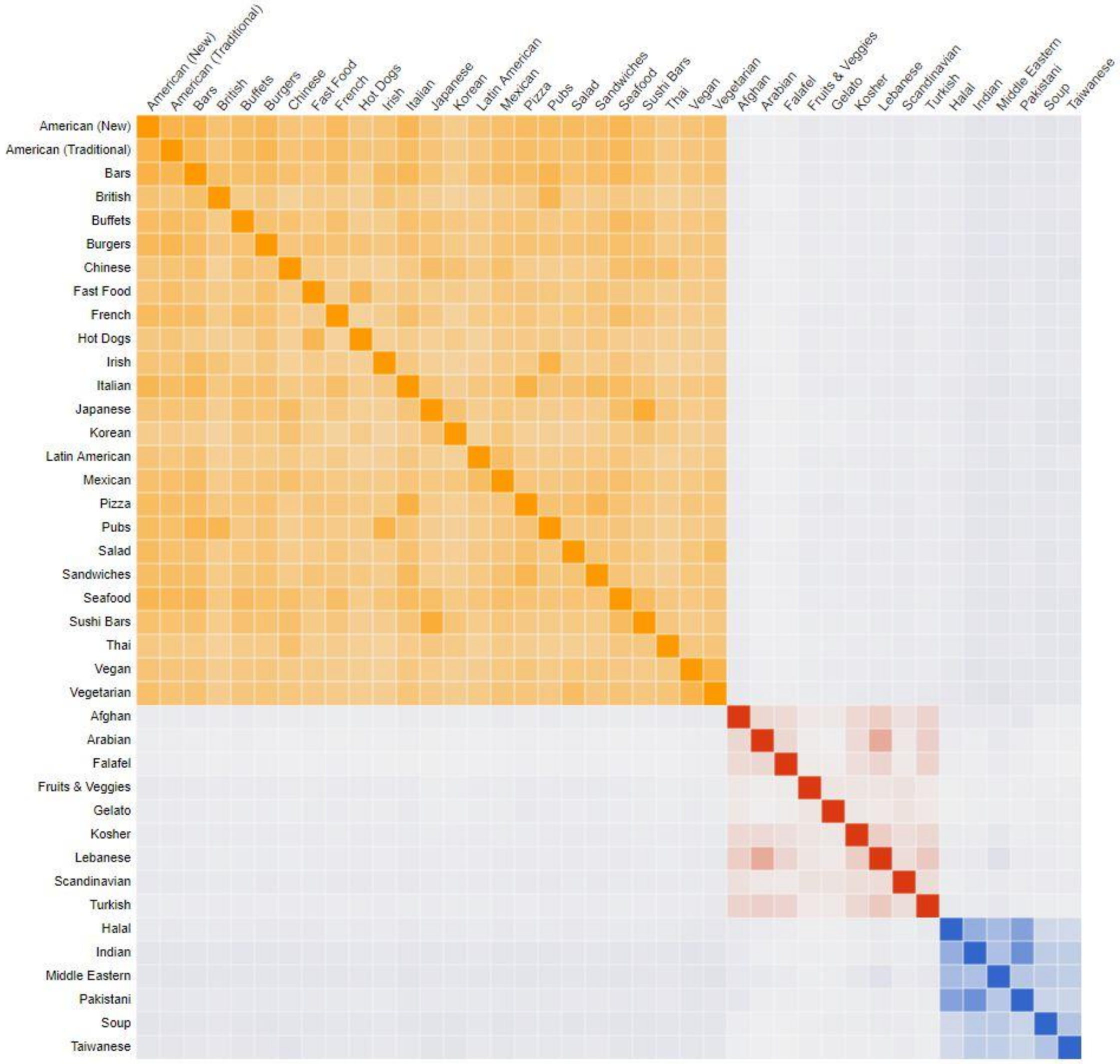
In this subtask, the results of the task 2.2 were taken and clustering was performed to visualize the cuisine clusters. Scikit-learn K-Means and Agglomerative Clustering clustering was used to cluster the similarity values in varied number of clusters to determine any hidden correlation in the matrix similarity. This task was divided into 2 parts (discussed in subsequent sections):

- Clustering with K-Means clustering
- Clustering using LDA with K-Means and Agglomerative Clustering

4.2 Task 2.3 A: Incorporating K-Means Clustering in Cuisine Map

For visualization, both similarity matrices and a forced directed graph for cluster proximity were generated in all cases.

At first the number of clusters was kept to three, while other settings were retained as in Task 2.2. The results are visualized in **Figure 6** and **Figure 7**.



Max Features	10000	Max df	0.8
Norm	L2	Min df	3
IDF used	TRUE	# Clusters	3

Figure 6, K-Means Clustings with 3 clusters

Here one can see the three clusters, which roughly correspond to be those of 'Western Cuisines', 'Middle Eastern/African Cuisines' and 'Asian/Subcontinental' cuisines. The similarity is more clearly visible in cluster graph, where the distance between the two nodes (representing cuisines) is proportionate to the similarity between them; i.e. if the cuisines are similar or related, they would be aligned close to one another in the graph as shown in **Figure 7**.

As one can see in **Figure 7**, the orange clusters contained 'Pakistani', 'Indian', and 'Halal', which seemed quite reasonable. Similarly, cuisines such as 'Lebanese', 'Turkish', 'Falafel' and 'Arabian', being in the blue cluster is also reasonable.

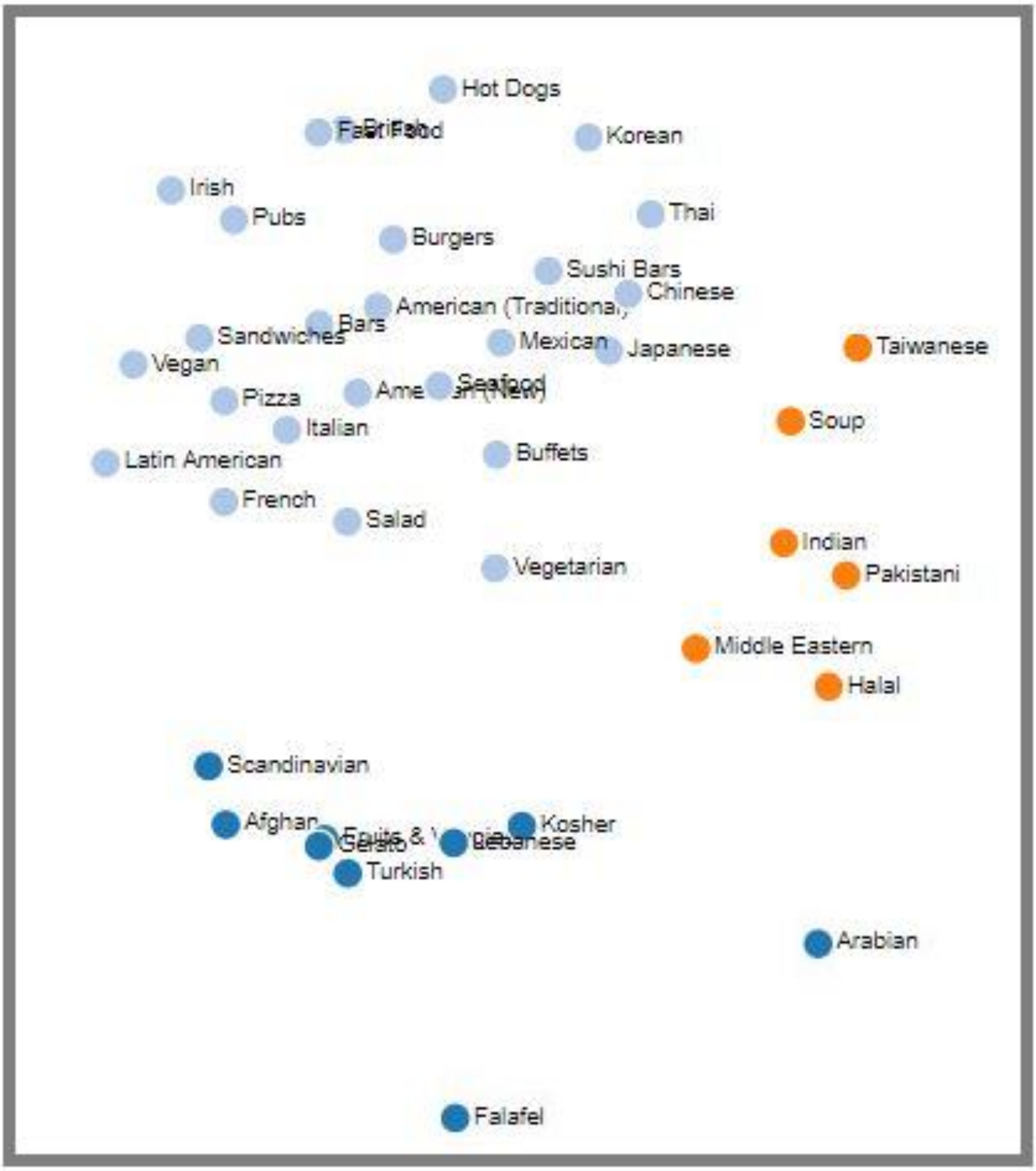
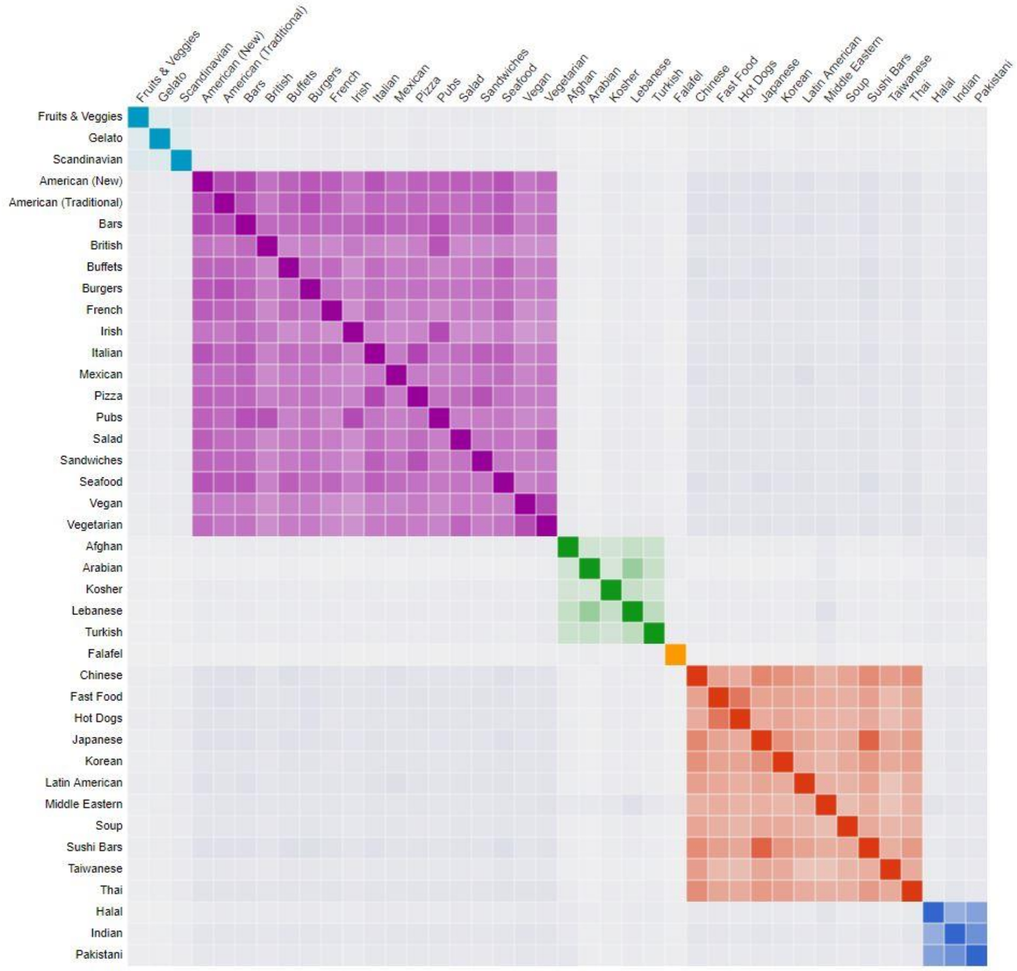


Figure 7 Cluster Chart with 3 Clusters

However, the cuisines in cyan cluster were not entirely separable, and called for increasing the number of clusters. For example, common sense dictates that American New and American Traditional cuisines should be in the same cluster, and that Japanese, Korean and Sushi Bars may also be placed in same cluster, but all of them being in the same cluster does not seemed very right. Similarly, it can be sees that placing Vegan and Vegetarian cuisines in the same cluster as of Burgers, Hotdogs and Fast food is also not very accurate. Therefore, in the numbers of clusters were increased to 6 and 9, and one can see that the clustering become more and more logical. The visualizations for 6 clusters are shown below in **Figure 8** and **Figure 9**, and the one for 9 clusters are shown in **Figure 10** and **Figure 11**.

Following sequence of charts show the effects of increasing the number of clusters, where one can see every split of a bigger cluster in a smaller one grouped more similar cuisines together in smaller clusters. For example, when 6 clusters were generated, Asian cuisines such as 'Chinese', 'Japanese' and 'Thai' got their own cluster. Similarly with 9 clusters, 'British', 'Irish' and 'Pubs' got together to become one cluster.



Max Features	10000	Max df	0.8
Norm	L2	Min df	3
IDF used	TRUE	# Clusters	6

Figure 8 Cuisines Clustering with 6 Clusters

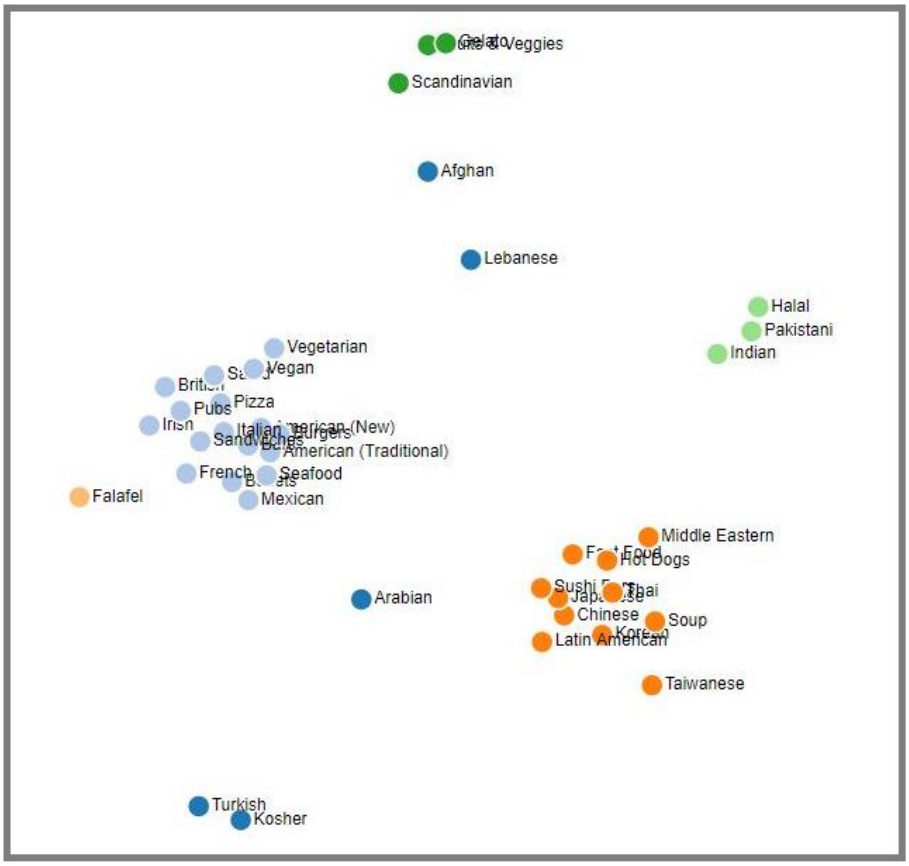
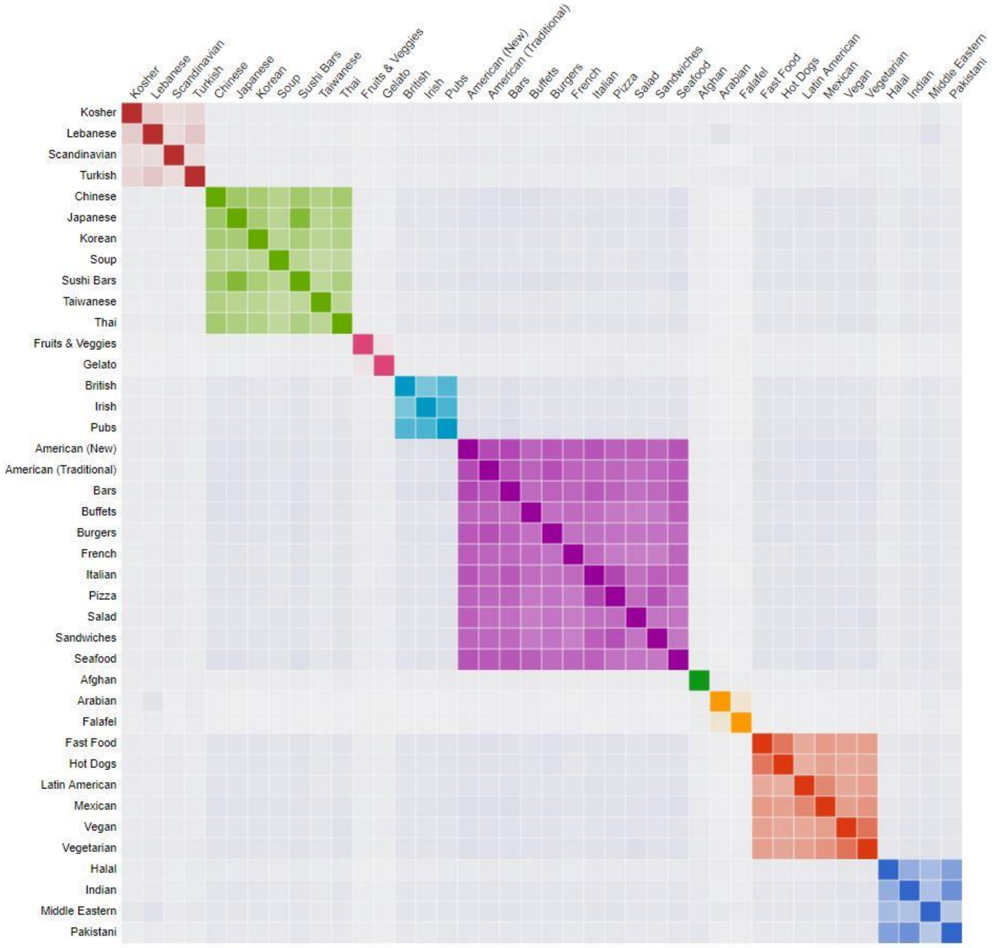


Figure 9 Cluster chart with 6 clusters



Max Features	10000	Max df	0.8
Norm	L2	Min df	3
IDF used	TRUE	# Clusters	9

Figure 10 Cuisines Clustering with 9 Clusters

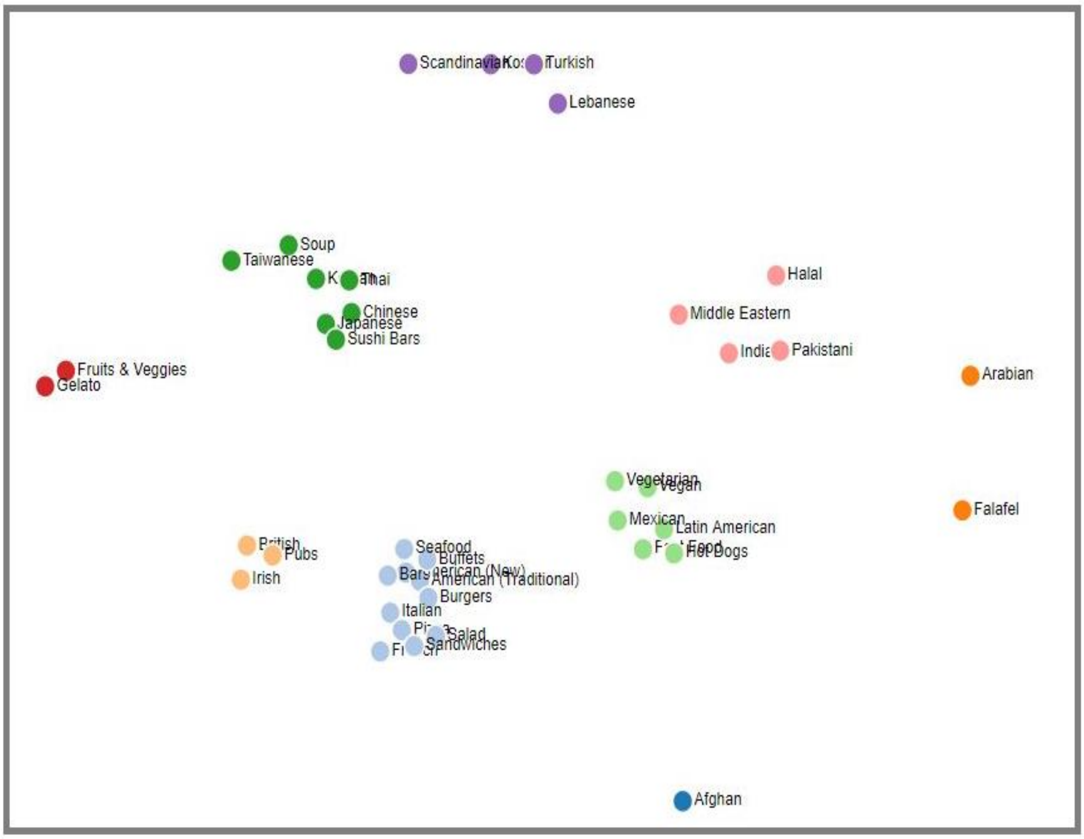


Figure 11 Cluster chart with 9 clusters

4.2 Task 2.3 B: Using LDA modeling with Clustering in Cuisine Map

In this subtask, we improve the clustering diversity with LDA. The number of Topics was set to 50, while the max feature setting was reduced to 2500 features and the IDF was set at a value of 0.5. The LDA performed 100 iteration with 10 workers on a cluster of 10, based on reviews groups of 200 concatenated reviews for a category with a total of 5000 reviews per category. The result of this analysis are visualized in **Figure 12** and **Figure 13**.

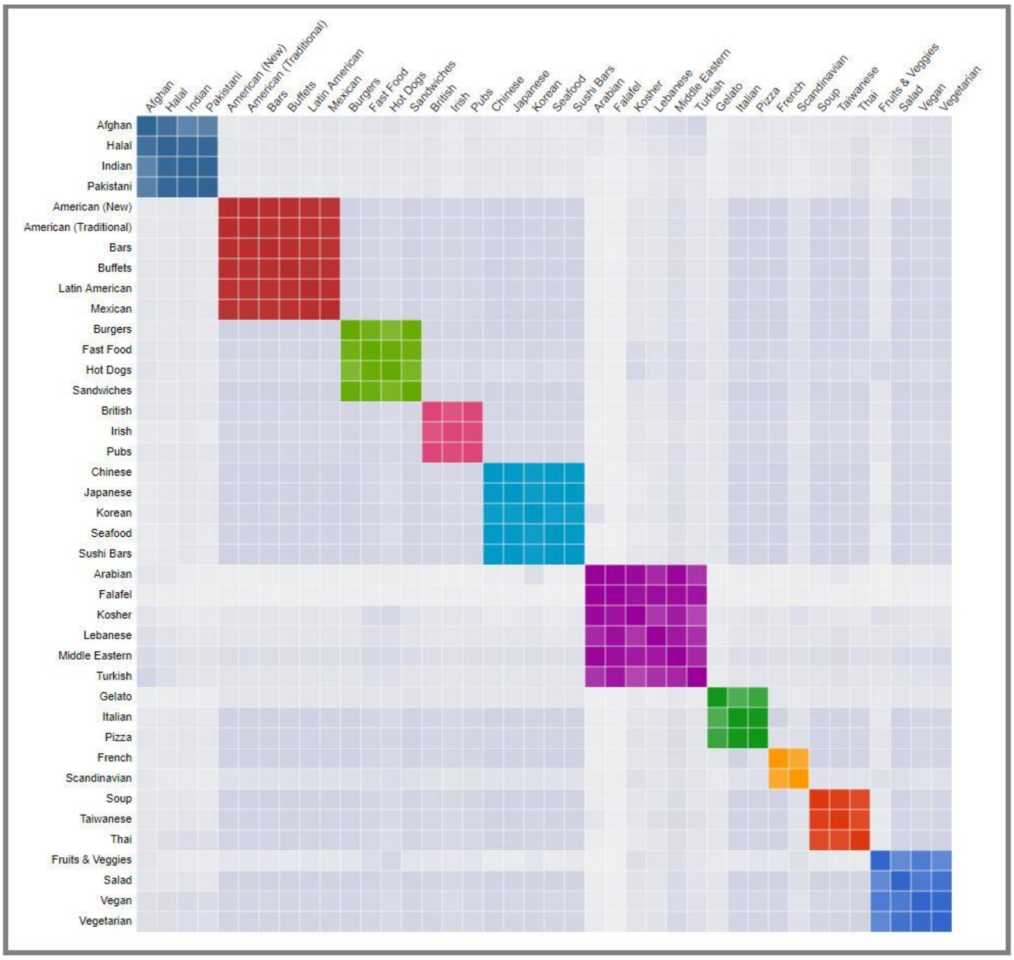


Figure 12 Clustering with LDA, Similarity Matrix

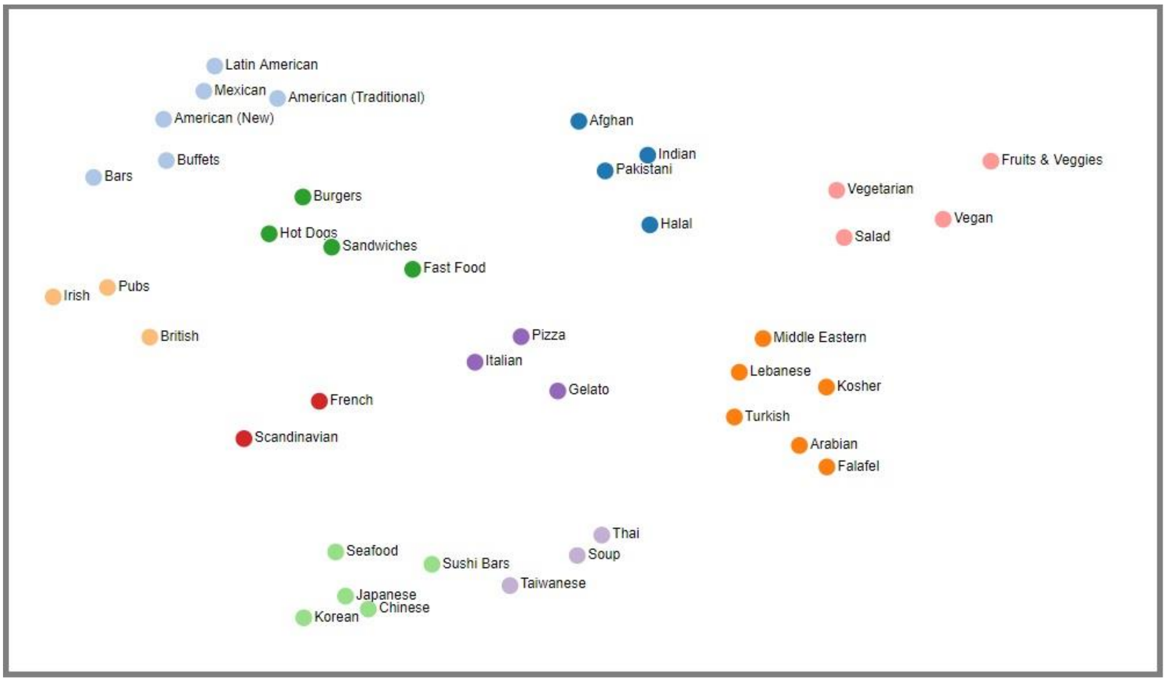


Figure 13 Clustering with LDA, Cluster chart

As can be seen, clustering with LDA produced reasonably accurate results where correlation between different cuisines, cuisine types and styles were clearly visible.

To further analyze the relation, agglomerative hierarchal clustering was used with LDA topic modeling. The results were used to generate Cluster Dendrogram as shown in **Figure 14**, and shows two distinctive categories of food (among others), namely Continental and Oriental (or Western and Eastern).

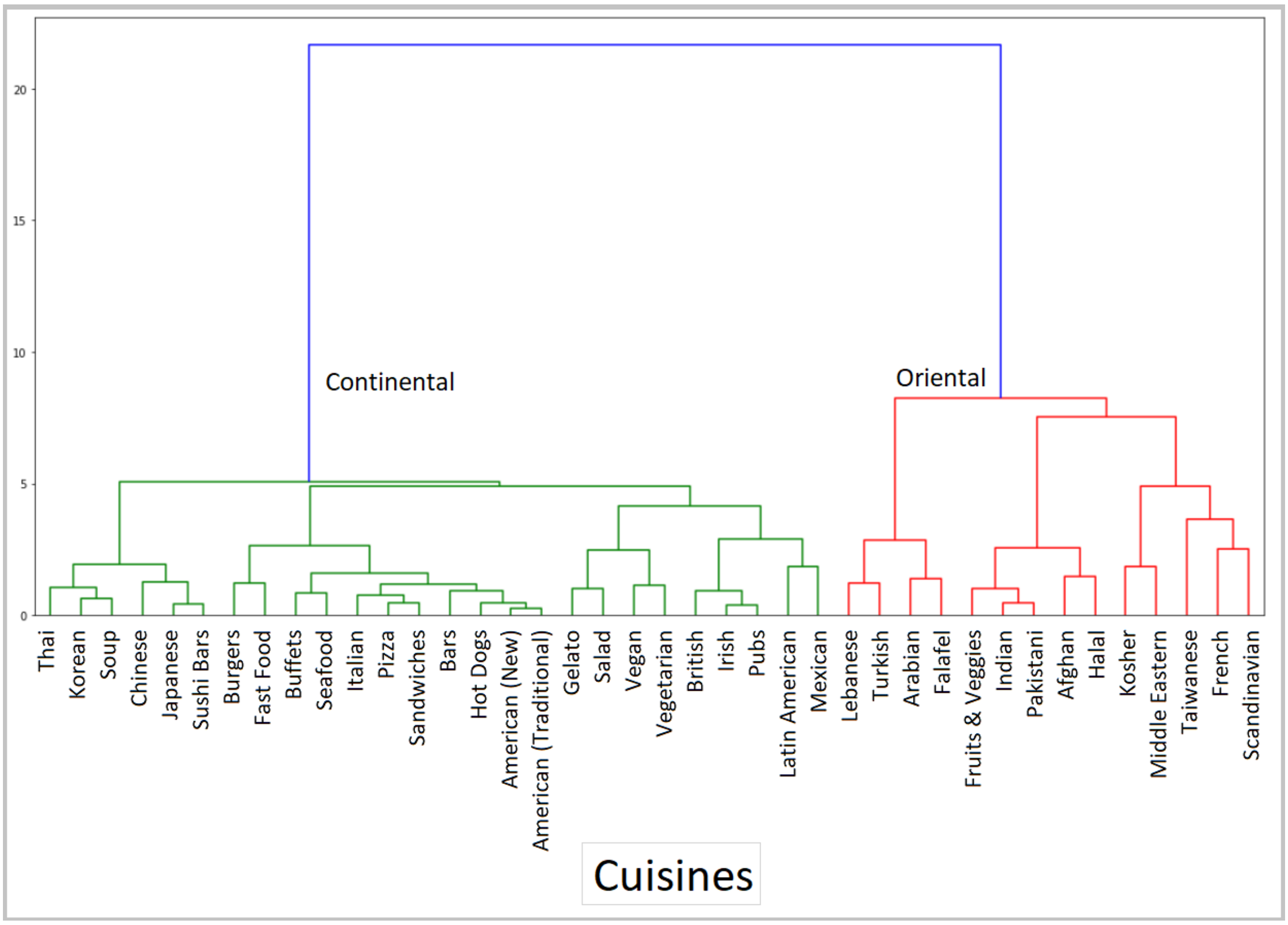


Figure 14 Agglomerative Clustering Dendogram

5. Conclusions

As can be seen, cuisine clustering can be improved using:

1 Processed text with specific POS lemmatized tokens and with stop-words removed
2 Using TF-IDF vectorization
3 Changing the settings of Max-Features and Min & Max DF
4 Using clustering algorithms on similarity calculations such as K-Means clustering.
5 Using topic modelling such as LDA with clustering.