## Cuisine Clustering and Map Construction:

A random sample of 30 cuisines (categories) from the sample data (containing reviews for all categories) that was available for download for Task 2 was considered for cuisine clustering and map construction.

**Task 2.1:** All the reviews for respective cuisine (category) were concatenated into one line. These lines (documents) were then sent to term frequency vectorizer (from **python gensim**) to fit and transform the text into term frequency **(TF)** form.
Following parameters were given to initialize the vectorizer:-
max_df=0.5 (Max document frequency), max_features=10000, min_df=2 (Min document frequency), stop_words='english', use_idf=False (No Inverse Document Frequency)

Once the term frequencies for 30 documents (one for each category) was obtained then cosine similarity matrix was calculated for these 30 documents.

This cosine similarity matrix was fed to **R** for visualization. **heatmap** library was used to visualize the matrix. Here is the output. X and Y axis represents the cuisines (categories) and each cell in the chart represents the similarity strength. Opacity is used to represent similarity. The darker the cell the more similar are the corresponding cuisines on X and Y axis. We can see that all the diagonals are dark as they represent similarity of a cuisine to itself.

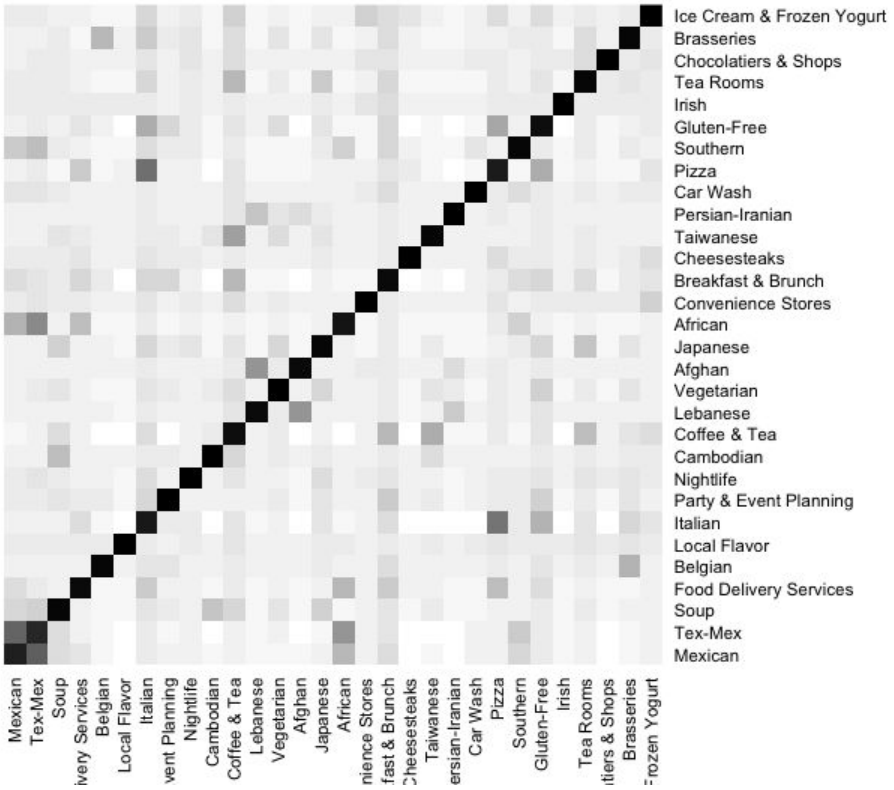


**Fig (1) - cuisine map using TF for cosine similarity**

In the above visualization we can clearly see some of the obvious similarities are reasonably dark - e.g. **Mexican and Tex-Mex**, **Italian and Pizza, Coffee & Tea and Breakfast and Brunch.** But some are definitely not that correct - Tex-Mex and African or Taiwanese and Coffee and Tea

**Task 2.2: Following changes were applied in sequence to represent the text better.**

a. **While initializing the term frequency vectorizer, the inverse document frequency was set to True,** this helped in weighing the more important terms correctly. All the other parameters and process was same as in 2.1, visualization output is given in **Fig 2 (a)**
b. LDA Topic Modelling was applied on the 30 documents (each representing the entire set of reviews for particular cuisine). Following were the parameters used for LDA: number of topics=100 and number of iterations=1000. The cosine similarity was calculated using the **topic distribution per document ($\pi_{d,i}$)** instead of TF/IDF. The visualization output is given in **Fig 2(b)**
   The document similarity matrix was then again sent to **R** for visualization.

For both 2(a) and 2(b) - ggplot2 library was used for visualizations. **ggplot** allowed to display the color dimension easily using **alpha** parameter for fill and **col** parameter for colors.

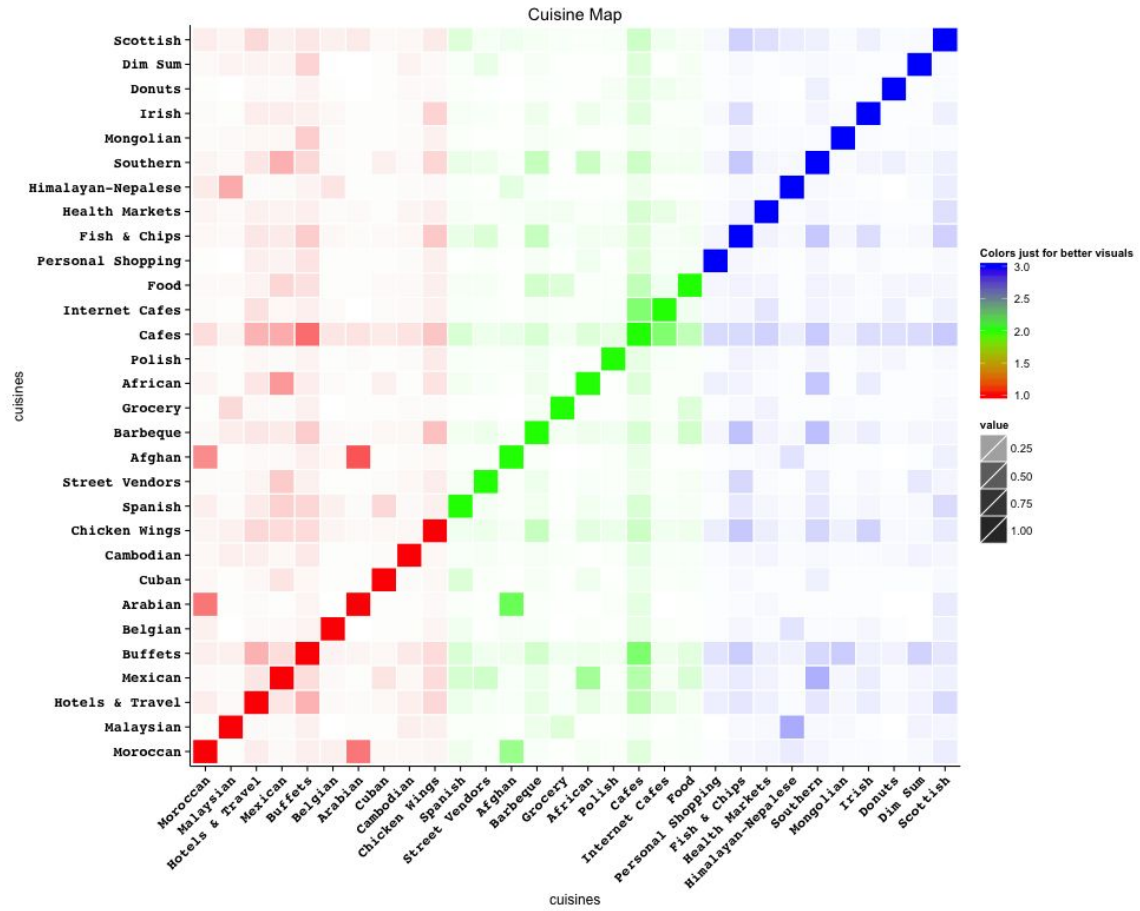


**Fig 2 (a) - cuisine map using TF-IDF for cosine similarity**

Please note that the colors here are only for better distinction of elements. It does not represent any cluster as this is task **2.2.**

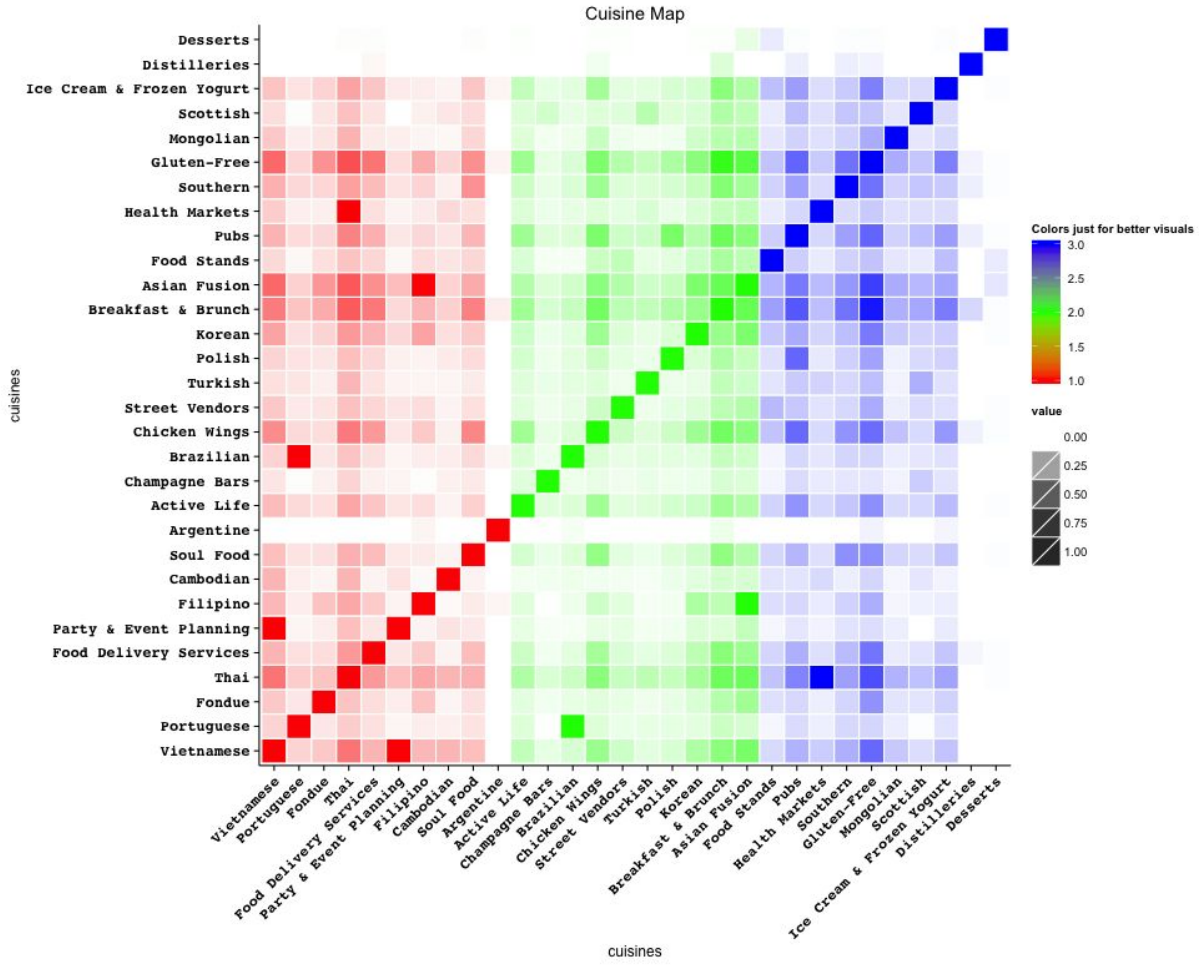


**Fig 2 (b) - cuisine map from LDA clustering - 100 topics**

**Varying the similarity function:**

The output from LDA topic modelling (with number of topics = 10) was used to vary the similarity function. Instead of using all the 10 topics from document-topic distribution ($\pi_{d,i}$), we used only first 5 topics from each document while calculating the cosine similarity.

All the other parameters/visualization process was same as for Fig 2 (b).

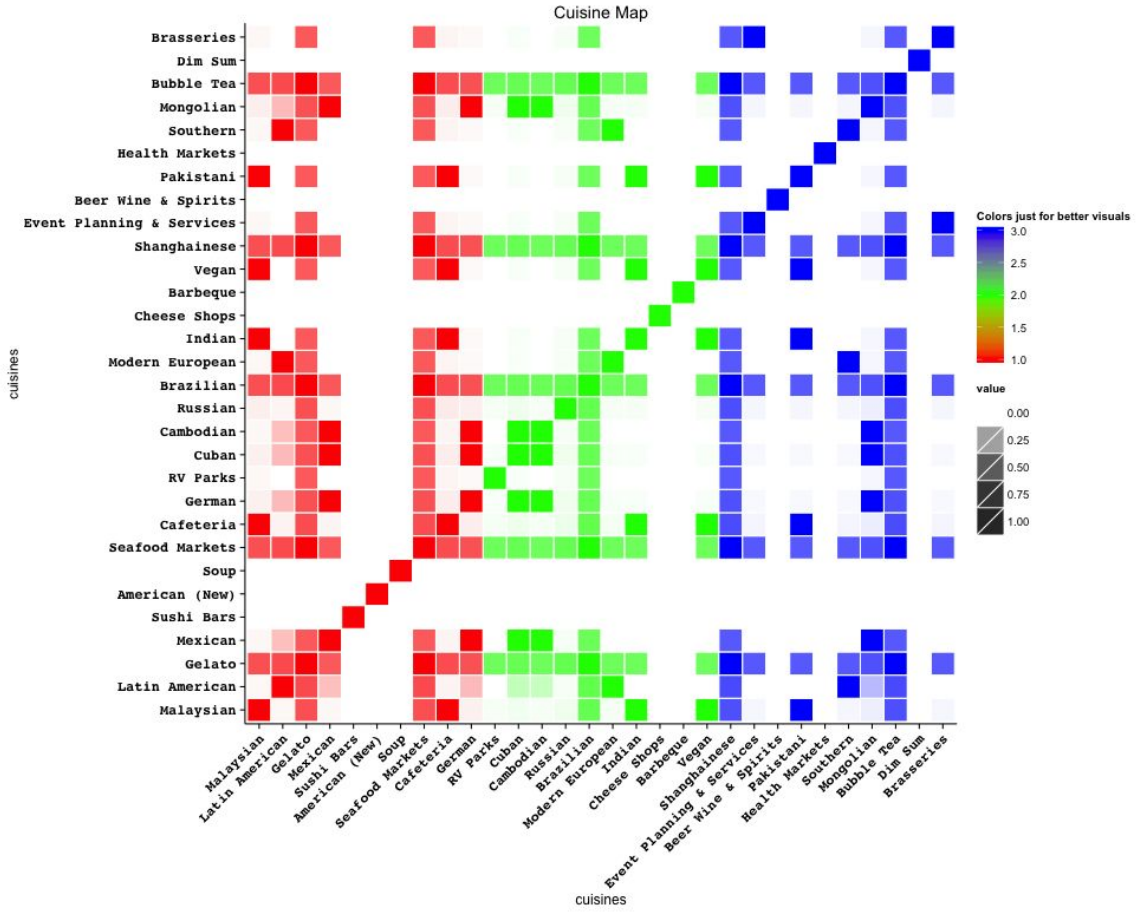


**Fig 2 (c) - Task 2.2 similarity function changes (only 5 topics considered)**

We can clearly see because we only considered five topics, we are able to emphasize the similarity on certain cuisines and de-emphasize some other cuisines. Around 4-5 cuisines are now not similar to any cuisine because the topic they represented the most was taken out of similarity calculations. But for most of the other cuisines we can see the similarities are more. Again colors here are only for better visuals.

## Task 2.3

In this task, the objective was to discover the cuisine clusters and visualize it using cuisine map that will show clusters as well the similarity among all cuisines considered.

For this task, again sample of 30 cuisines were taken and LDA Topic mining was applied with same parameters as in **2.2 (b)** using 100 Topics and calculated the cosine similarity by using the document-topic distribution i.e. ($\pi_{d,i}$).
Throughout this task, same set of cuisines are used but similarity clustering algorithms (it will be mentioned in visualization title).

**Clustering:**
Since the cuisine relationships can be complicated and we don't want to rely on false bias or intuition, a relative measure called **"Silhouette co-efficient"** was used to evaluate quality of clusters. It was decided not to use external measures due to lack of expert labelled data.

Partitioning Around Medians **(PAM)** algorithm and **Hierarchical clustering (**Lance–Williams algorithms from hclust in R**)** were selected for clustering. PAM was selected as it is less sensitive to initialization as compared to K-means. **Cluster** package in **R** was used to apply the PAM and hierarchical clustering algorithms. The **similarity** matrix was converted into **distance** matrix using **distmatrix** function in **hopach** library. Distance was calculated using cosine of the angle values (d="**cosangle**" parameter) from similarity matrix.

The distance matrix was then given to **PAM** and **hclust** function to cluster the cuisines.

The following parameters were given to the PAM and hclust:
**pam:** distance Matrix, **k** =(number of clusters) , diss = TRUE (indicating the matrix passed as data is dissimilarity matrix) , keep.diss = TRUE (indicates dissimilar elements will not be filtered)

**hclust:** distance Matrix (by converting to **dist** type structure) and (method was kept was default - "complete" linkage). hclust function uses agglomerative

To identify an optimal K value for passing to pam and hclust algorithms, average **Silhouette** width values (silhouette coefficient) were calculated for different values of **K** for **pam** and **cutree. Fig 3 (a)** shows plot of silhouette values for different K using pam and **Fig 3 (b)** shows plot of silhouette values for different K using cutree (hclust) and silhouette functions.

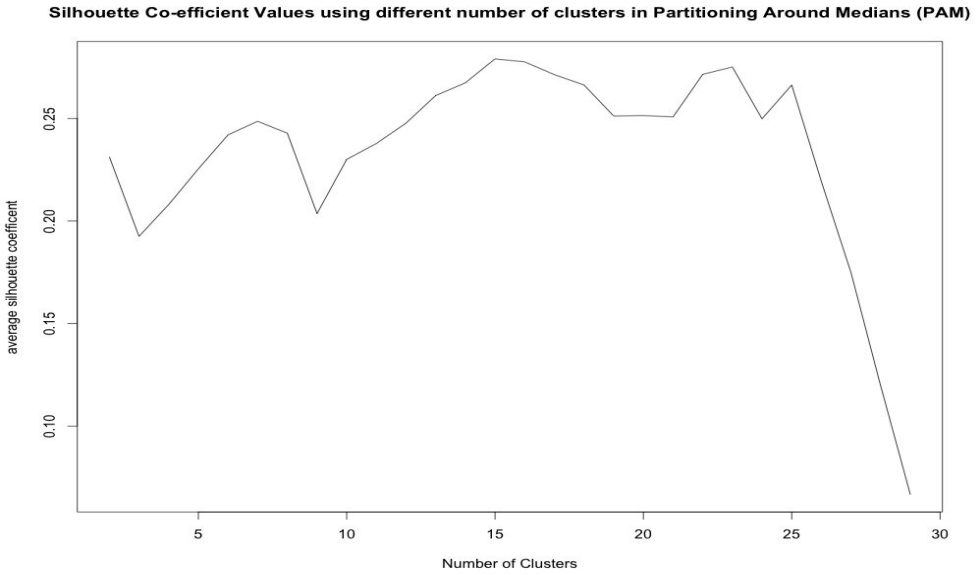


**Fig 3 (a) - Plot of Silhouette avg width for different Number of Clusters in PAM**

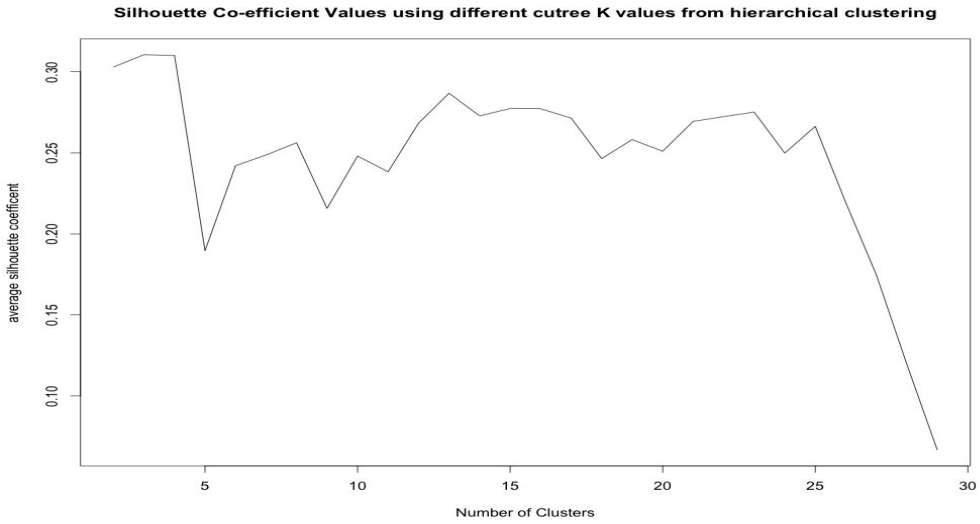


**Fig 3 (b) - Plot of Silhouette avg width for different number of clusters using hclust/cutree**

**From 3 (a)** , it was observed that around 15 clusters gives a reasonable silhouette value (0.3). This is not good but it seems reasonable considering the sample has variety of cuisines. Using K=15 following cuisine map was obtained using **PAM and ggplot2 library in R**

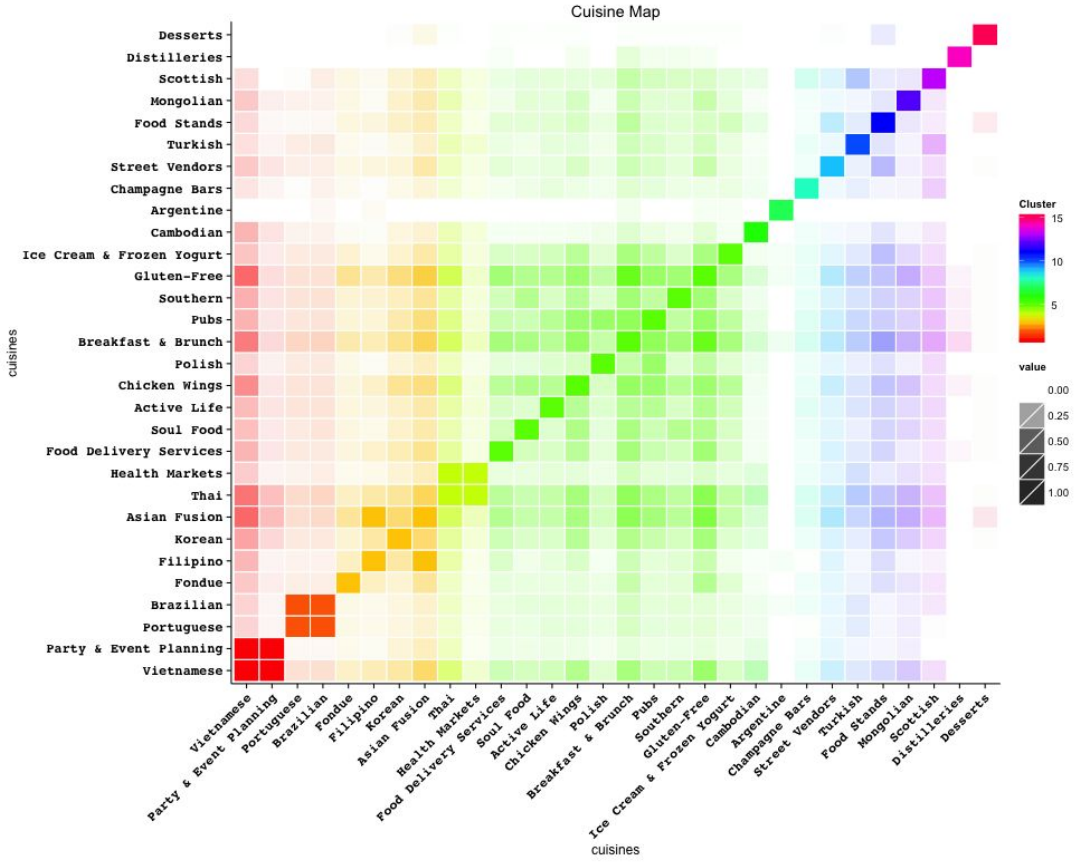


**Fig 3(c) - PAM clustering (k=15) with similarity matrix from 100 topics (from LDA)**

**In the Fig 3 (c) clusters are grouped by color. The opacity of color represents similarity strength of a cuisine on X-axis with a cuisine on Y-axis.**
We can definitely see **green** as a big meta category which consists of comfort food like food trucks, street vendors, breakfast and Brunch, Pubs, Soul Food etc. There is also **orange** meta category which consists of **Asian Food like Filipino, Asian Fusion and Korean.** There are lot of single cuisine clusters which could not fit in with other cuisines.

To get to more bigger meta categories instead of small clusters, K was reduced while using **cutree** to get output from **hierarchical** clustering. **Silhouette** plot for cutree (hclust) indicated that there are possible two peaks at **K=12 and K=15.** K=12 has smaller silhouette value than value at K=15 but some trade-off was acceptable considering we wanted a cuisine map (more meta categories). Here is the cuisine map for **K=12 value in cutree**

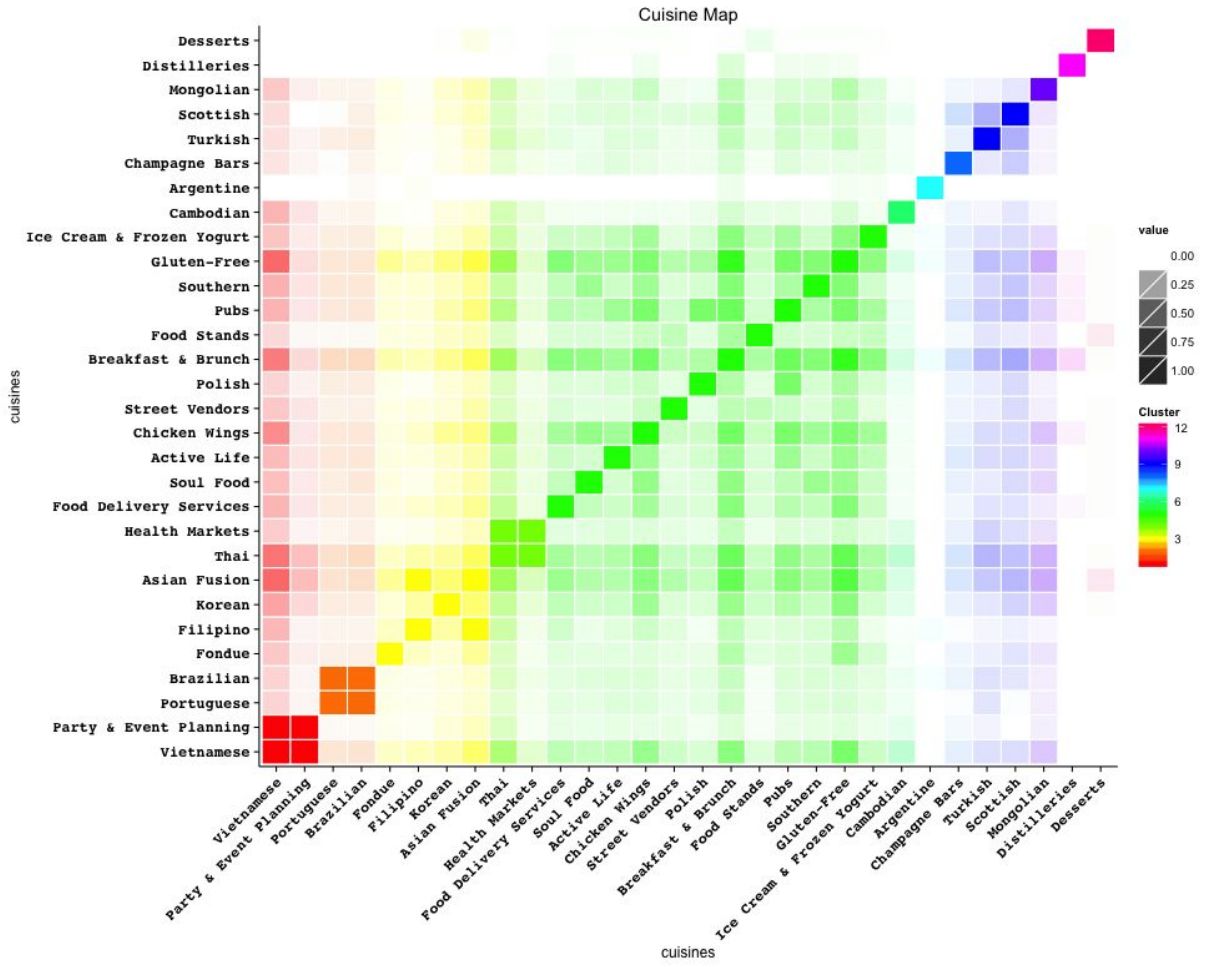


**Fig 3 (d) - Hierarchical Clustering (hclust) k=12 cutree - (default method=complete linkage) using similarity matrix from LDA topic mining (100 topics)**

**In the Fig 3 (d) clusters are grouped by color. The opacity of color represents similarity strength of a cuisine on X-axis with a cuisine on Y-axis.**

The above figure 3(d) gave a better picture to suggest possible 4 meta categories: **South American** (See Orange:Brazilian/Portuguese cuisines), **Asian** (See Yellow:Filipino,Korean and Asian Fusion), **Comfort Food** (see Green: Health Market, Soul Food, Food Stands, Pubs, Southern, Breakfast and Brunch..) and **European** (See Purple: Turkish/Scottish)

**Conclusion:**

The clustering definitely helps to visualize the cuisines with a different point of view. It gives an idea on the meta categories and also helps user in discovering new cuisines which are similar to a known liked cuisine. The silhouette coefficient helped in giving an approximate optimal number of clusters to select for performing clustering algorithms. It was also a good learning experience on how every step in clustering can impact the end result (cuisine map). All the steps from text representation, term frequency-inverse document frequency (TF-IDF), topic mining, similarity functions and clustering algorithms and their evaluations are all very important stages.Some experimentation to vary number of topics in LDA and use either subset (like in 2 (c) )or all of topics to calculate similarity matrix for clustering using PAM/hclust algorithms in R still could have been done but was left due to time constraints. All the variations can definitely give different and interesting results.