Data Mining - Capstone Project - Task 2

Tran Ho Thanh Dong

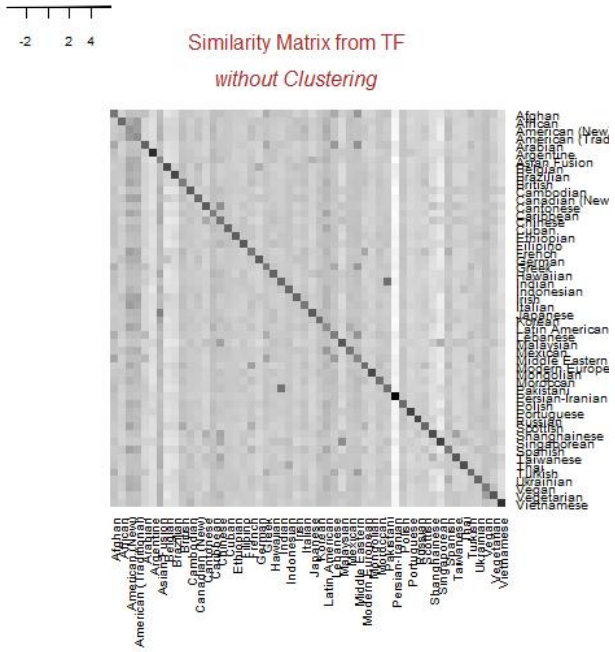
- ***R language is used to clean, explore , analyze and visualize data. The R allows to visualize the heatmap of data with and without Clusters incorporated***

Load and Prepare data

- A corpus of reviews from 52 cuisines with name of specific country such as Vietnamese, Thai, Indian... are created. **A Term-Document Matrix** is created based on the corpus in which **Terms** (bag of words) as **rows** and **Document** is represented by a vector of term -frequencies as **columns. Cosine** method is used to calculate similarity

Task 2.1 - Visualization of the Cuisine Map

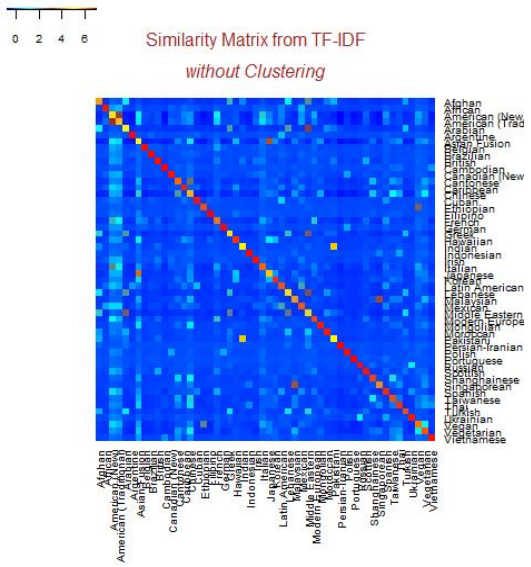
- A similarity matrix between above vector-documents (columns) is calculated with **term frequency** (**TF**) as **weight.**
- The similarity matrix is visualized in a heatmap graph in which the opacity of each cell is the similarity between two documents - with a higher opacity for a higher similarity

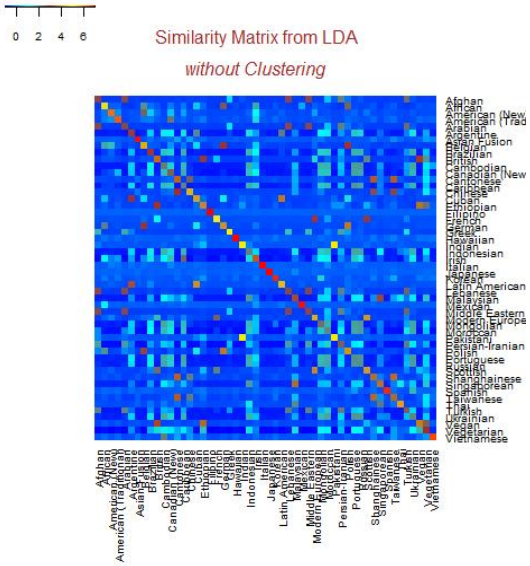


Task 2.2 - Improving the Cuisine Map

- The similarity matrix is calculated with **term frequency - inverse document frequency** (**TF-IDF**) as **weight**. The heatmap is almost the same in Task 2.1. However, the degree of similarity between documents seems lower in the Task 2.1 . Varying the weighting of terms from TF to TF-IDF did nor improve the quality of the cuisine map

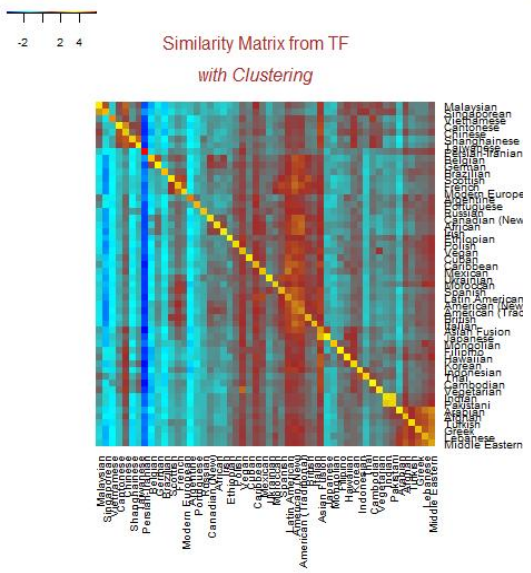
- A another method to represent a document and calculate similarity matrix is based on topics which are extracted from the corpus of reviews by **Latent Dirichlet Allocation (LDA)** models. **Document** is represented by a vector of **posterior probabilities of the topics.** A comparison between 02 plots shows that the cuisine map applied **Topic Model** displays groups of very similar documents in different area. Applying Topic Model improves the quality of the cuisine map.

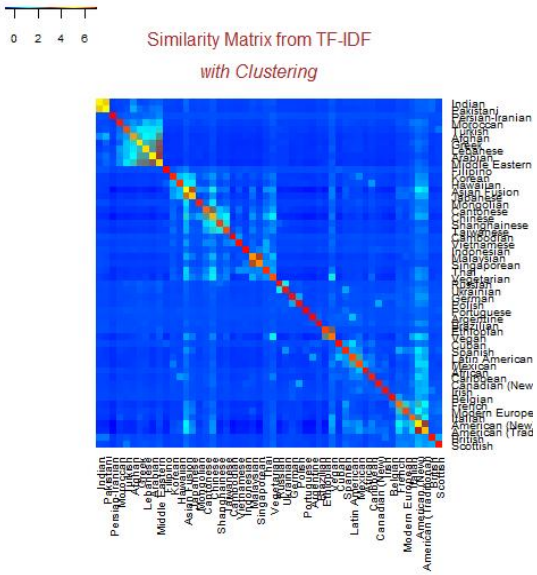




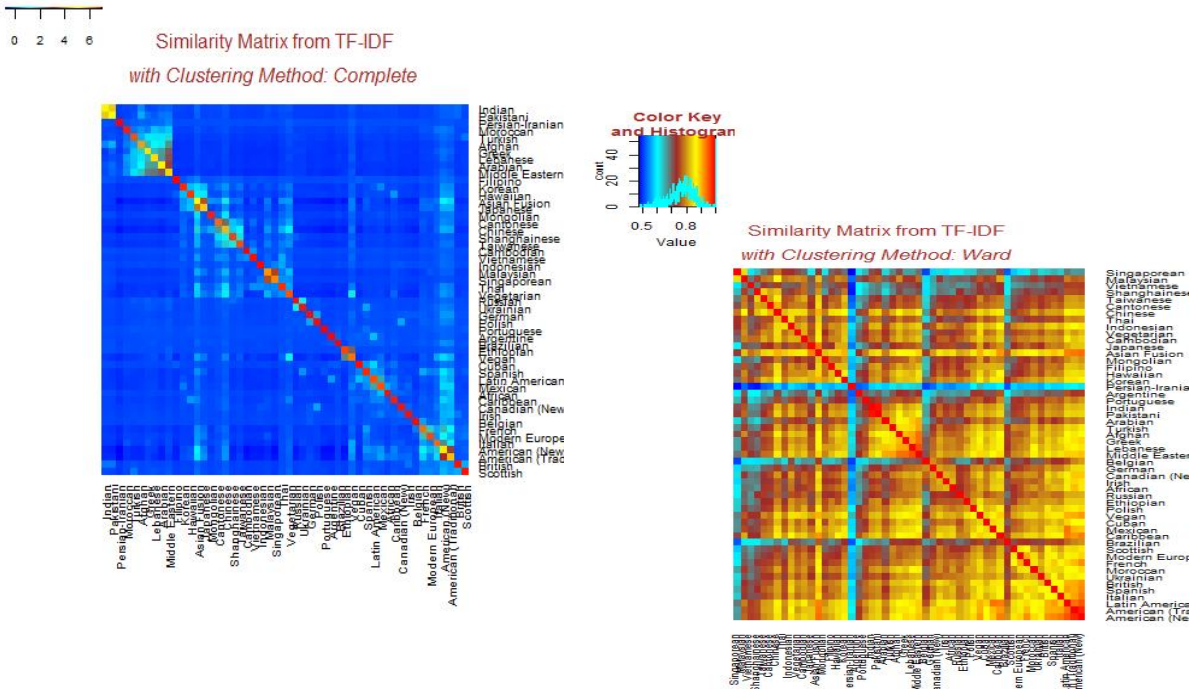
Task 2.3 - Incorporating Clustering in Cuisine Map

- As mentioned above, R allows to display similarity matrix with and without clustering. Please see the cuisine maps incorporated clustering.

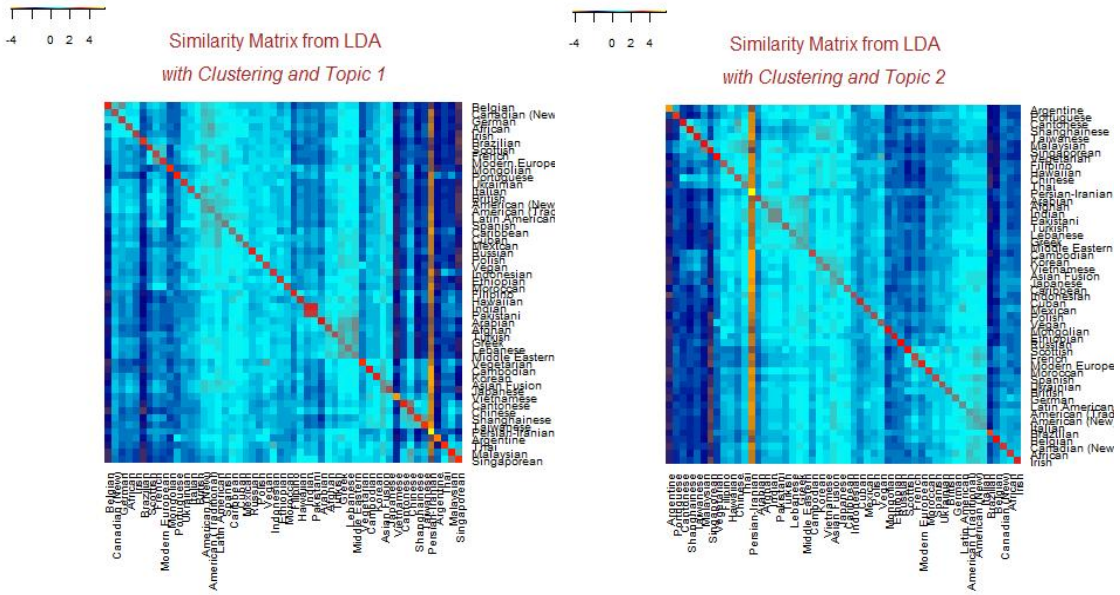




- Two Clustering Algorithm: **Complete** and **Ward** are used for **similarity matrix with TF-IDF weight**, the cuisine maps are absolutely different



- Topic 1 and Topic 2 extracted by LDA model is used to generate a matrix in which **Document** is represented by a vector of **posterior probabilities of terms** in a topic.



- The cuisine map display different clusters which represent main categories of cuisines such as (**Chinese, Cantonese, Shanghainese, Vietnamese, Taiwanese**), (**Greek, Lebanese, Afghan, Arabian**), (**Indian, Pakistan**)...