Data Mining Project - Week 2 - Cuisine Clustering and Map Construction

Data Mining Specialization - Coursera / University of Illinois at Urbana-Champaign

- Author: Michael Onishi
- Date: November, 2019

Introduction

The goal of this task is to mine the data set to construct a cuisine map to visually understand the landscape of different types of cuisines and their similarities. The cuisine map can help users understand what cuisines are available and their relations, which allows for the discovery of new cuisines, thus facilitating exploration of unfamiliar cuisines.

For this project, I used the python programming language along with some libraries such as pandas, matplotlib and sklearn.

I applied basically the same data cleansing and normalizing made on the task 1:

- Removed all the reviews not about restaurants
- Removed accents, http links, line breaks and tabs
- Converted to lowercase

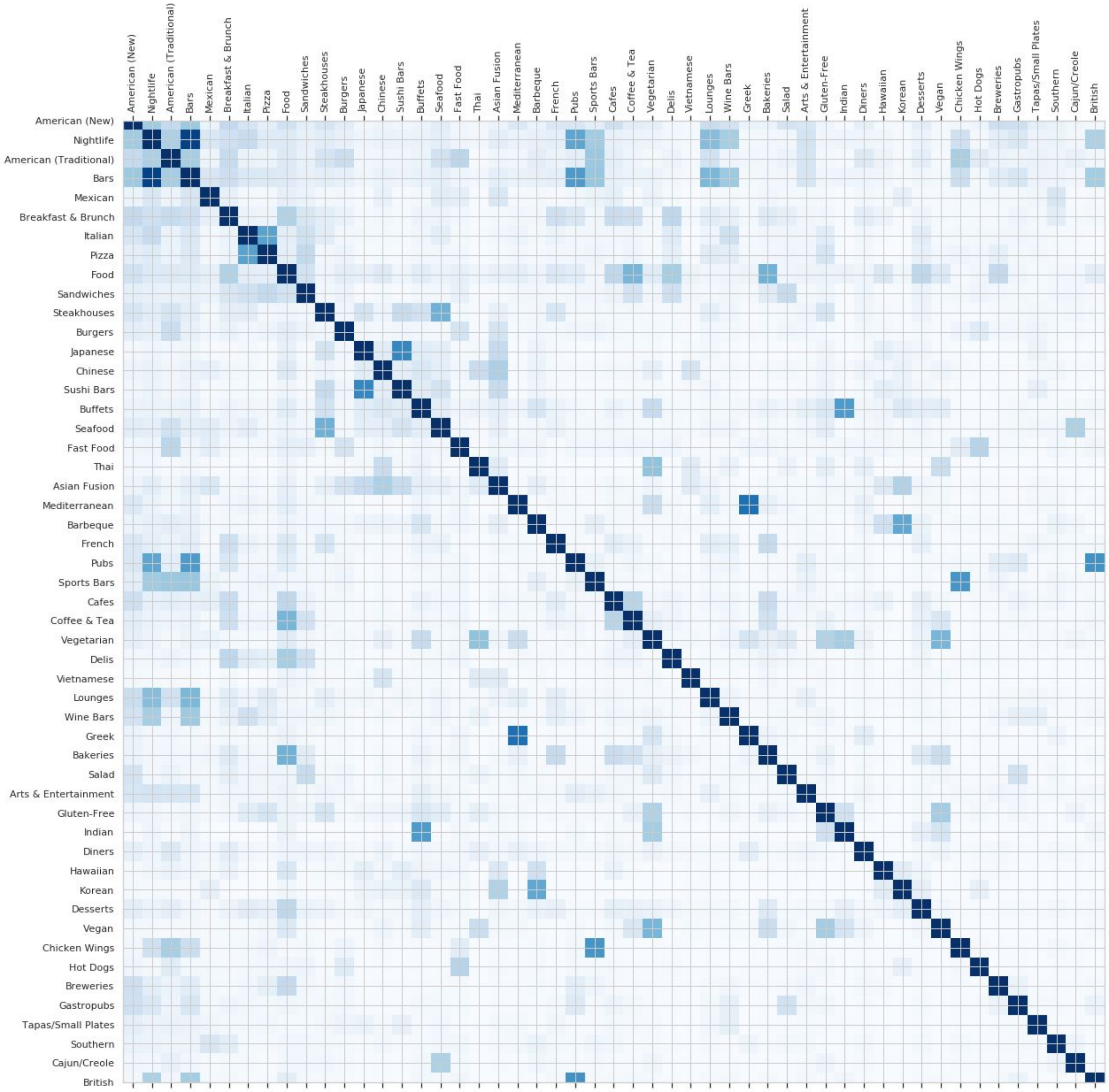
Additionally, I only considered restaurant categories with more than 5,000 reviews, because I would like to have a decent dataset size for every category and it would be easier to visualize less categories. With this filter, I got 51 categories and a total of 1,347,264 reviews.

The full notebook with all the steps and details taken here, making it possible to completely reproduce this work, is available at:
https://github.com/michaelonishi/coursera-data-mining-specialization/blob/master/c6-data-mining-project/task2/Data_Mining_Project_Week_2_Cuisine_Clustering_and_Map_Construction.ipynb

Task 2.1: Visualization of the Cuisine Map

Use all the reviews of restaurants of each cuisine to represent that cuisine and compute the similarity of cuisines based on the similarity of their corresponding text representations. Visualize the similarities of the cuisines and describe your visualization.

First I applied a TF-IDF vectorizer limited to 100,000 features (words) for appended review texts for each category. The first natural choice of cuisine map was plotted as a similarity matrix from the tf-idf-weighted document-term matrix (higher opacity means higher similarity):

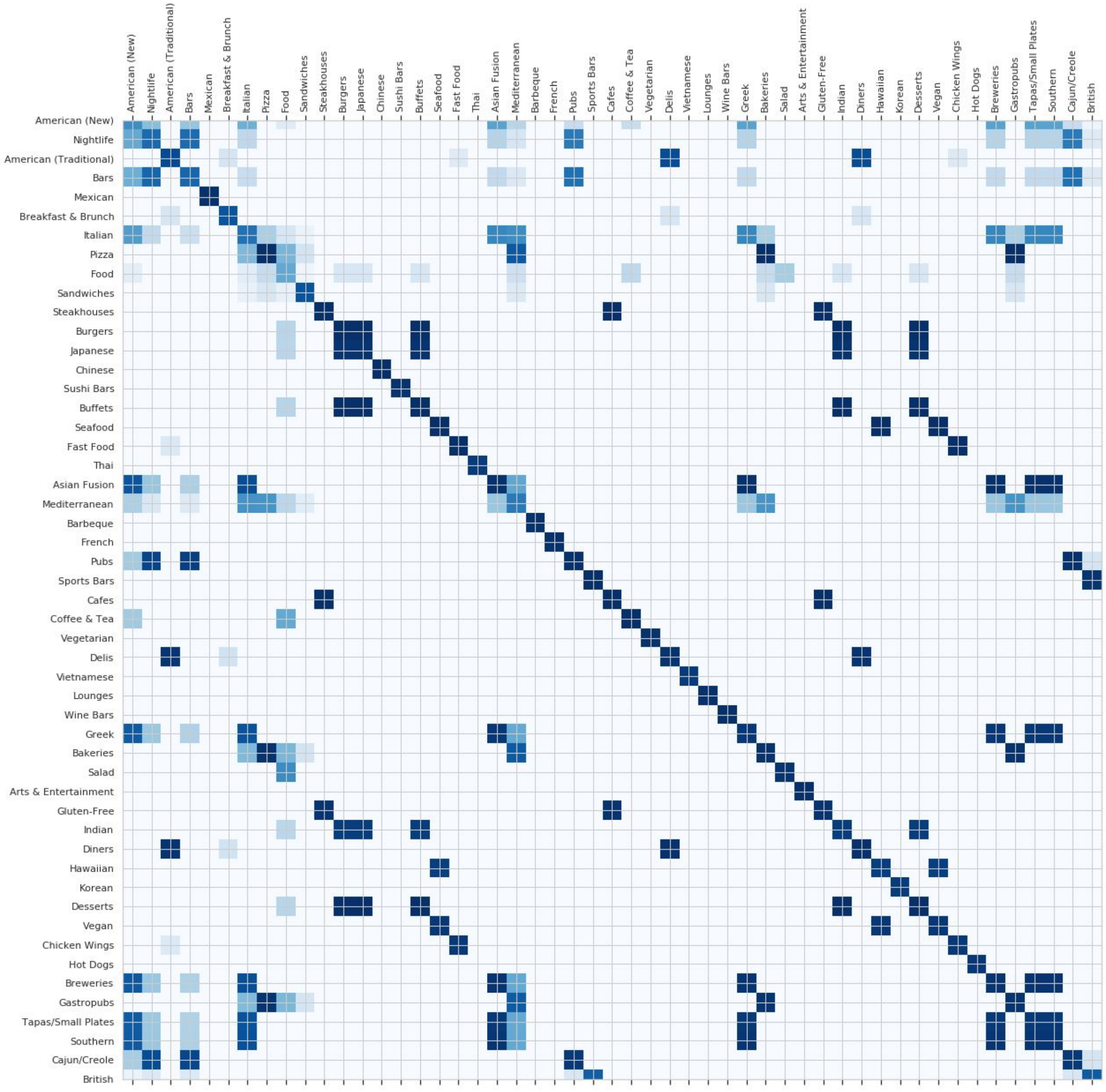


With this very simple visualization, we can already see some very interesting results as a high similarity between, for example:

- Bars and nightlife
- Greek and Mediterranean
- Chicken Wings and Sports Bar
- Sushi Bars and Japanese

Task 2.2: Improving Cuisine Map

After this, I used the same TF-IDF matrix as input to an LDA topic model with 50 topics (because I wanted to have many features to distinguish each category well) and got the following result:



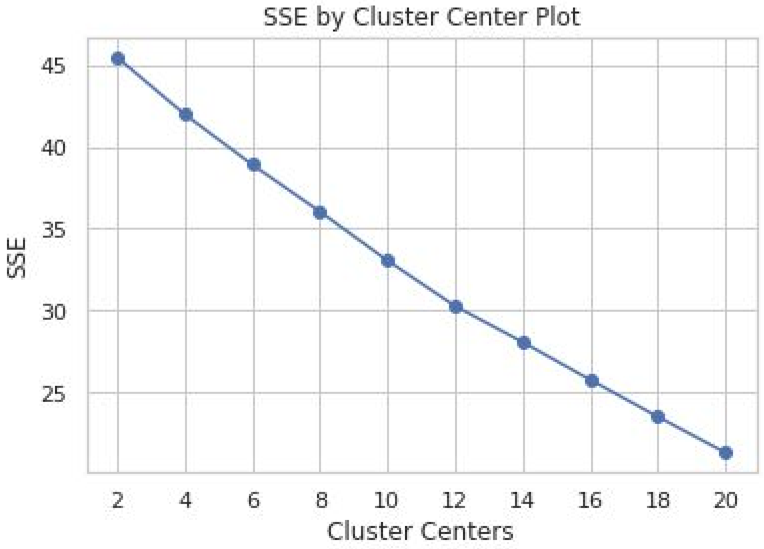
The results were somewhat similar, but with LDA the similarities are more evident and I would highlight the following similarities:

- Fast Food and Chicken Wings
- Sports Bars and British
- Pubs and Cajun/Creole
- Seafood and Vegan

Task 2.3: Incorporating Clustering in Cuisine Map

Use any similarity results from Task 2.1 or Task 2.2 to do clustering. Visualize the clustering results to show the major categories of cuisines. Vary the number of clusters to try at least two very different numbers of clusters, and discuss how this affects the quality or usefulness of the map. Use multiple clustering algorithms for this task.

First I used the K-Means clustering algorithm with the TF-IDF matrix, analysing the sum of squared errors (SSE) of the clustering results varying the number of clusters. Below I plotted a line graph showing how the SSE varied for number of clusters varying from 2 to 20.



Although the graph is almost linear, I chose 10 as a good number of clusters as it already gave very good results. Shown below:

Cluster 0 - Salad, Gastropubs

Cluster 1 - Steakhouses, Japanese, Sushi Bars, Buffets, Indian

Cluster 2 - Italian, Pizza, Sandwiches, Wine Bars

Cluster 3 - Nightlife, Bars, Pubs, British

Cluster 4 - American (New), American (Traditional), Burgers, Seafood, Fast Food, Sports Bars, Lounges, Arts & Entertainment, Diners, Chicken Wings, Hot Dogs, Tapas/Small Plates, Southern, Cajun/Creole

Cluster 5 - Mexican, Breakfast & Brunch, Food, French, Cafes, Coffee & Tea, Delis, Bakeries, Hawaiian, Desserts, Breweries

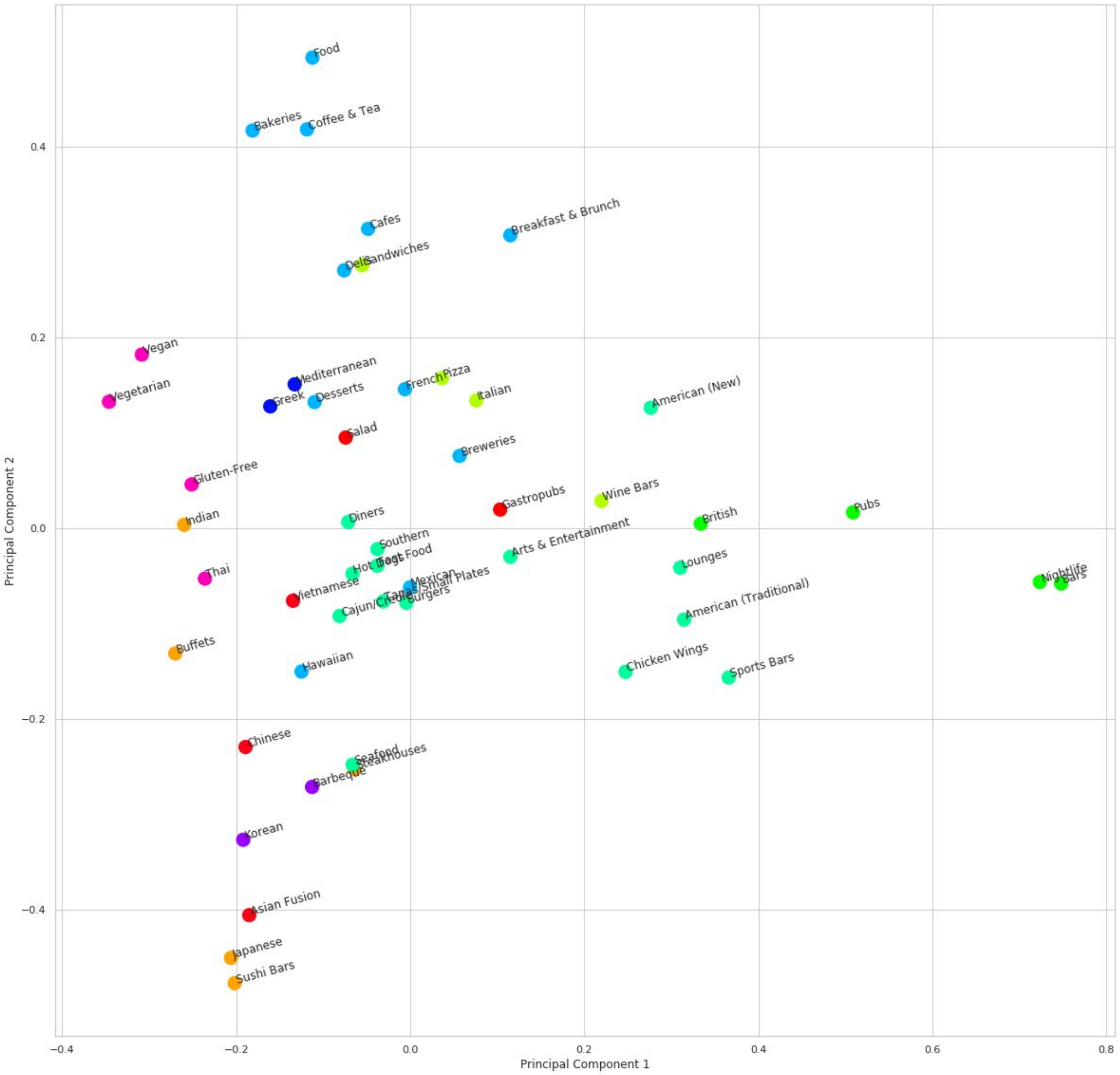
Cluster 6 - Mediterranean, Greek

Cluster 7 - Barbeque, Korean

Cluster 8 - Thai, Vegetarian, Gluten-Free, Vegan

Cluster 9 - Chinese, Asian Fusion, Vietnamese

We can clearly see that all the clusters have very reasonable selection of related cuisines. Next, to make clusters visualization possible, I applied a Principal Component Analysis algorithm to reduce the dimensionality of the matrix to 2 dimensions. Here I plotted a scatter plot with all the 10 clusters with distinct colors:



Next I tried another clustering method called Spectral Clustering. The results were similar:

Cluster 0 - Breakfast & Brunch, Food, French, Cafes, Coffee & Tea, Delis, Bakeries, Desserts

Cluster 1 - American (New), Mexican, Sandwiches, Burgers, Chinese, Fast Food, Asian Fusion, Vietnamese, Lounges, Wine Bars, Salad, Arts & Entertainment, Diners, Hawaiian, Hot Dogs, Breweries, Gastropubs, Tapas/Small Plates, Southern

Cluster 2 - Japanese, Sushi Bars

Cluster 3 - Buffets, Thai, Vegetarian, Gluten-Free, Indian, Vegan

Cluster 4 - American (Traditional), Sports Bars, Chicken Wings

Cluster 5 - Italian, Pizza

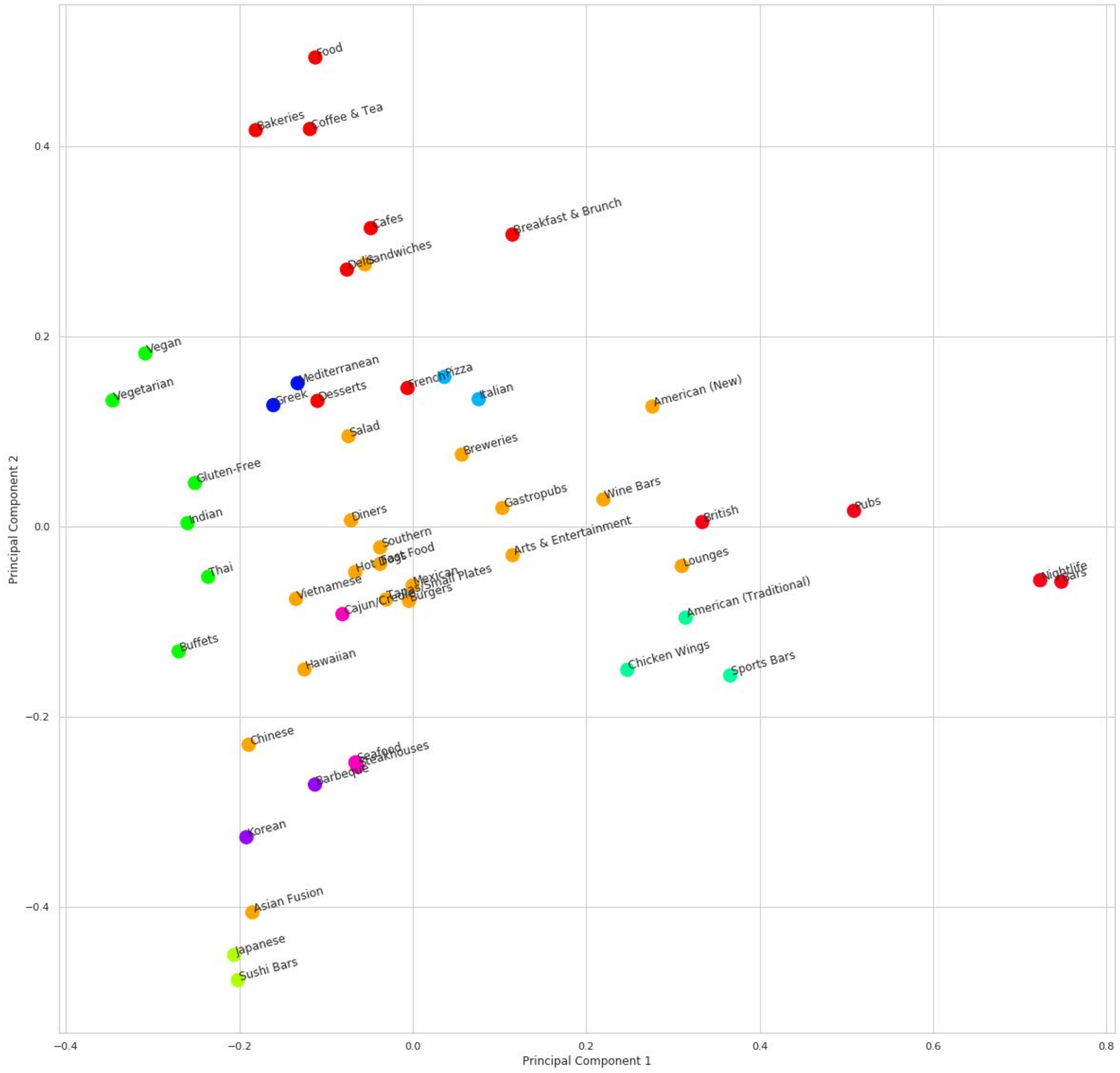
Cluster 6 - Mediterranean, Greek

Cluster 7 - Barbeque, Korean

Cluster 8 - Steakhouses, Seafood, Cajun/Creole

Cluster 9 - Nightlife, Bars, Pubs, British

But plotting it using the same method as before, we can see a clearly better visual separation:



Conclusion

This second project was also very interesting and challenging. It's difficult to visually compare so many objects. The similarity matrix proved to be very useful to make a first contact with the similarities and dissimilarities among all the cuisines. The LDA similarity matrix helped filtering some noise and showing the relations more clearly.

The clustering models applied were very simple and fast to run. The most difficult part was choosing the number of clusters. The results were quite impressive, showing very good relationships and grouping together similar cuisines.