Battle of neighborhoods

Mhatipog

IBM Data science capstone project

Clusters in the world cuisine

Coursera

Table of Contents

Abstract

In this work, we have used the technique of KNN to cluster cuisines based on the similarity of the ingredients typically used. Even though geographical coordinates were not used in the clustering, the resulting clusters, when shown on a map appeared mostly as geographically coherent, with understandable borders and mostly understandable exceptions. We found 7 cuisines, which we named the Indian Ocean cuisine, the Western cuisine, the (East-) Asian cuisine, the Latin cuisine, the Arabic cuisine, the Mediterranean cuisine, and the Bangladeshi cuisine. The largest surprise was to see the Bangladeshi cuisine appear as a cluster on its own. This was also the case when the number of clusters was imposed to be 6. Therefore, this project suggests to the touristic sector and to tourists that Bangladeshi cuisine is unique. How unique? Tasting the Bangladeshi cuisine is like tasting 1/7th cuisines of the world.

Keywords: World cuisine, data science

Clusters in the world cuisine

Battle of neighborhoods

Introduction/Business Problem

Before the COVID-19 pandemic, tourism was a booming sector across the globe. Also, after the pandemic, it is expected to recover. Because people are curious about seeing places, doing new things, experiencing different cultures, and finally, trying out different local cuisines.

The search of new tastes drives people to tourism and the touristic sector, for example the marketing department of a tourism agency can think of mapping out the similarities/differences between the different local cuisines in the world. Which cuisine is truly unique? Which countries have similar cuisines? The answers to these questions can form the basis of a marketing argument and is already interesting on its own.

Therefore, the primary target audience of this project are the companies and government agencies with an interest of having a great restart of the touristic sector after the pandemic. In the second place, the project also targets the attention of people seeking to try new cuisines, as a help to guide them towards a destination with tastes they have not tasted before.

Methods/Data

To tackle the above-mentioned problem, I will use the following data.

Clusters in the world cuisine

Data of recipes

A researcher named Yong
Yeol Ahn scraped tens of thousands of food recipes (cuisines and ingredien
ts) from three different websites. For more information on Yong-
Yeol Ahn and his research, I refer to Flavor Network and the Principles of
 Food Pairing[i].
I will be using the version of the data hosted on the IBM server[ii].
It is a dataset of 57,691 recipes, with data on the cuisine, as well as wh
ether 384 ingredients exist in the recipe or not.

Foursquare API

Foursquare API will be for two purposed,
geocoding the coordinates of

* Obtaining the coordinates of the said cuisine, and
* Exploring the restaurants in the location of interest

Data preparation

A few basic operations have been performed on the data. The most notable one is the manual correcting of the cuisine names to give them a consistent country tag. In this process, some general tags are specified. For example, Cajun_Creole is assigned to New Orleans, and African is assigned to Congo. A total of 43 entries have been manually corrected. For the whole list of corrections, please see the notebook.

Then, the data is casted in the dataframe with the format of interest, namely in the format where each country is assigned a culinary profile based on the percentage for each possible ingredient that they use in their dishes, as in figure 1 below.

SPAIN
olive_oil (61%) garlic (56%) onion (46%) bell_pepper (35%)

SWEDEN
wheat (74%) butter (70%) egg (59%) cream (27%)

SWITZERLAND
butter (70%) wheat (65%) egg (45%) pepper (30%)

TANZANIA
onion (63%) ginger (36%) garlic (36%) cumin (27%)

TEXAS
butter (57%) wheat (48%) egg (41%) corn (29%)

THAILAND
garlic (59%) fish (52%) cayenne (47%) cilantro (41%)

TURKEY
onion (62%) garlic (62%) tomato (43%) bell_pepper (37%)

UNITED KINGDOM
wheat (60%) butter (59%) egg (48%) milk (38%)

UNITED STATES
butter (41%) egg (40%) wheat (39%) onion (29%)

VENEZUELA
garlic (56%) onion (54%) cayenne (51%) tomato (41%)

VIETNAM
fish (73%) garlic (72%) rice (49%) cayenne (43%)

Figure 1 Examples of culinary profiles

Results

KNN Clustering

The KNN clustering algorithm is implemented to cluster these 69 cuisines into several groups by the similarity in their ingredient profiles.

As the KNN algorithm requires to specify the number of clusters beforehand, we need first to answer this question. For this purpose, the elbow method is used on the result shown in Figure 2.

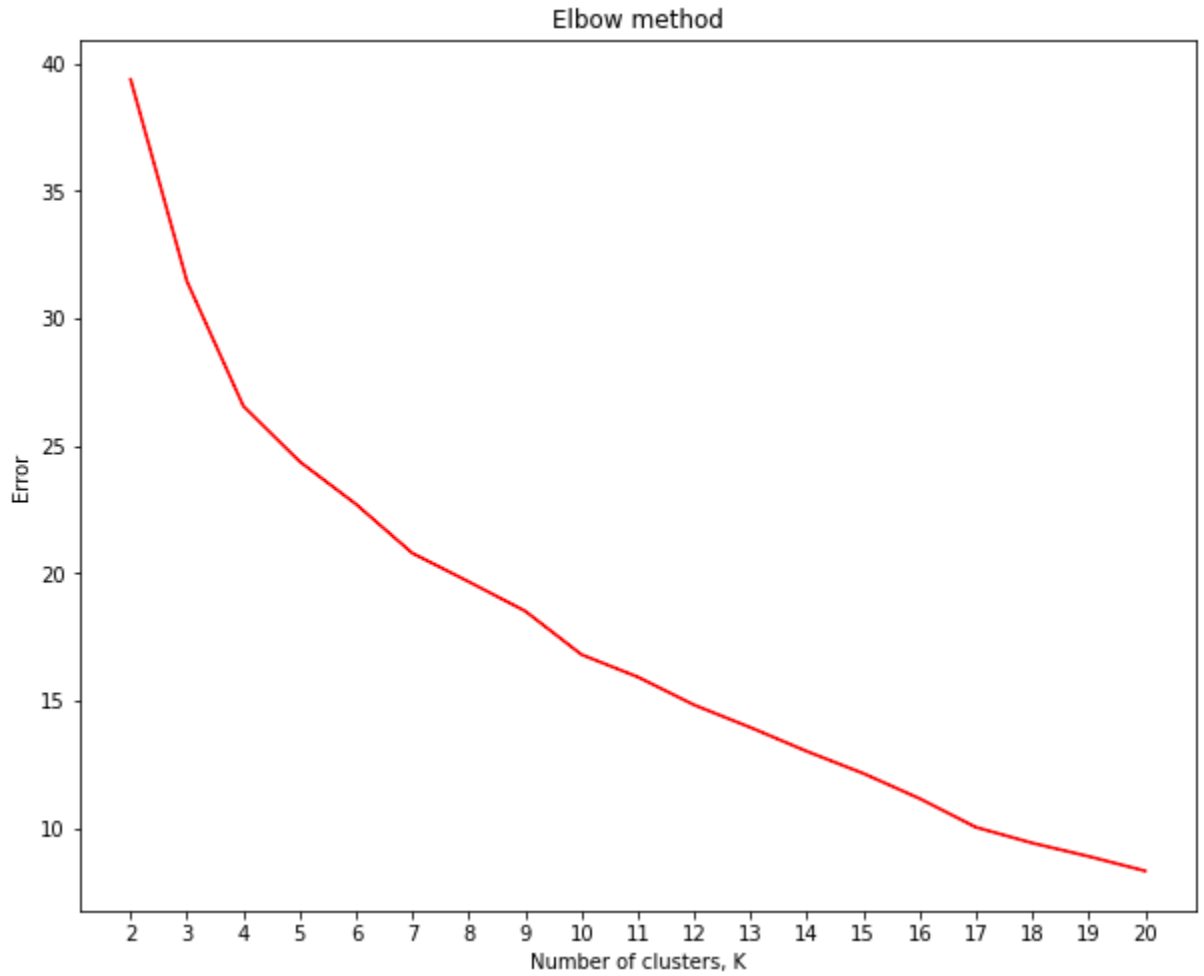

Figure 2 Elbow method

The high numbers on the vertical axis are expected, as we are comparing 384 ingredients. It seems that the elbow cannot be unambiguously defined visually. Still a good candidate is 7 for the following reasons,

- While at 7 we do not have a clear elbow, the decrease for k>7 is less than linear.
- The error seems reasonable (on its scale), and more importantly
- We aim to have a relatable picture for tourists. It is conceptually nice to divide the world map in seven pieces, as there are for example also (usually) 7 continents.

Clusters on world map

The clusters can be adequately represented on a world map, as shown in Fig. 3.

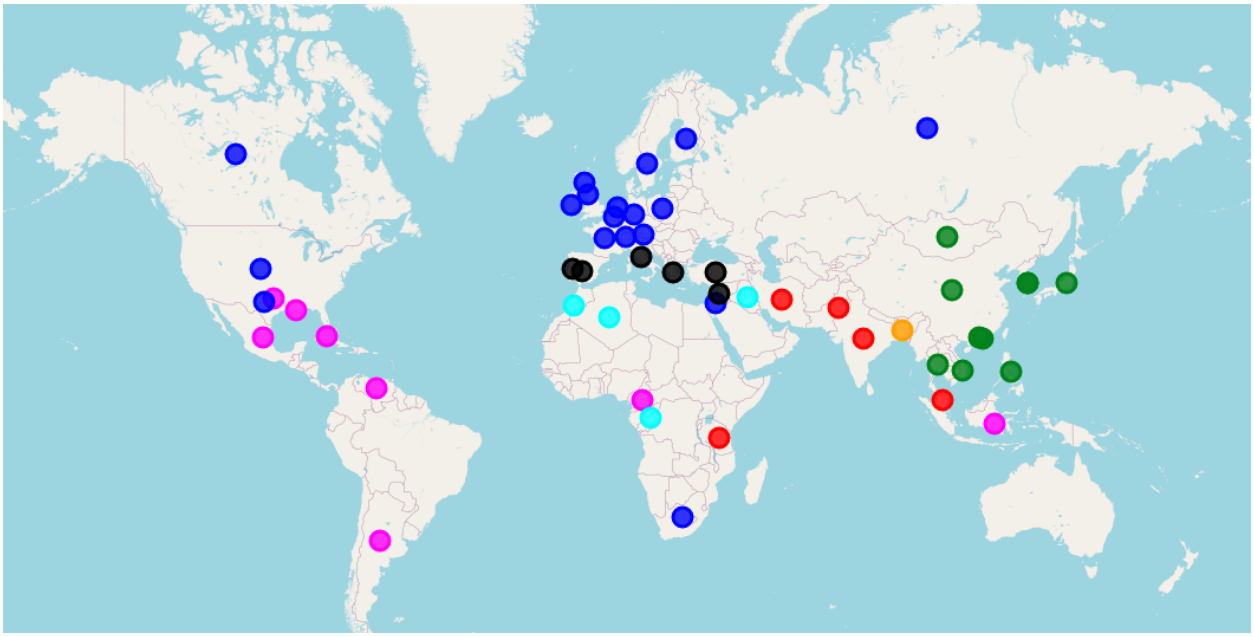

Figure 3 World map of culinary clusters

This interesting result will be discussed in the discussions section. But we already remark that the Bangladeshi cuisine is a cluster on its own. Figure 4 shows the number of countries each cluster is accommodating. They are 18, 10, 6, 5, 4, 3 and 1 for respectively the blue, green, magenta, green, black-, red-, cyan- and orange-colored clusters.

Clusters in the world cuisine

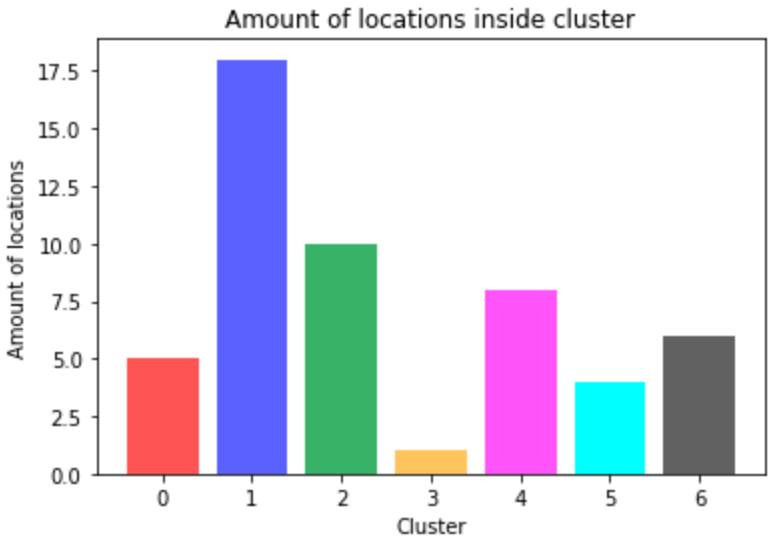

Figure 4 Number of countries in cluster

Bangladeshi restaurants in Tokyo

We take a side-step to ask the following question: why do we need to go to Bangladesh to taste its cuisine while there are restaurants of all cuisines everywhere?

The answer is: because there are not. Restaurants serving Bangladeshi food seem to be rare, even in very cosmopolitan cities, such as London, New York City, Amsterdam, and Los Angeles. For example, in Figure 5, we see the Bangladeshi restaurants in world's most populous city, Tokyo. There are 7 of them, separated from each other with distances which take over an hour (except the three which are a bit closer together).

Clusters in the world cuisine

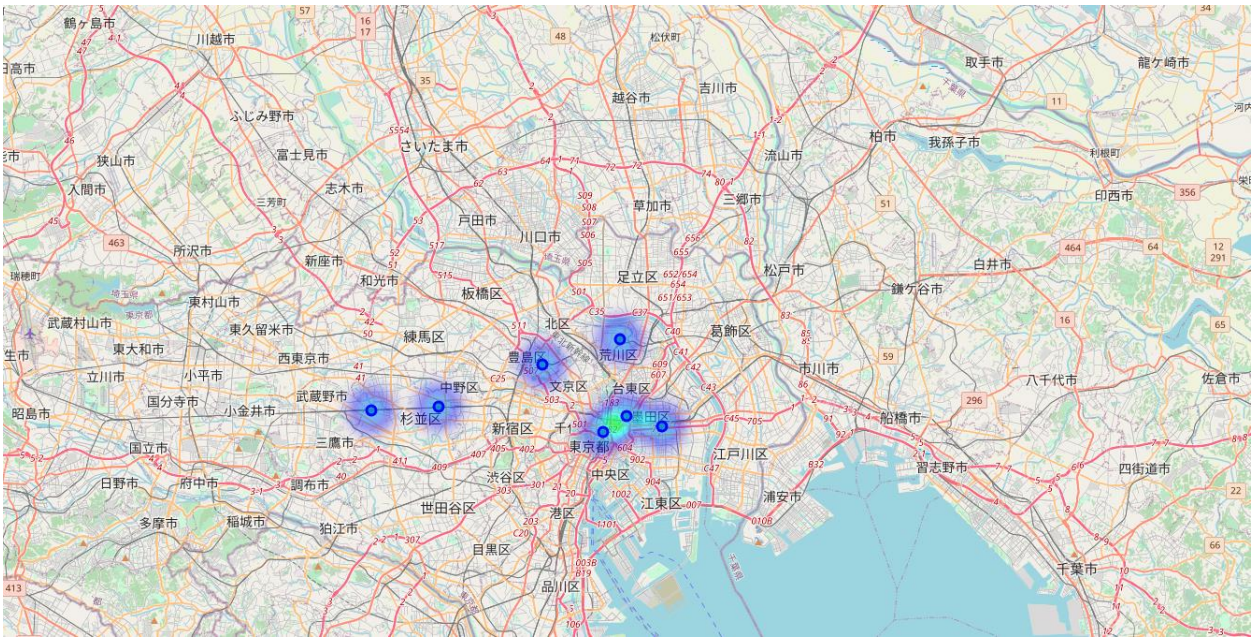

Figure 5 Bangladeshi restaurants in Tokyo

Discussion

For the clustering, can see that the world map of cuisines can be consistently grouped by similarity into seven clusters. Even though the geographical proximity was not fed into the algorithm, the result that came out clearly showed that the cuisines are mostly geographically grouped. As regions which emerge out of this grouping are huge, even continent-sized, this fact cannot be explained by mere similarities in natural geography and vegetation, but rather by a combination of it with cultural influences (which in its turn is also largely influenced by geography).

The following clusters appeared:

1. The **Indian Ocean cuisine** consists of India, Pakistan, Iran but also Malaysia and Tanzania. While the grouping of the first three countries is expected, the latter two country also use similar ingredients, which I was not really expecting.
2. The **Western cuisine** is the largest group, probably because the website at the origin of the dataset made the finest distinction between western countries. They mostly form a geographical cluster, while the exceptions are expected. These exceptions are South-Africa and Israel, which have been receiving an important number of settlers from Western countries.
3. The **(East-) Asian cuisine** is strictly geographically coherent. It consists of the cuisines of Japan, China, Korea. But also, countries like Vietnam and the Philippines share the main ingredients.

4. The **Bangladeshi cuisine** is a unique one. This cuisine ended up as its own cluster. Even when k was selected to be 6, this cuisine was still a cluster on its own. Geographically, it is at the border of two clusters, namely the Indian Ocean cuisine and the East-Asian cuisine. Therefore, it is most likely influenced by both.
5. The **Latin cuisine** is the grouping of the cuisines mainly of Latin-American countries, but surprisingly it seems that Indonesia and Eastern Africa also share many ingredients. While this grouping could be coincidental, it is interesting to note that geographically, we can collect them all in a close curve which does not intersect any other cuisines.
6. The **Arabic/African cuisine** groups Northern Africa and Iraq. The dot placed on Congo in fact correspond to recipes labeled as 'African' and is the dot with the least accurate placing. So, take that dot with a grain of salt.
7. The **Mediterranean cuisine** shows the culinary binding effect of the Mediterranean Sea. Most of the countries around this sea, from Spain, to Italy, to Turkey, mostly eat very similar ingredients. The only country in it which has no Mediterranean coast, is Portugal.

KNN Clustering

Conclusion and outlook

In this work, we have used the technique of KNN to cluster cuisines based on the similarity of the ingredients typically used. Even though geographical coordinates were not used in the clustering, the resulting clusters, when shown on a map appeared mostly as geographically coherent, with understandable borders and mostly understandable exceptions.

We found 7 cuisines, which we named the Indian Ocean cuisine, the Western cuisine, the (East-) Asian cuisine, the Latin cuisine, the Arabic cuisine, the Mediterranean cuisine, and the Bangladeshi cuisine.

The largest surprise was to see the Bangladeshi cuisine appear as as a cluster on its own. This was also the case when the number of clusters was imposed to be 6.

Therefore, this project suggests to the touristic sector and to tourists that Bangladeshi cuisine is unique. How unique? Tasting the Bangladeshi cuisine is like tasting 1/7th cuisines of the world.

Obviously, this project is limited in scope and the results need to be taken with a lot of salt. To improve its quality, data sources should be extended in size, special care needs to be taken that they contain no bias with respect to our questions, and different clustering techniques need to be compared to check that the result does not depend on the arbitrary choice of it.

---

[i] http://yongyeol.com/papers/ahn-flavornet-2011.pdf

[ii] https://s3-api.us-geo.objectstorage.softlayer.net/cf-courses-data/CognitiveClass/DS0103EN/labs/data/recipes.csv