Recipe Analysis

Gaurav Burman, Mayank Jaggi, Sumeet Shinde

Abstract- The goal is to predict the cuisine of recipes based on its ingredients. Unique ingredients to a cuisine and support vector classifier algorithm are used for prediction. The relationship between geographic proximity of cuisines and common ingredients used is determined. A recommender system is developed, recommending a cuisine to the user based on his/her ingredient preference.

Keywords- Correlation, K-means, PCA, Support Vector Classifier, Naive Bayes

I. INTRODUCTION

Food is not only a necessity but also reflects the culture of the respective society. In this project, we attempt to understand the similarities and differences of various cuisines through the ingredients that go into a particular recipe of that cuisine and what they represent from the perspective of different cultures around the world.

The dataset is a JSON file containing recipe id's, cuisine and ingredients with a total of 39774 rows. Figure 1 gives a snapshot of the first 5 rows of the dataset. The recipes are spread across 20 cuisines which are listed in Figure 2 below.[1]

	id	cuisine	ingredients
0	10259	greek	[romaine lettuce, black olives, grape tomatoes, garlic, pepper, purple onion, seasoning, garbanzo beans, feta cheese crumbles]
1	25693	southern_us	[plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn meal, milk, vegetable oil]
2	20130	filipino	[eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, soy sauce, butter, chicken livers]
3	22213	indian	[water, vegetable oil, wheat, salt]
4	13162	indian	[black pepper, shallots, cornflour, cayenne pepper, onions, garlic paste, milk, butter, salt, lemon juice, water, chili powder, passata, oil, ground cumin, boneless chicken skinless thigh, garam masala, double cream, natural yogurt, bay leaf]

Figure 1. Snapshot of Recipe Dataset

Sr No	Cuisine	Sr No	Cuisine
1	brazilian	11	jamaican
2	british	12	japanese
3	cajun_creole	13	korean
4	chinese	14	mexican
5	filipino	15	moroccan
6	french	16	russian
7	greek	17	southern_us
8	indian	18	spanish
9	irish	19	thai
10	italian	20	vietnamese

Figure 2. Cuisines in the Recipe Dataset

II. CUISINE CLUSTERING

K-means clustering is an elegant unsupervised learning algorithm that partitions a data set into K distinct, non-overlapping clusters. We group the cuisines having similar ingredients using K-means and PCA (Principal Component Analysis) producing 3 clusters as shown in Figure 3. This is implemented by using tf-idf (Term Frequency-Inverse Document Frequency) followed by reducing the dimensions to 2 using PCA. This reduced data is fed into the K-means algorithm.[2]

The results obtained using 2 Principal Components explains only 39% of the total variance in the data. This information is not sufficient to make any reliable inferences. We note that by increasing the number of Principal Components to 5, we are able to explain only 62% of the variance. However, an increase in the dimensionality of the reduces the results' interpretability making the clustering technique less effective.

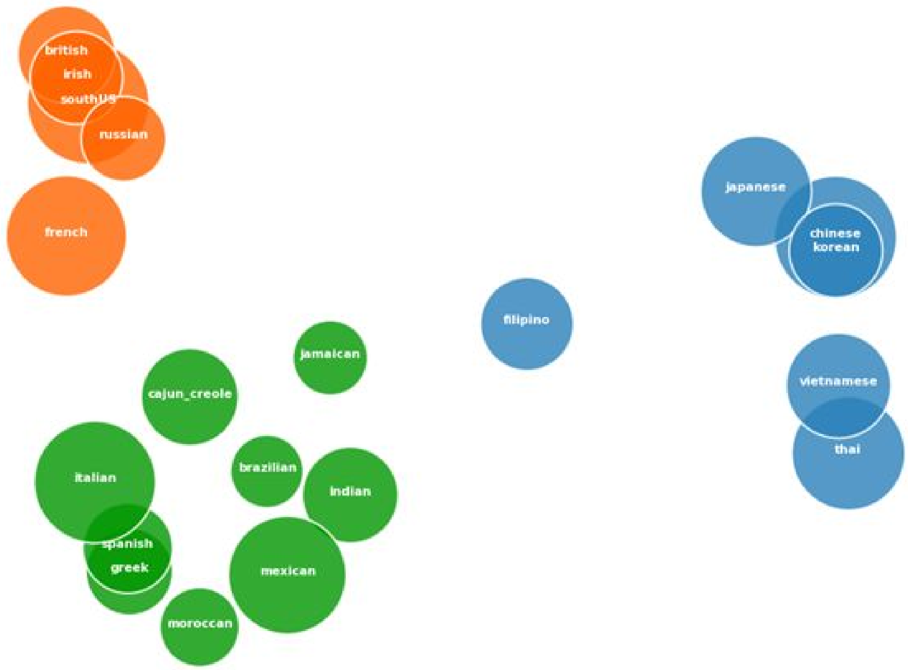

Figure 3. Cuisines grouped into 3 clusters using K-means

III. EXPLORATORY DATA ANALYSIS

As mentioned earlier, the dataset has 39774 recipes distributed across 20 cuisines as shown in Figure 4. To determine the correlation between cuisines, we use the count of ingredients shared in the recipes between two cuisines. The higher the count of common ingredients, the stronger is the relation between the two cuisines. The correlation between cuisines is summarized in Figure 5. For Brazilian cuisine, the top three cuisines that share ingredients are Mexican, Italian, and Southern US. Similarly, it is determined for the remaining 19 cuisines. The information of all the pairs of cuisines (60 pairs) can be seen in Food_Analysis Python Notebook. We calculated the correlation between cuisines based on its geographical proximity by using the latitude and longitude coordinates of the country of origin of the cuisine. The distance between the two cuisines location is calculated and used as a parameter to determine the relation between cuisines. Smaller the distances higher is the geographic proximity. The relations are summarized in Figure 6. The proximate pairs of cuisines can be found in Food_Analysis Python Notebook. Additionally, we determined whether geographic proximate pairs of cuisine share similar ingredients. Using two data frame which lists 60 pairs each (one associated with geographic proximity and other with common ingredients), we calculate the common pairs. The results are summarized in Figure 7. Around 22 pairs exist in both the data frames.[3]

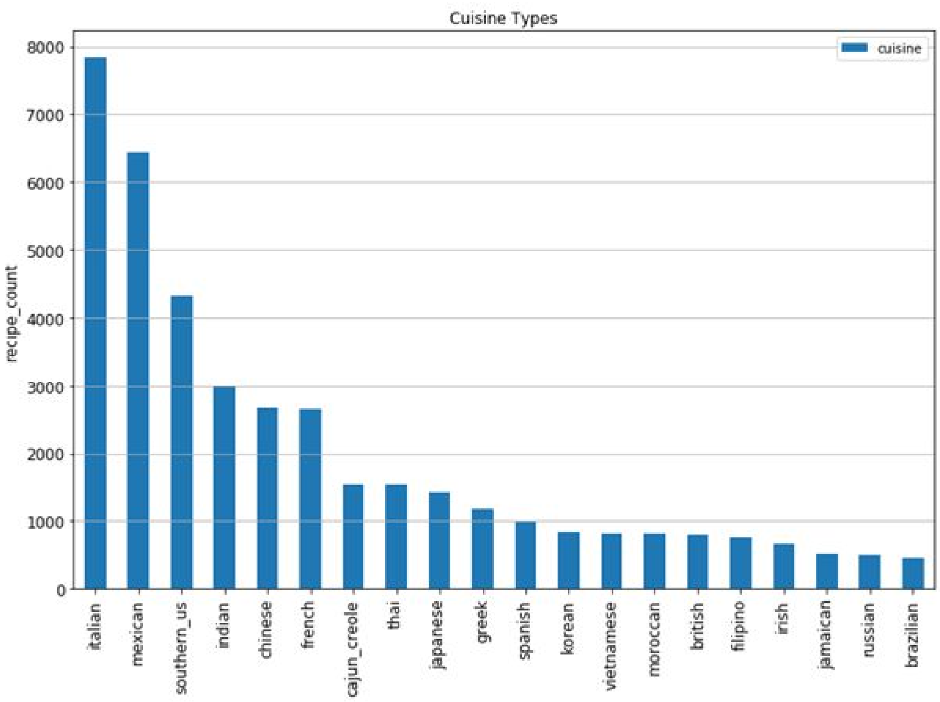

Figure 4. Distribution of recipes across cuisines

	brazilian	british	cajun_creole	chinese	filipino	french	greek	indian	irish	italian	jamaican	japanese	korean	mexican	moroccan	russian	southern_us	spanish	thai	vietnamese
brazilian	852	432	527	524	414	582	433	533	402	649	440	470	352	663	418	374	641	511	486	412
british	432	1165	586	590	401	812	522	621	584	846	443	515	355	749	482	515	819	572	465	395
cajun_creole	527	586	1573	774	532	953	695	751	571	1178	578	655	455	1121	567	535	1140	756	674	547
chinese	524	590	774	1791	655	880	626	879	524	1046	563	903	670	1062	601	508	996	659	949	815
filipino	414	401	532	655	947	561	417	570	360	636	438	565	440	663	397	369	642	464	578	532
french	582	812	953	880	561	2095	843	893	736	1535	585	761	518	1242	717	663	1286	936	745	613
greek	433	522	695	626	417	843	1196	717	493	989	457	547	402	882	618	486	836	668	575	457
indian	533	621	751	879	570	893	717	1661	564	1058	599	809	506	1054	701	541	1003	707	842	650
irish	402	584	571	524	360	736	493	564	996	770	416	461	328	712	453	466	764	533	434	353
italian	649	846	1178	1046	636	1535	989	1058	770	2921	662	872	585	1626	787	692	1548	1039	878	704
jamaican	440	443	578	563	438	585	457	599	416	662	877	499	365	707	435	385	684	519	508	429
japanese	470	515	655	903	565	761	547	809	461	872	499	1439	595	852	515	469	810	574	791	668
korean	352	355	455	670	440	518	402	506	328	585	365	595	898	597	371	340	557	422	567	520
mexican	663	749	1121	1062	663	1242	882	1054	712	1626	707	852	597	2675	741	657	1521	953	914	741
moroccan	418	482	567	601	397	717	618	701	453	787	435	515	371	741	974	429	703	601	549	445
russian	374	515	535	508	369	663	486	541	466	692	385	469	340	657	429	872	643	507	417	375
southern_us	641	819	1140	996	642	1286	836	1003	764	1548	684	810	557	1521	703	643	2452	911	817	661
spanish	511	572	756	659	464	936	668	707	533	1039	519	574	422	953	601	507	911	1262	588	486
thai	486	465	674	949	578	745	575	842	434	878	508	791	567	914	549	417	817	588	1376	761
vietnamese	412	395	547	815	532	613	457	650	353	704	429	668	520	741	445	375	661	486	761	1108

Figure 5. Correlation between cuisines based on the count of common ingredients

	brazilian	british	cajun_creole	chinese	filipino	french	greek	indian	irish	italian	jamaican	japanese	korean	mexican	moroccan	russian	southern_us	spanish	thai	vietnamese
brazilian	0	5766	5077	10763	11255	5698	6042	8746	5714	5717	3626	11538	11269	4773	4603	7175	5024	5060	9982	10454
british	5766	0	4644	5057	6672	213	1486	4170	288	891	4682	5940	5504	5549	1253	1554	4849	784	5923	5739
cajun_creole	5077	4644	0	7074	8455	4821	6124	8172	4356	5489	1453	6593	6933	963	4853	5699	213	4812	9105	8518
chinese	10763	5057	7074	0	1773	5104	4731	2347	5145	5048	8344	1301	592	7741	6176	3599	7209	5728	2049	1446
filipino	11255	6672	8455	1773	0	6677	5986	2955	6805	6458	9879	1862	1628	8837	7638	5132	8532	7243	1373	1089
french	5698	213	4821	5104	6677	0	1303	4093	485	687	4801	6035	5571	5715	1128	1545	5025	654	5868	5715
greek	6042	1486	6124	4731	5986	1303	0	3112	1774	653	6008	5906	5291	7010	1722	1386	6327	1472	4923	4922
indian	8746	4170	8172	2347	2955	4093	3112	0	4398	3677	8777	3627	2910	9108	4830	2698	8374	4519	1812	1866
irish	5714	288	4356	5145	6805	485	1774	4398	0	1172	4427	5956	5563	5264	1336	1737	4562	901	6126	5910
italian	5717	891	5489	5048	6458	687	653	3677	1172	0	5360	6124	5572	6364	1181	1477	5690	847	5487	5426
jamaican	3626	4682	1453	8344	9879	4801	6008	8777	4427	5360	0	8024	8299	1463	4402	6082	1423	4543	10235	9734
japanese	11538	5940	6593	1301	1862	6035	5906	3627	5956	6124	8024	0	718	7026	7158	4647	6672	6687	2860	2278
korean	11269	5504	6933	592	1628	5571	5291	2910	5563	5572	8299	718	0	7491	6670	4105	7042	6211	2311	1701
mexican	4773	5549	963	7741	8837	5715	7010	9108	5264	6364	1463	7026	7491	0	5604	6661	752	5632	9784	9171
moroccan	4603	1253	4853	6176	7638	1128	1722	4830	1336	1181	4402	7158	6670	5604	0	2577	5029	475	6642	6606
russian	7175	1554	5699	3599	5132	1545	1386	2698	1737	1477	6082	4647	4105	6661	2577	0	5912	2137	4390	4186
southern_us	5024	4849	213	7209	8532	5025	6327	8374	4562	5690	1423	6672	7042	752	5029	5912	0	5002	9250	8655
spanish	5060	784	4812	5728	7243	654	1472	4519	901	847	4543	6687	6211	5632	475	2137	5002	0	6326	6239
thai	9982	5923	9105	2049	1373	5868	4923	1812	6126	5487	10235	2860	2311	9784	6642	4390	9250	6326	0	614
vietnamese	10454	5739	8518	1446	1089	5715	4922	1866	5910	5426	9734	2278	1701	9171	6606	4186	8655	6239	614	0

Figure 6. Correlation between cuisines based on geographic proximity

Sr No	Cuisine	Similar_Cuisine	Nearby_Cuisine
1	brazilian	mexican	mexican
2	british	french	french
3	cajun_creole	southern_us	southern_us
4	cajun_creole	mexican	mexican
5	greek	italian	italian
6	greek	french	french
7	irish	french	french
8	italian	french	french
9	jamaican	mexican	mexican
10	jamaican	southern_us	southern_us
11	japanese	chinese	chinese
12	korean	chinese	chinese
13	korean	japanese	japanese
14	mexican	southern_us	southern_us
15	moroccan	italian	italian
16	moroccan	french	french
17	russian	italian	italian
18	russian	french	french
19	southern_us	mexican	mexican
20	spanish	french	french
21	vietnamese	chinese	chinese
22	vietnamese	thai	thai

Figure 7. Pairs of Cuisine- Highly correlated based on geographic proximity and common ingredients

In our analysis, we notice that there are quite a few ingredients that are shared amongst multiple cuisines (Figure 8) and a few that are unique to a cuisine. So, we use these unique ingredients to predict the cuisines for our test dataset (9944 observations). We predicted cuisines for only 1144 recipes out of 9944 recipes in the test data. The number of predictions is low because ingredients used in the recipes in the test data are not unique to a cuisine.

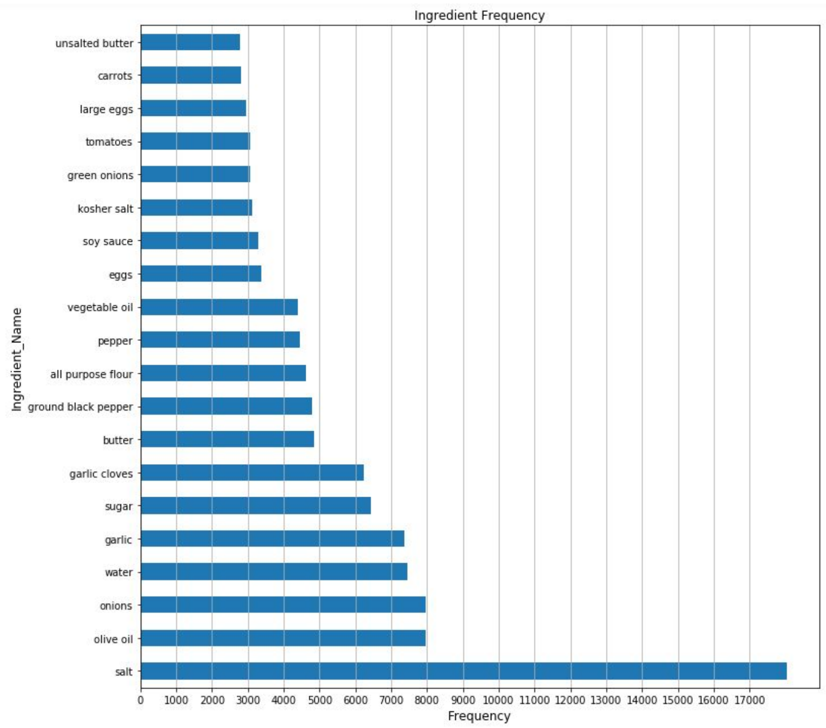

Figure 8. Frequency of top 20 ingredients across cuisine

IV. MODELING AND PREDICTION

To get better results than earlier, we use the Linear Support Vector Classifier algorithm. The algorithm builds a hyperplane (a line in two dimensions) classifying the dataset (here into different cuisines). Using this algorithm, we get a prediction accuracy of 80% (score calculated on kaggle) on our test data. The predicted cuisine for the recipes in the dataset in the submission CSV file.

V. RECOMMENDER SYSTEM

The traditional recommendation system requires some kind of ratings to predict what the user might prefer. Here, we don't have such a rating. Therefore, we created our own rating as follows:

- Find the highest occurring 5 ingredients in a cuisine.
- Rate the ingredients from range 1 to 5 with the most frequent ingredient as 5.
- Make a data frame such that the first column is the list of cuisines and the rest are the ingredients.

We can make a recommendation by finding the probability of a person liking a cuisine by taking his/her inputs as the rating for the ingredients.

We decided to implement the Naive Bayes Algorithm. The Algorithm does not need a large data set to predict as it is based on Bayes Theorem and it is simply calculating the probability of a person liking a cuisine given his/her preference for ingredients.

While implementing we considered only 3 cuisines and their top 5 ingredients. We want to demonstrate that such a technique will work as a recommender system without complicating and compromising the interpretability of the model. The developed model is scalable vertically (more cuisines) as well as horizontally (considering more ingredients per cuisines).

VI. CONCLUSION

We analyzed the Recipe dataset and grouped the 20 cuisines into 3 clusters and observed the relation between cuisines based on the common ingredients and its geographic proximity. Our findings suggested that there is not a strong correlation between geographic proximity of the cuisine and the common ingredients used as only 22 pairs amongst 120 pairs (not distinct) of cuisine were common in both methods. We predicted the cuisine of recipes based on its ingredients using unique ingredients technique and Linear Support Vector classifier algorithm with the latter giving a prediction accuracy of 80%. Finally, we developed a recommender system, recommending the cuisine to a user based on the ingredients he/she likes.

REFERENCES

[1] Recipe Dataset by Yummly. Available at: https://www.kaggle.com/c/whats-cooking/data
[2] An Introduction to Statistical Learning with Applications in R by Gareth James, Robert Tibshirani.
[3] Food_Analysis Python Notebook (Attached with the report)