Analyzing the similarities of World Food by Using Online Recipes

Sahithya Sridhar sahsrid@iu.edu

Prajakta Patil patilpr@iu.edu

Nishad Tupe ntupe@iu.edu

Indiana University, Bloomington

Abstract

People around the world are passionate about their food culture and they tend to think their food is most unique. Food and nutrition data provide an invaluable information about the culinary culture around the world and shows strong similarity between the taste preferences. Using a database of over 9000 recipes and more than 20 cuisines we analyze ingredients and their nutritional values and use this knowledge to assess the predictability of recipes from different cuisines. We also determine how the food of different countries far separated geographically are similar. By performing different analyses on the food data set of different countries, this project studies and shows similarities between the different food cultures by studying the top ingredients, and the common ingredients used in different cuisines and their nutritional values.

Keywords

World-Cuisine, Ingredients, Nutritional Value, Recipe, Pandas, Scikit-Learn.

1. Introduction

Food is most basic human need. People developed different food based on their geographical locations and ingredients that were available to them locally. World is now becoming a global village as more and more people can travel around the world easily. With so many people travelling around the world and with so many cooking channels trying new dishes, and with the general acceptability to try different cuisines, the passion for food around the world has increased. Food is becoming global too and ingredients are being used intercontinentally. If work is done to study, visualize and show similarities between different food cultures; it can help people feel how they are connected. Some work has been done in network science to visualize most used ingredients in 57k recipes divided in different categories such as vegetables, dairy, cereal etc. (Ahn et al, 2011). In addition, visualization work is also done to show most common ingredients among different regions of world. In our project we have tried to use these ideas, search most famous recipes from different parts of world and visualize common ingredients among them. We also tried to visualize eating habits around world and find which diet is most nutritious. This visualization may help people make small changes to their diets so that they can benefit from it.

2 Literature Review:

Recipe recommendation has gained tremendous commercial importance in the recent years. Data scientist in company's like Hello Fresh, Blue Ribbon delves into algorithms for suggesting recipes to their customers based on recipe rating and cooking history. This project focuses on finding subtle similarities in the recipes based on overlapping ingredients by their appearance in various international cuisine.

Today's world food network analysis has been revolutionary from macro to micro level analysis. In past research, (Ahn et al, 2011) found how recipes around world vary depending on flavor compounds shared by ingredients. (Stefnar, 2016) Analyzed German startup mymuesli.com data and found interesting flavors combination on how their customer combined the ingredients they offer through custom mix muesli per order. Stefnar discovered ingredient grouping using the radial network visualization that depicts the relationship of mix of fruits, nuts, and seeds with muesli. With his analysis, he found that fruits are most popular ingredient among other combinations. (Teng et al, 2012) capture relationship between ingredients by establishing a complement network and a substitute network. In another study of Macedonian recipes of south European cuisine (Bogojeska et al, 2016) found strong influences of Middle Eastern and Eastern European cuisines. The project delivers a criterion which shows the contribution of specific ingredient.

In this project, we have tried to find common ingredients shared by different cuisine around world and see if there are certain communities of countries that share similar ingredients to classify them as a group. This is biggest contribution of our work, as we also performed simple analysis on nutritional value of different cuisine.

3. Data Collection and processing

We collected data for our project from Kaggle website which is one of well-known site for finding datasets related to different topics. We downloaded Yummly dataset from this website. This dataset has list of ingredients of different cuisines from around world and id for those ingredients. We got data for over 9000 recipes for 20 countries. Our analysis is done on this full dataset but, we also took 25 random recipes from each country to see how their cuisine breaks into vegetarian and non-vegetarian dishes.
Our dataset consists of:

Ingredients: The ingredients always did not appear in the same way. There where ingredients that are written in different formats and terminology. E. g. salt, coarse salt.

Cuisine: We used the train-test method to determine the cuisine name from the ingredients present.

Nutrition: As nutrition data was not available in the original dataset, we used data from USDA website to get the nutritional value for the ingredients present.

We used above data in .csv, .json, .pajek formats to analyze and visualize it in multiple ways.

Initial approaches involved the use of NLTK library and we cleaned raw data to remove Punctuation, remove digits, change everything to lower case, stem the ingredients, Lemmatization, remove white space, remove unnecessary features such as %, \, /, &, (,), ., ”, \\.

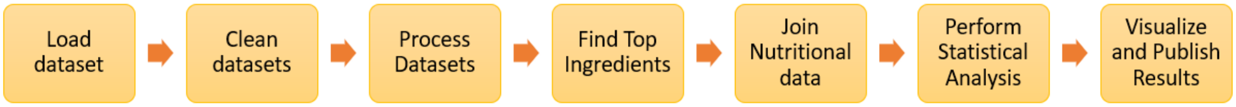

Load dataset → Clean datasets → Process Datasets → Find Top Ingredients → Join Nutritional data → Perform Statistical Analysis → Visualize and Publish Results

Image shows general flow of our work.

4. Methods

We wanted to see most common ingredients, eating patterns among different cuisines. First, we characterized different cuisines around the world by their ingredients. Then, we trained a SVM classifier and used deep learning models to predict a cuisine from its ingredients. This enabled us to discover the similarity across different cuisines based on their ingredients.

We applied following methods to help make our observations and hence draw conclusions from them.

1. Centrality measurements
   a) Degree centrality
   b) Closeness centrality
2. Community detection
3. Count, Term Frequency Inverse Document Frequency (tfidf) vectorizer, and Linear SVC
4. Logistic Regression, Multinomial Naïve Bayes and Random Forest algorithm
5. Hypothesis testing on degree assortivity of network
6. Statistical analysis on nutritional value of cuisines

5. Results

Applying degree centrality on network helps us find relative use number of ingredients in different cuisine. Countries with more ingredients show up as large nodes and those with relatively less ingredients show up as smaller nodes. As seen in Fig. 1., Italy had most number of ingredients and hence is the largest node. While Brazil, Russia and Jamaica are some of the smallest nodes.

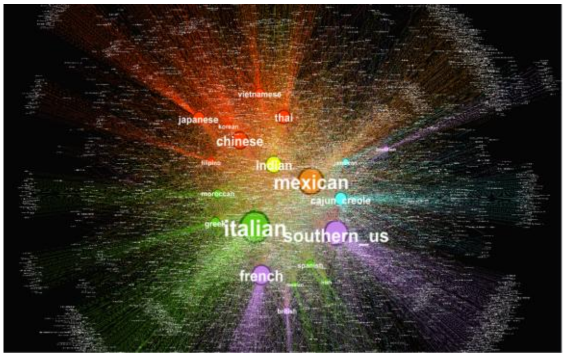


Fig. 1. Cuisine network based on ingredients

Closeness centrality helped us show most common ingredients among all countries. Below image is taken by applying closeness centrality to random 25 recipes from data. As we can see here, all countries use same basic ingredients such as salt, olive oil, black-pepper, sugar, onions, garlic, egg, butter etc. We see same most common ingredients in full data as well.

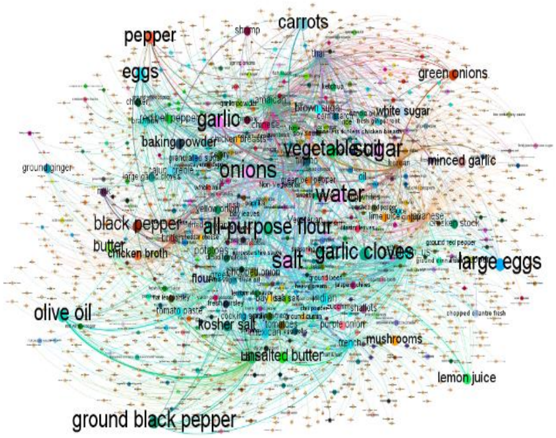


Fig. 2. Most common ingredients in world cuisine

In addition to centralities, community detection on same data helped us show countries that share most ingredients in same color and hence in same community. From this emerged 6 distinct group of countries that share lot of ingredients. As seen in Fig. 1. these groups are,

1. Italy-Greece-Morocco-Spain-Ireland-Russia (Green)
2. Britain-France-Southern US-Brazil (Purple)
3. China-Thailand-Vietnam-Japan-Korea-Philippines (Red)
4. Cajun and Jamaican (Sky blue)
5. Mexico (Orange)
6. India (Yellow)

Calculating edge weight for ingredients on limited data of 25 random recipes per country shows, top 3 cuisines with non-vegetarian food are France, Britain and Thailand. While top 3 vegetarian cuisines are India, Brazil and Greece.

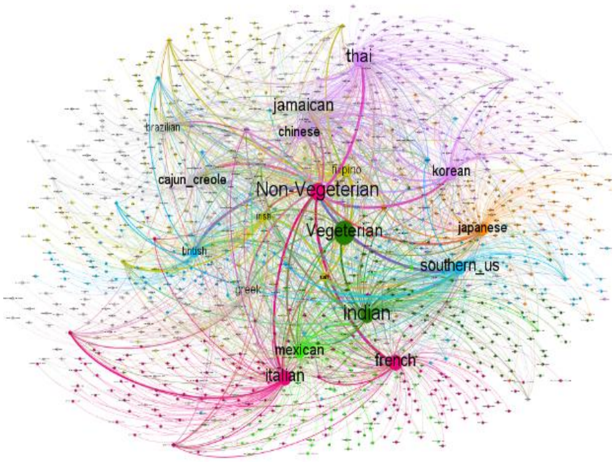


Fig. 3. Vegetarian vs Non-vegetarian diet

Applying count vectorizers on data shows Italian cuisine had most number of ingredients for roughly 2000 dishes while India was at 4th place for number of ingredients for roughly 800 dishes.

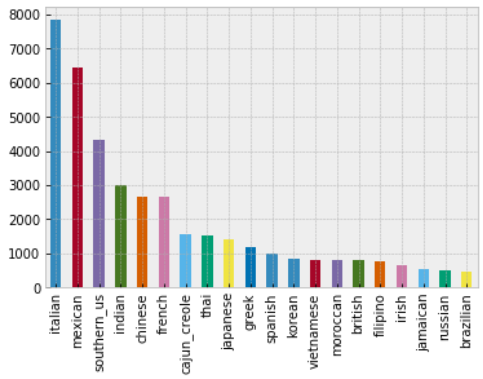
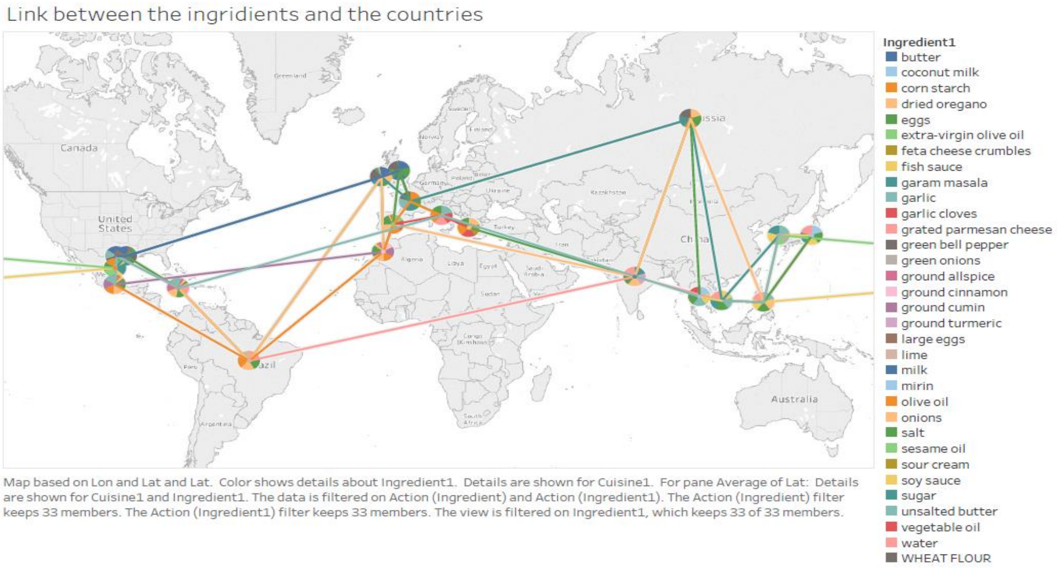


Fig. 4. Countries ranked by number of ingredients not-normalized for number of recipes

Fig. 5. Below shows the diversity of ingredients in different countries. We used tableau to get the below image. It is an interactive visualization which shows the connection between the cuisine and the ingredients that are used. As shown by colored lines connecting different countries, we see network developing based on unique ingredient shared by set of respective two countries.

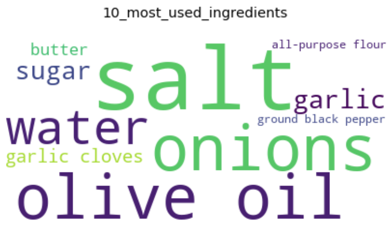
We can see that when an ingredient is selected it shows all the country that uses these ingredients.



Map based on Lon and Lat and Lat. Color shows details about Ingredient1. Details are shown for Cuisine1. For pane Average of Lat: Details are shown for Cuisine1 and Ingredient1. The data is filtered on Action (Ingredient) and Action (Ingredient1). The Action (Ingredient) filter keeps 33 members. The Action (Ingredient1) filter keeps 33 members. The view is filtered on Ingredient1, which keeps 33 of 33 members.

Fig. 5. Network based on diversity of ingredients in different countries

We also showed most top 5 ingredients used in all the cuisines across the world. We can see salt, garlic, onions, sugar, all-purpose flour are used widely. It was imperative to look at least used ingredients as well. Looking at these ingredients in Fig. 6., we can see they all appear unique. This makes sense why they are least used.



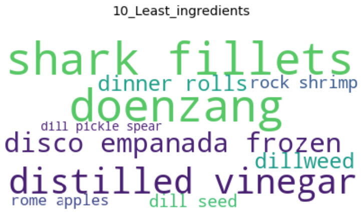


Fig.6 Network of top 5 ingredients from all countries

Fig. 6. Least used ingredients in world cuisine

Below figure shows how some of most common ingredients such as butter, eggs, garlic are widely used by all countries in relatively same proportion while some such as feta cheese, fish sauce are used specifically by selected countries.

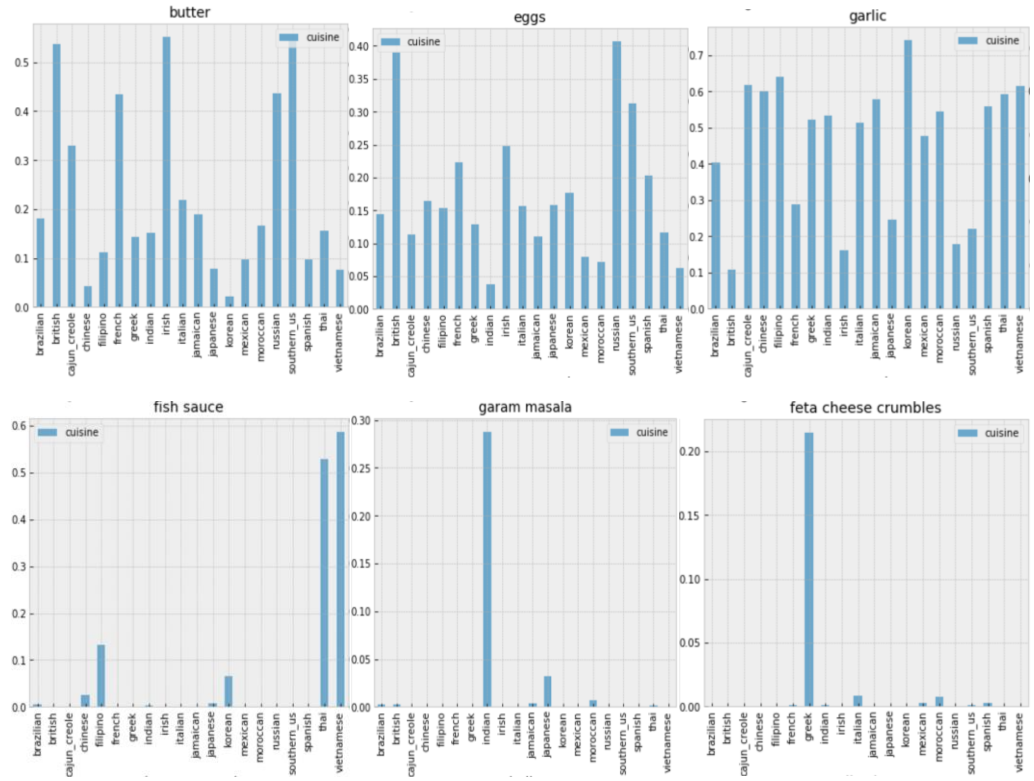


Fig. 7. Usage of most common ingredients across different cuisines

We wanted to check most common ingredients for countries to look for their eating habits. Fig. 8. below shows top 10 ingredients used by Indian cuisine. Information like this can be developed further to link it to health data for different countries. We also looked at other similar cases, which we do not present here due to space constraints.

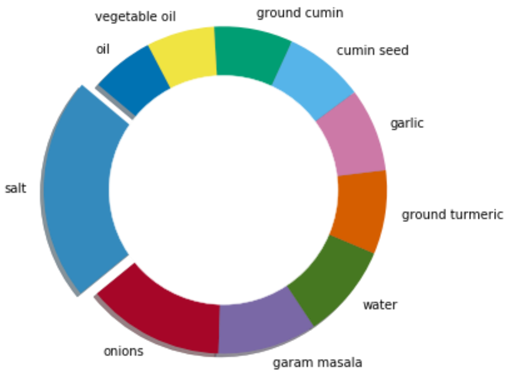


Fig. 8. Top 10 ingredients in Indian cuisine

We used different machine learning algorithms for classification of ingredients into different cuisines and checked their accuracy for how efficiently they performed. We used scikit-learn to perform our classification. First, we encoded our features to a matrix that the machine learning algorithms in scikit learn could use. This was done using a count vectorizer. Now that we have our feature matrix, we still need to encode the labels that represent the cuisine of each recipe. This is done with a label encoder. We can check the result by inspecting the encoders classes. We then trained our logistic regression on dataset and then checked how it performed on test data. It turns out it performed with a 78% accuracy. We also trained the data set using the random forest method, Multinominal Naïve Bayes and found that the logistic regression performed the best.

Classifier	Logistic Regression	Multinominal Naïve Bayes	Random Forest
Accuracy	78%	70%	56%

5-fold cross validation results for logistic regression

Precision	Recall	F1-Score
0.78	0.79	0.78

As seen below in Fig. 9., confusion matrix based on logistic regression shows Greek, Thai and Moroccan cuisine is difficult to predict as they share lot of ingredients with other cuisines. This confirms our findings from community detection as Greece, Thailand and Morocco share their ingredients with other cuisine and hence it is difficult to make prediction for them based on ingredient.

We wanted to see if highly connected i.e. most common ingredients co-occur with other highly used ingredients only or do they also get used with other less common ingredients. This would help us say if all recipes have some basic common taste or if recipes are unique in terms of taste

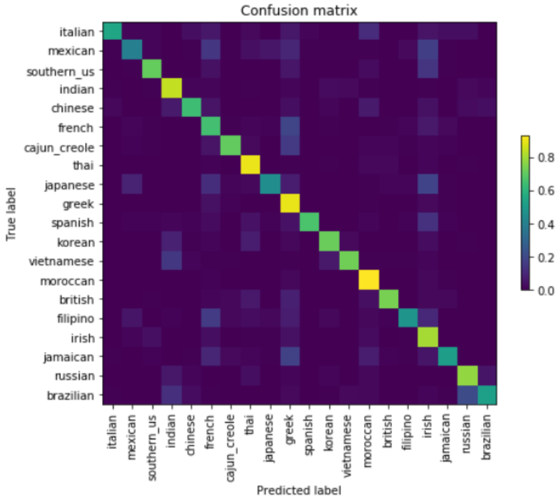
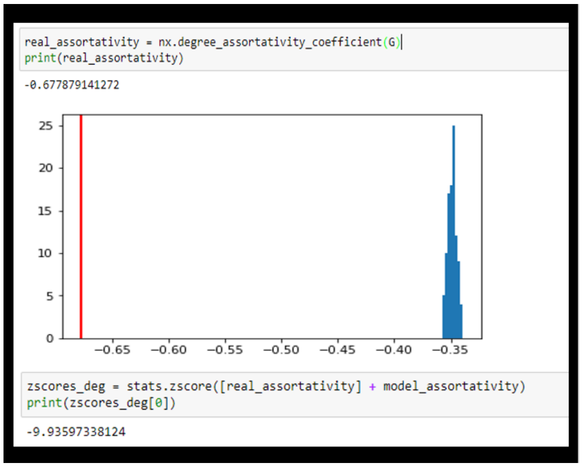


Fig. 9. Confusion matrix for cuisine prediction

We got degree assortivity value of -0.678. This means most common ingredients are also present in recipes with not so common ingredients. Meaning, recipes are unique in how different ingredients are used together and hence generating unique flavors. We also checked if this degree assortivity is statistically significant or not. Z-score of 9.94 is significantly lower than expected value. Meaning, if we are to look at another food network of similar degree sequence; there is no guarantee that it will have same degree assortivity.



```
real_assortativity = nx.degree_assortativity_coefficient(G)
print(real_assortativity)
```

```
-0.677879141272
```

```
zscores_deg = stats.zscore([real_assortativity] + model_assortativity)
print(zscores_deg[0])
```

```
-9.93597338124
```

Fig. 10. Results for degree assortivity hypothesis

After finding common ingredients and their usage by countries, we were interested in seeing nutritional contents of each cuisine. This was a difficult task and there was no data available showing how all recipes used said ingredients in what quantity. But, we used USDA data and nutritional contents of each ingredient per 100-gram basis.

We Then linked this with our top 5 ingredient list for each country. Fig. 12. shows how different cuisine and their top 5 ingredient rank in terms of protein and calorie contents.

We can see Italian food has maximum protein content and Brazilian food has least protein content from top 5 common ingredients for them. Italian food has parmesan cheese as one of most common ingredients and is reason for high protein content. French cuisine has most calories from its top ingredients as it has butter and olive oil in top 5 ingredient list.

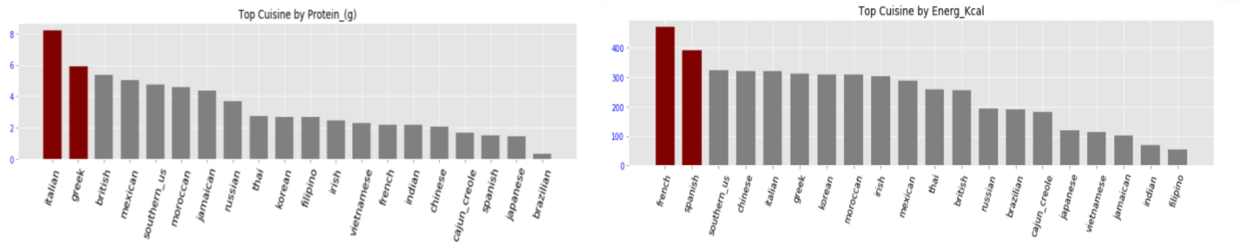


Fig. 12. Cuisine by Protein and calorie content of top 5 ingredients

6. Discussion

Looking at list of top 5 ingredients for each country/cuisine, we see that some ingredients have 2 names e.g. garlic and garlic cloves, olive oil and extra virgin olive oil. Raw data did not have information on how much quantity of every ingredient is used in each recipe. Because of this, it was not possible to make accurate analysis on nutritional content of different cuisine.We would be interested to see if this work can be taken further to see if such data can be used to link ingredients from different cuisine and its relation to quality of life around world. This can also possibly show our usage of ingredients that are harmful to nature because of their adverse contribution to global warming or other environmental issues. On positive side, it can help us plan effective production and transportation of ingredients.

7. Conclusion

Our project has been able to find network from world cuisine and show how we share different ingredients.

We presented a large-scale study of user-generated recipes on the web, their ingredients, nutrition, similarities across countries.

We were able to find common ingredients among different cuisine. Most common ingredients are salt, olive oil, black-pepper, sugar, onions, garlic, egg, butter. We were also able to find 6 communities of countries that share most ingredients. Our analysis shows India has most vegetarian dishes while France had most non-vegetarian dishes.

All data, codes and visualization files for this project can be found here.

8. Acknowledgements

We would like to thank Professor Y.Y. Ahn for support and lessons on network science in this project. We would also extend our appreciation to Kaggle and USDA website for hosting ingredient and nutrient datasets which is used in this project. We also acknowledge various online resources which helped understand Python and Gephi.

9. References

[1] Xiaowen Lin, Qian Dang, and Megan Konar - https://pubs.acs.org/doi/pdf/10.1021/es500471d, April 2014

[2] Cereal Ingredient Network Visualization — Student Work @ Pratt Institute School of Information, http://studentwork.prattinfoschool.nyc/blog/coursework/information-visualization/cereal-ingredient-network-visualization/1098/rsif.2014.0623 , November 7 ,2016

[3] Flavor network and the principles of food pairing. Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow & Albert-László Barabási , https://www.nature.com/articles/srep00196, 15 December 2011

[4] Truth \& Beauty - Müsli Ingredient Network, Mortiz Stefnar , http://truth-and-beauty.net/,

[5] Bogojeska A., Kalajdziski S., Kocarev L. (2016) Processing and Analysis of Macedonian Cuisine and its Flavours by Using Online Recipes. In: Loshkovska S., Koceski S. (eds) ICT Innovations 2015. Advances in Intelligent Systems and Computing, vol 399. Springer, Cham

[6] Good to Eat: Riddles of Food and Culture - Marvin Harris - Google Books. https://books.google.com/books/about/Good_to_Eat.html?id=WV8WAAAAQBAJ&printsec=frontcover&source=kp_read_button#v=onepage&q&f=false

[7] https://dl.acm.org/citation.cfm?doid=2380718.2380757

[8] United States Department of Agriculture, https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-

[9] https://flothesof.github.io/kaggle-whats-cooking-machine-learning.html

[10] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[11] https://www.analyticsvidhya.com/blog/2018/02/the-different-methods-deal-text-data-predictive-python/

[12] https://www.kaggle.com/c/whats-cooking