Data Science Presentation – What's Cooking?

- By Ameya Deolalkar

Introduction:

- This project solves a **Kaggle Competition: What's Cooking?**. Project leverages Natural Language processing(NLP) techniques to predict a cuisine from a list of ingredients
- **Problem Statement: Use recipe ingredients to categorize the cuisine**

The usual….:

- Read .json data into pandas dataframe, define the features and dependent variable

[{ "id": 10259, "cuisine": "greek", "ingredients": ["romaine lettuce", "black olives", "grape tomatoes", "garlic", "pepper", "purple onion", "seasoning", "garbanzo beans"]

	cuisine	id	ingredients
0	greek	10259	[romaine lettuce, black olives, grape tomatoes...
1	southern_us	25693	[plain flour, ground pepper, salt, tomatoes, g...
2	filipino	20130	[eggs, pepper, salt, mayonaise, cooking oil, g...
3	indian	22213	[water, vegetable oil, wheat, salt]
4	indian	13162	[black pepper, shallots, cornflour, cayenne pe...

y = Prediction Piece

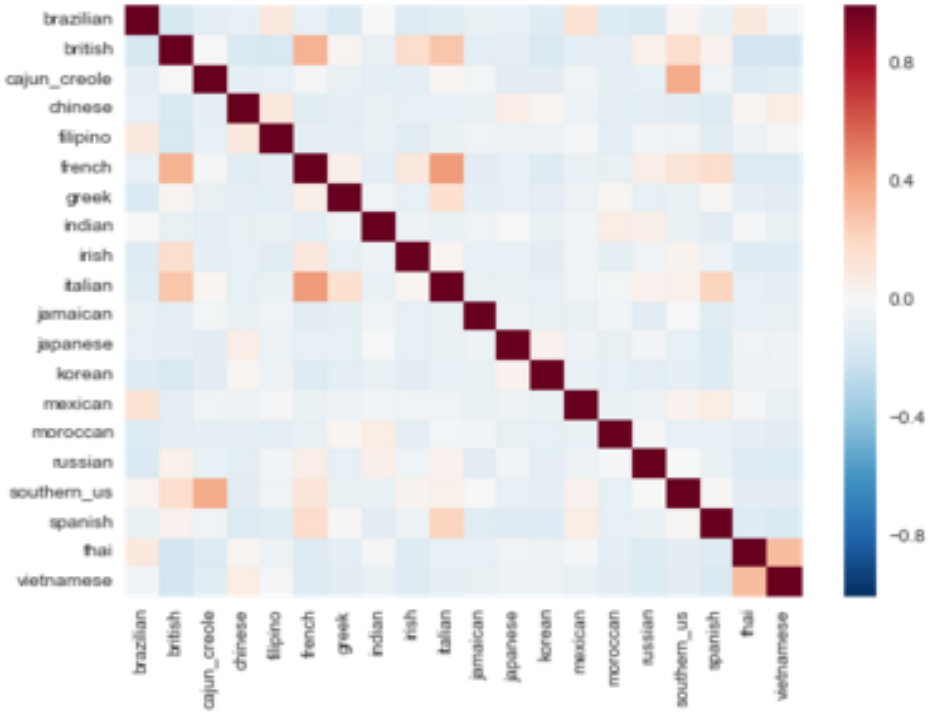
X = Analysis Piece

Implementations

Implementation 1:

The initial approach is to implement a Naive Bayes model on training and testing data tokenized using CountVectorizer

Evaluation metric: accuracy score, Result: Accuracy score 0.7198 (71.98%)



Implementations

Experiments to reduce error:

- **Use of stop_words –** Stop top 10 repeating words

  Result – Accuracy score: 0.7269
- **Use of max_df = 5000, max_features = 2100**

  Result – Accuracy score: 0.7272 (72.72%)

Implementation 2:

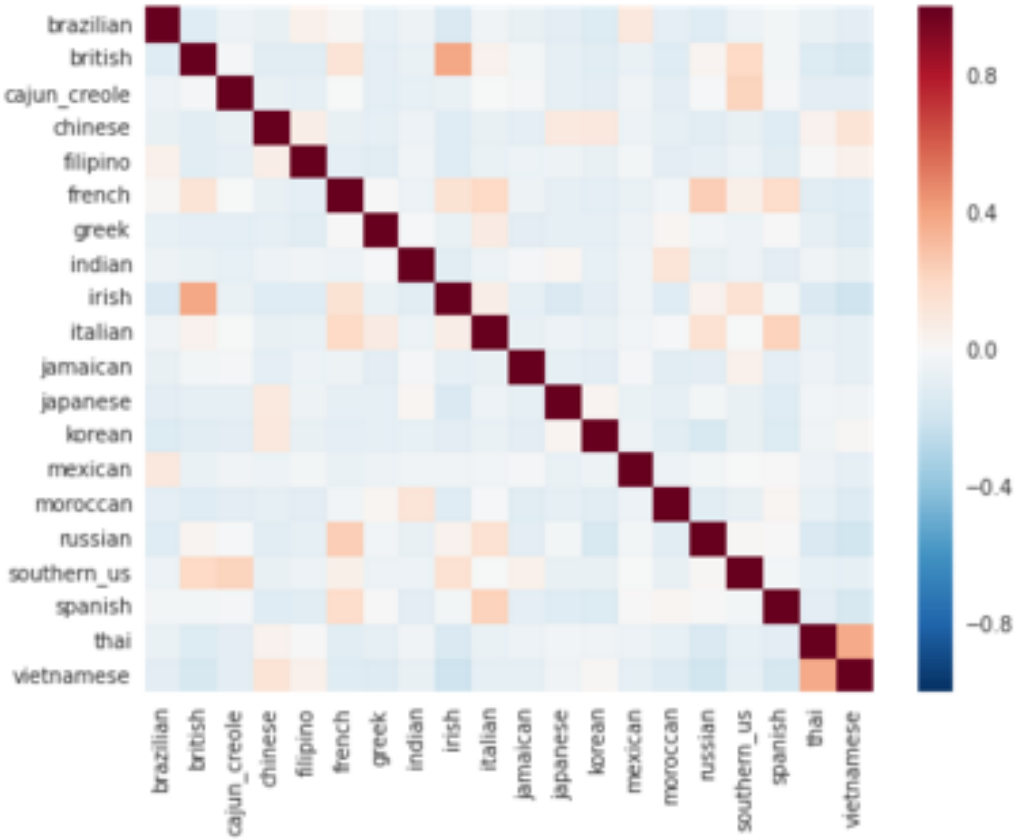
Feature Engineering – length of ingredient list

Logistic Regression – Has lower asymptotic error when number of features is large

TF-IDF Vectorizer – Frequency of an ingredient appearing in a cuisine is compared to its frequency across all cuisines

Implementations

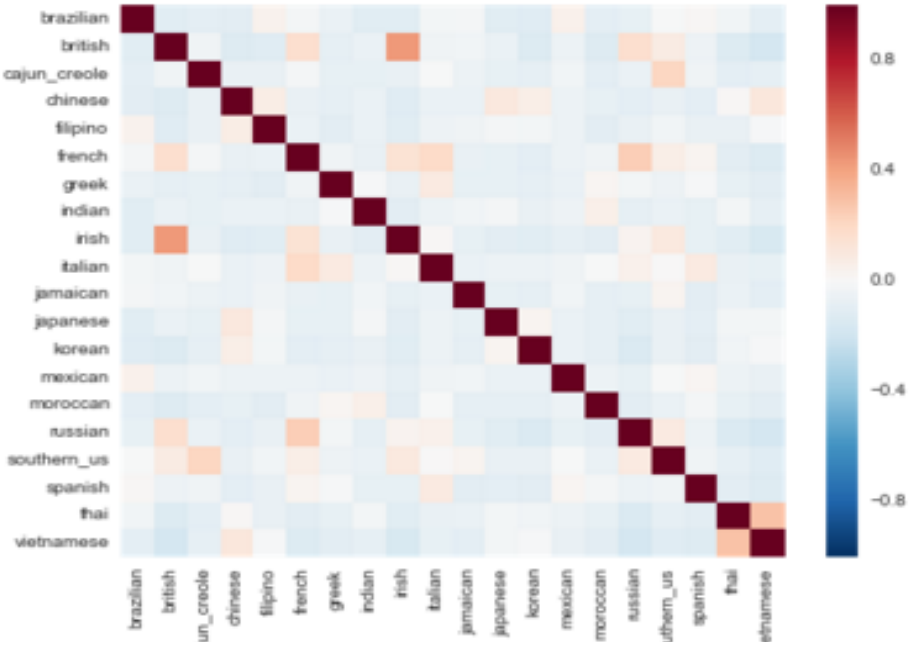
Result – Accuracy score: 0.7331 (73%)



Implementations

Experiments to reduce error:

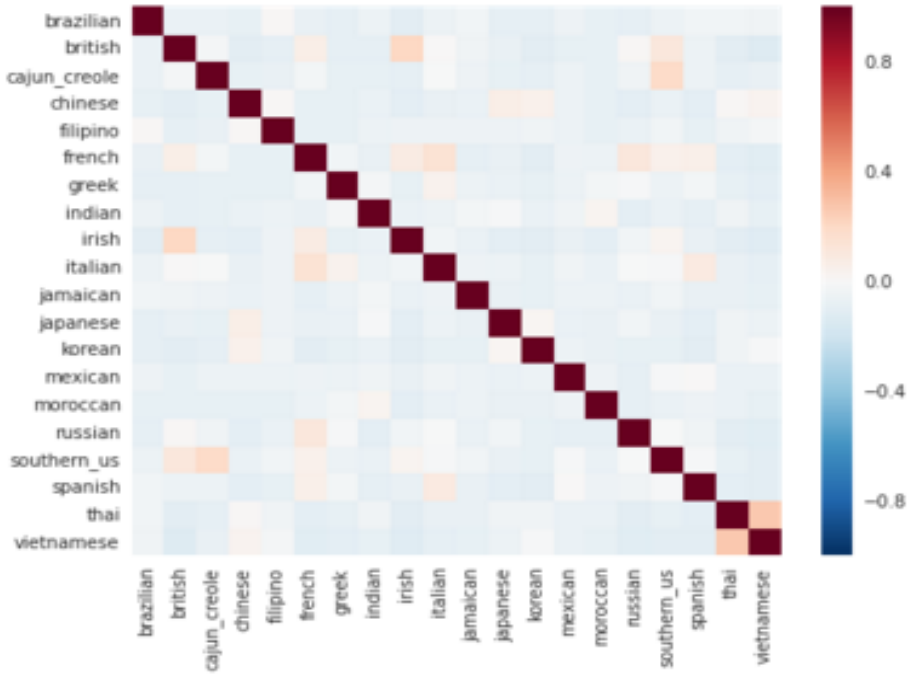
- Find top 20 common ingredients between cuisines that are being confused the most
- Set max_features to 700 which is approximately the mean of the frequency of occurrence of top 20 common ingredients in the cuisines selected
- Iterate for different max_features to find the best value
- Result: Accuracy Score - 0.7560 (75.60%)



Implementations

Implementation 3:

- **One-vs-rest classifier:** The strategy consists in fitting one classifier per class. For each classifier, the class is fitted against all the other classes.
- Result: Accuracy Score – 0.7865 (78.65%)



Conclusion

Progress Timeline