Where is the chef from?

Identify a recipe's country of origin from list of ingredients

Manish Pandit (manish7), Annie Pitkin (apitkin), Hermann Qiu (hq2128)
Department of Computer Science, Stanford University

Objective

How well can a machine guess a recipe's origin from just the list of ingredients?

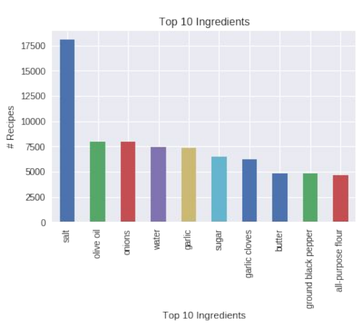
- Inspired by a recent kaggle challenge to analyze the strongest geographic and cultural associations - region's local foods
- There are overall thousands of features(ingredients) in the dataset
- Multi-class classification problem

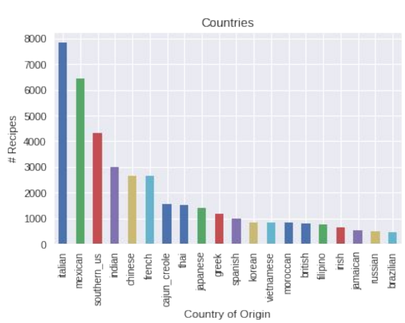
Dataset

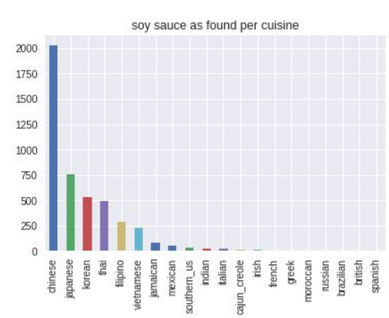
- Large public dataset with approximately 40,000 recipes and 6,800 unique ingredients from 20 countries.
- Each recipe uses a very small subset of ingredients.
- Disproportionate distribution of number of recipes across counties.

Feature Engineering

- Ingredients contains words that include brand names and descriptive preparatory steps.
- Design matrix leverages text feature extraction.
- Converted raw ingredients to a matrix of TF-IDF features.
- No need for data clean up since we are dealing with ingredients as text tokens.
- TF-IDF efficiently manages the sparsity of data. This not only counts but also returns the normalized count based on how many times an ingredient appears in all the recipes, for example, soy sauce is used most frequently in east Asian cuisines.

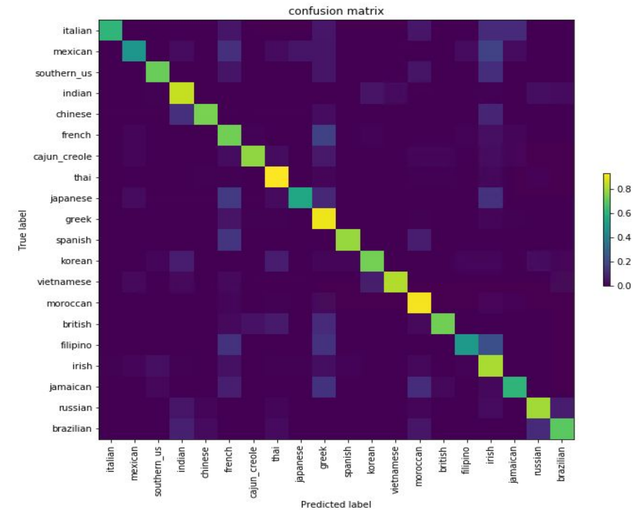






Classifiers and Results

- Split Labeled Dataset into Training and Test data with a 95% - 5% split.
- Trained and evaluated 7 different classifiers using sklearn, received accuracy scores between 64% and 79%.
- Used a Grid Search and 5-fold cross-validation scheme to tune hyperparameters on 4 of the best models.
- Highest accuracy score was now **82%** using SVM with 'C' = 10, 'gamma' =1, 'kernel' = rbf.
- Based on the noticeably small improvement via hyperparameter optimization, we realize we must generate more features.



Classifier	*Acc on Test Set*	*HP Opt*	*CV #*	*CV Score on Training Set*	*New Acc*
MLP Neural Net	***77%***	***Yes***	***-***	***-***	***80%***
Logistic Reg	***79%***	***Yes***	***5***	***79%***	***80%***
SVM	***79%***	***Yes***	***5***	***81%***	***82%***
Decision Tree	***64%***	***No***	***-***	***-***	***-***
Passive Agg	***74%***	***Yes***	***5***	***75%***	***75%***
Random Forest	***66%***	***No***	***-***	***-***	***-***
Multinomial NB	***69%***	***Yes***	***5***	***73%***	***74%***

Error Analysis

- Initial baselines were established without pre-processing the data.
- Attempts to homogenize the ingredients by removing purality, brand names and descriptive tokens were not effective in improving performance.
- Converting raw ingredients tokens to a matrix of TF-IDF features improved performance of all classifiers significantly.

Future Work and Citations

- We have generated several more features, yet did not have time to train on them prior to today.
- We have been experimenting with recursive feature selection algorithms, and will put them into practice on our new feature vectors.

1. *Multiclass classification.* https://en.wikipedia.org/wiki/Multiclass_classification
2. *Softmax function:* https://en.wikipedia.org/wiki/Softmax_function
3. *Naive Bayes Classifier:* https://en.wikipedia.org/wiki/Naive_Bayes_classifier
4. *Mohamed, Aly (2005). "Survey on multiclass classification methods" (PDF). Technical Report, Caltech.*