Team Clementine

Cuisine Predictor

Zhi Jiang
Jimin Kim
Jason Lee
Yumeng Li
Arpan Pal

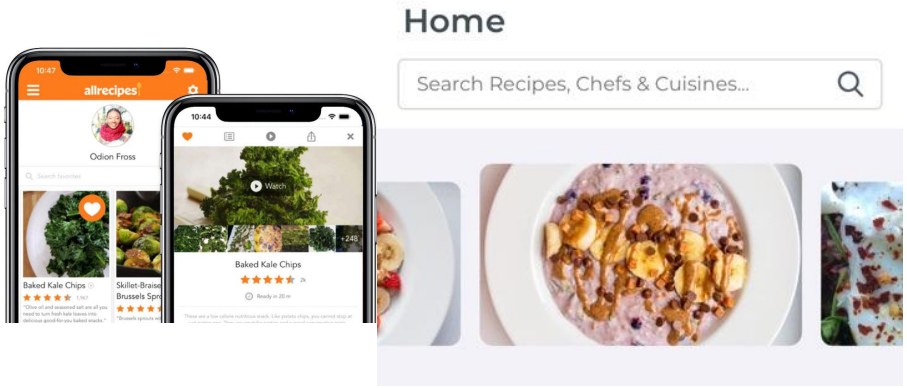
Problem

Given a list of ingredients, predict what type of cuisine the dish is.

Machine Learning

Stakeholders

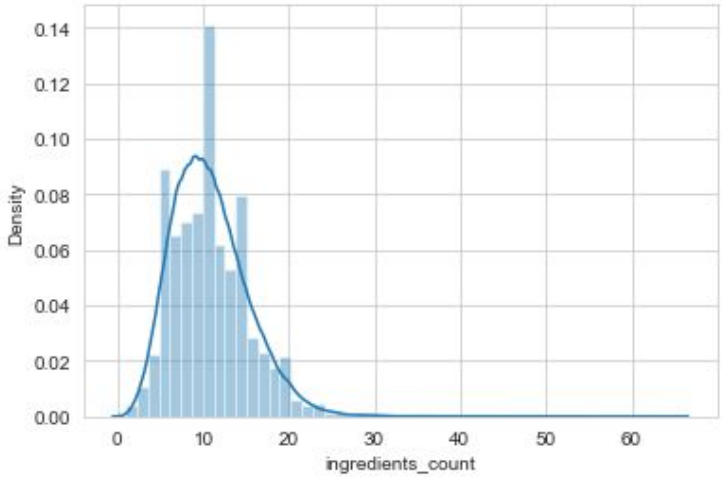
- Cookbook Services
  - When users share their recipes in a social media platform, the service can label them by cuisine and make a filter by cuisines.
- For individuals who wants to use ingredients in their fridge.
  - Recommend which cuisine is possible to make using those ingredients.



Data

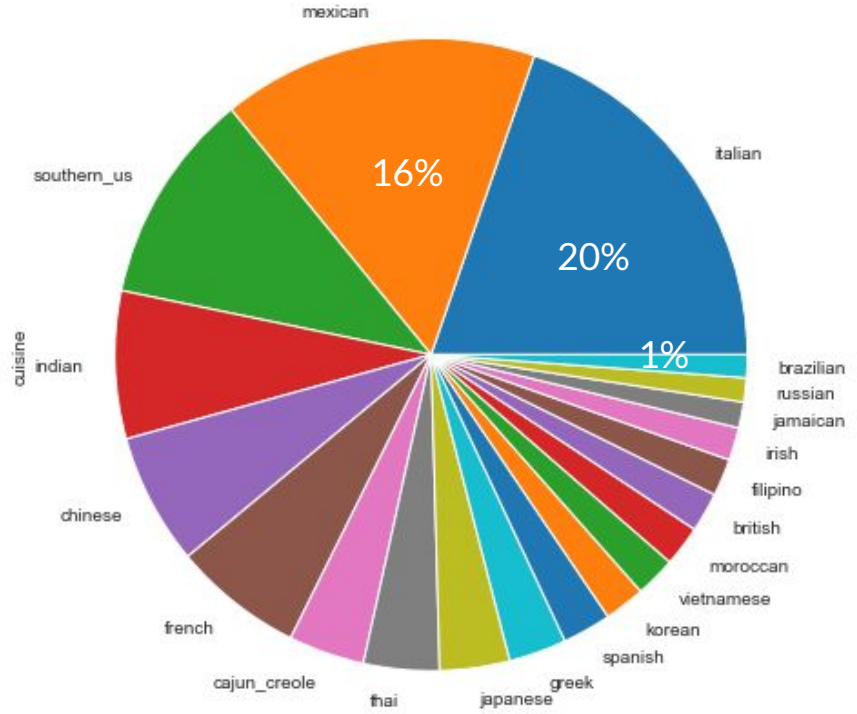
- A Kaggle competition data set provided by Yummly
- 39774 recipes (= # of rows)
- 20 cuisines
- 10.77 ingredients per recipe on average
- 6990 unique ingredients

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



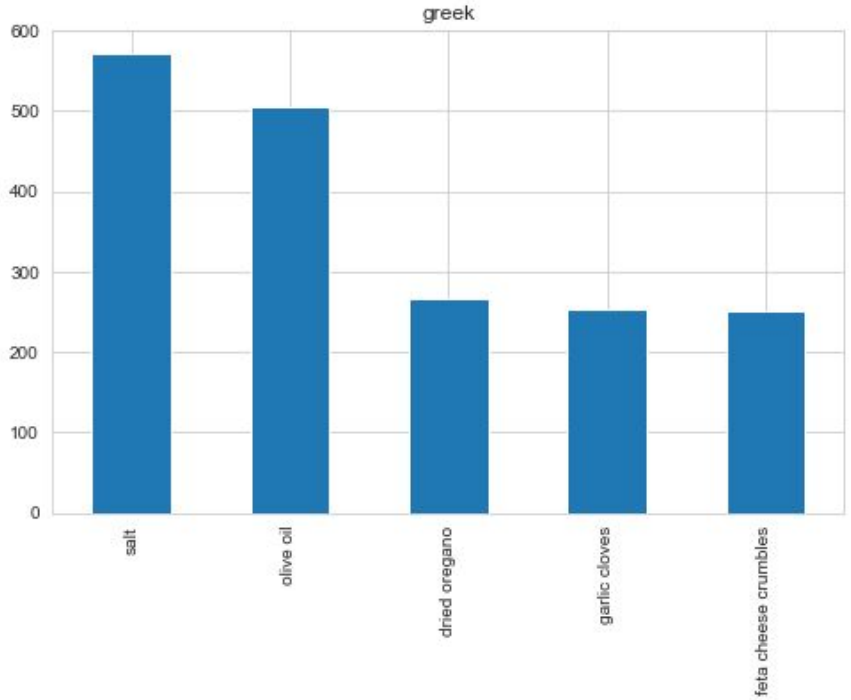
Exploratory Data Analysis

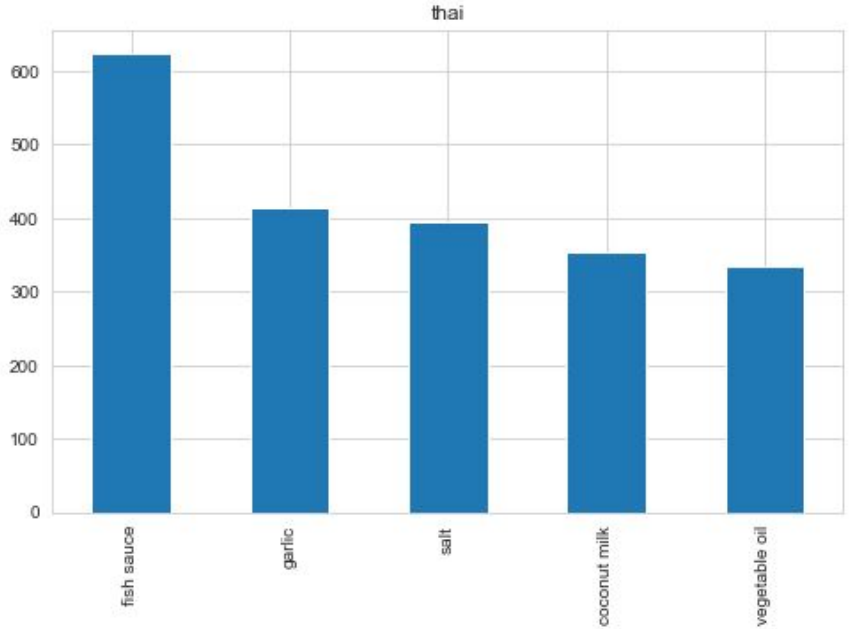
Proportion by cuisine in the training set



Exploratory Data Analysis

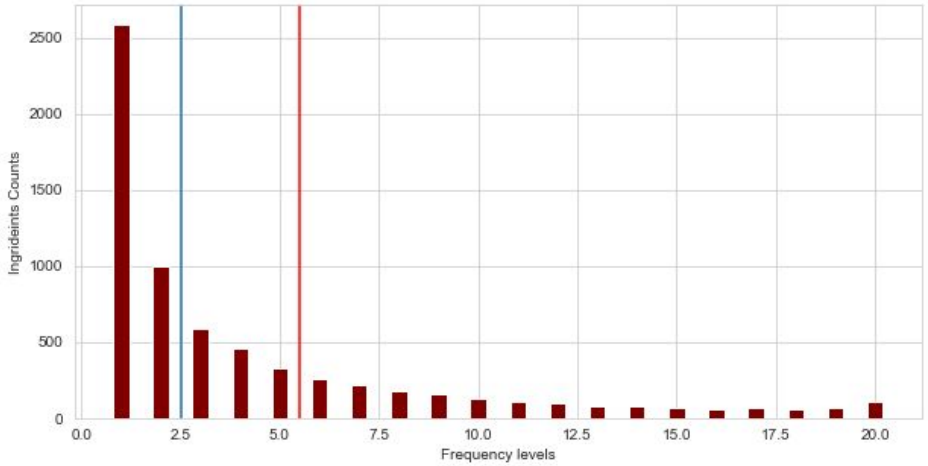
Top 5 ingredients given a cuisine





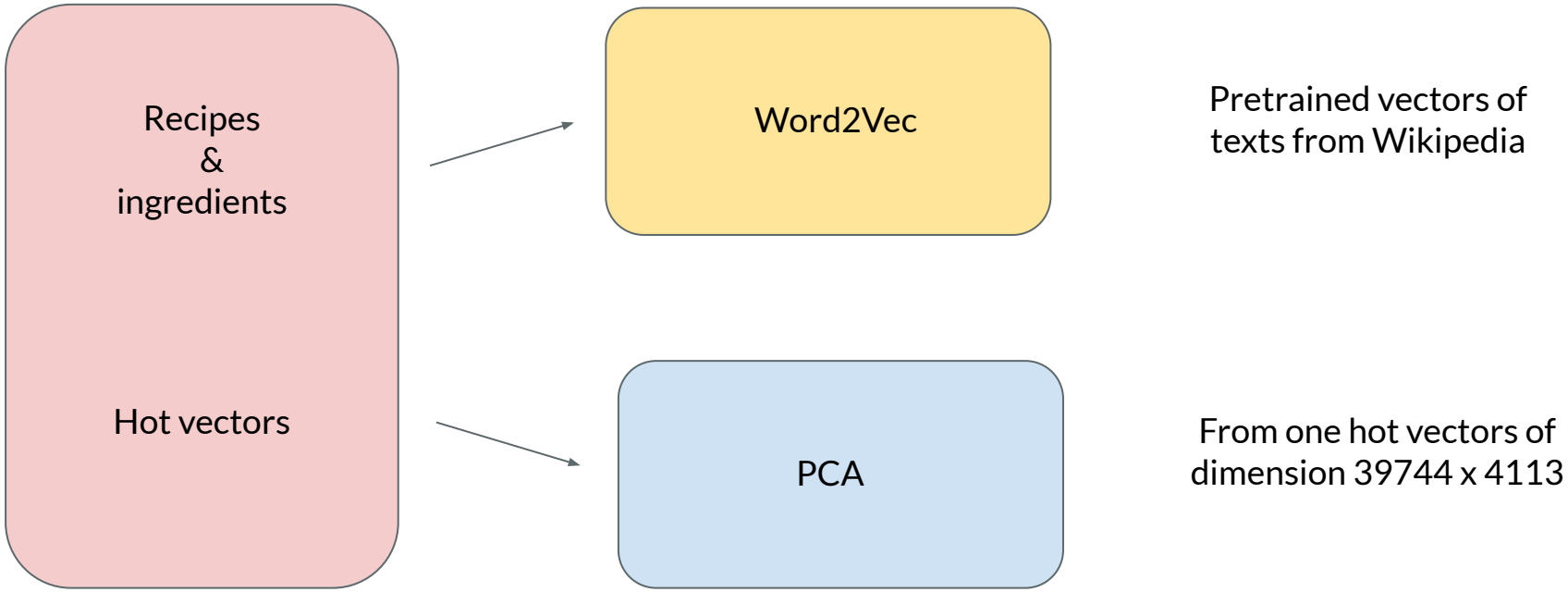
Data Setup

	id	cuisine	ingredients
8990	41124	indian	[butter]
22119	41135	french	[butter]



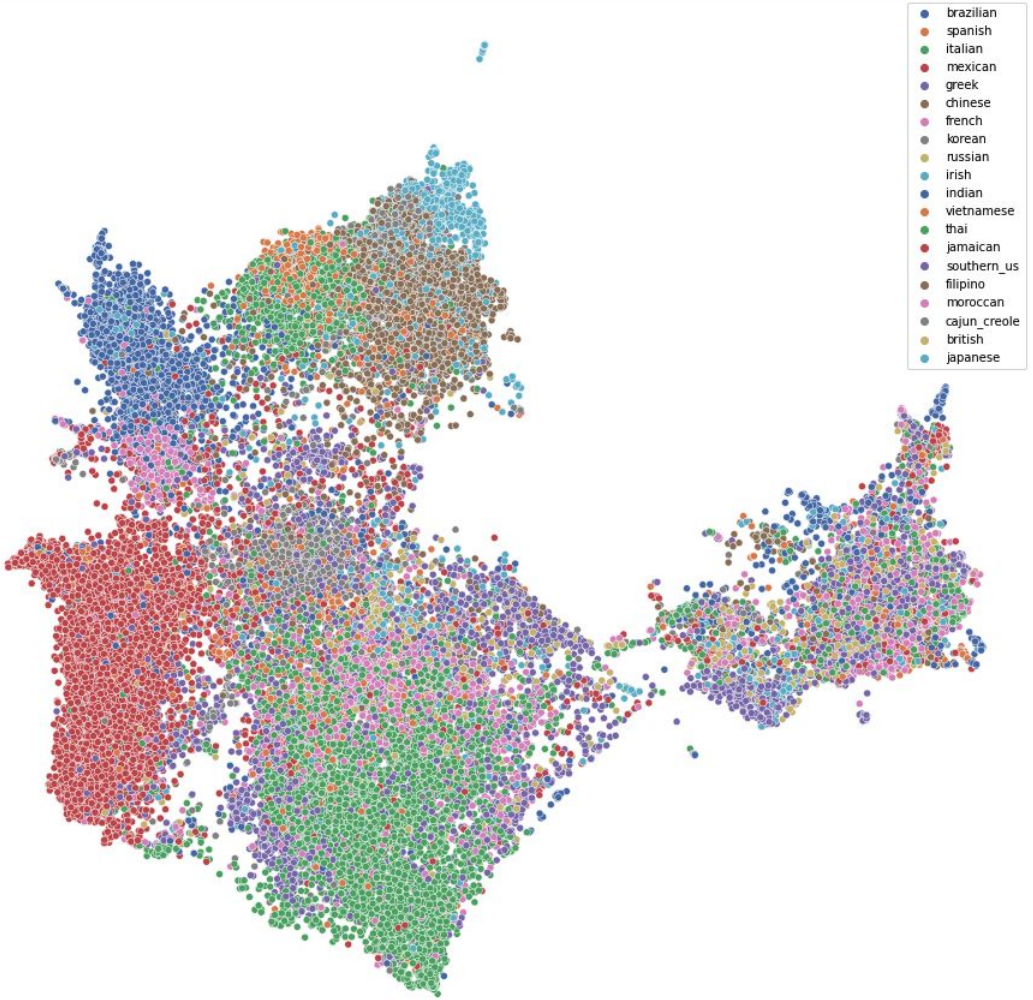
- All ingredients are converted to lowercase.
- We remove 3 pair of data points with the same set of ingredients but labeled under two different cuisines.
- Extract ‘Special ingredients’ which are exclusive to each cuisine.

Data Embedding

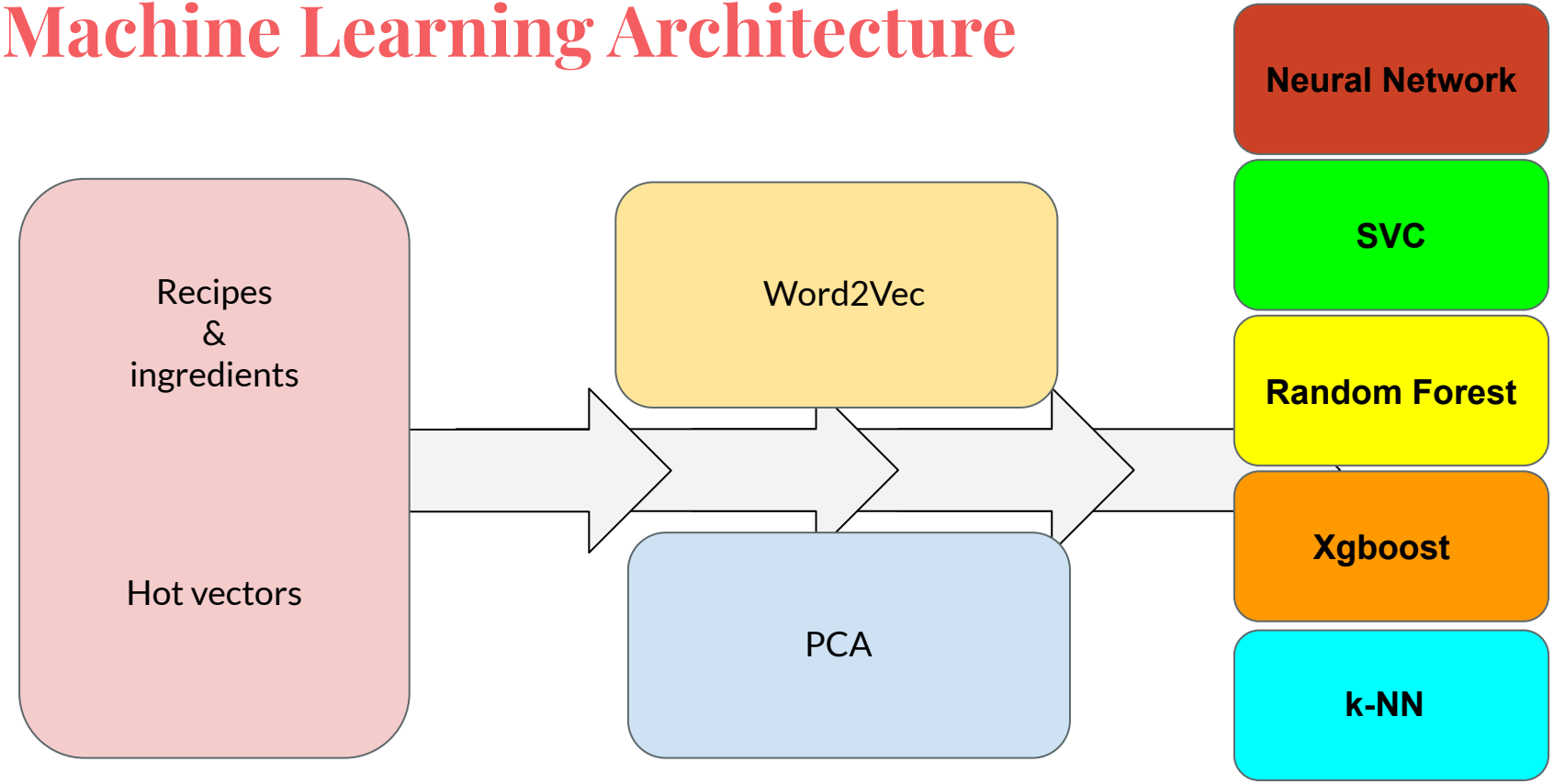


Data Embedding

UMAP visualization of 20-cuisine types of the training data after W2V-300 embedding



Machine Learning Architecture



Performance by Accuracy (on validation set)

Model \ Data	W2V-100	PCA-170	W2V-200	W2V-200 &PCA141	W2V-300	W2V-300 &PCA196
Fine Tuning SVC	0.7510	0.7010	0.7633	0.7750	0.7678	0.8001
Random Forest			0.6504			
Xgboost			0.7237	0.7227		
1-NN	0.7490	0.7470				
2-NN					0.7467	0.7332

Highlight

Identifying cuisines using raw ingredients will be valuable to companies that build recipe databases.

- Big data (40k rows, 7k ingredients), fast and easy to categorize (20 min)
- Up to 80% accuracy
- Word2Vec is powerful even for new data points (external validity).
- Tried out many different models.

Future Work

- Expand data by adding other features like the amount of ingredients, cooking time, or images of ingredients.
- Decision Tree can suggest the significance of the ingredients.
- By applying better natural language processing packages, we may achieve better models.
- Evaluate the model on unseen ingredients.