Predict Type of Cuisine Using Logistic Regression

Lujin Zhao

What I have:

	cuisine	id	ingredients
0	greek	10259	[romaine lettuce, black olives, grape tomatoes, garlic, pepper, purple onion, seasoning, garbanzo beans, feta cheese crumbles]
1	southern_us	25693	[plain flour, ground pepper, salt, tomatoes, ground black pepper, thyme, eggs, green tomatoes, yellow corn meal, milk, vegetable oil]
2	filipino	20130	[eggs, pepper, salt, mayonaise, cooking oil, green chilies, grilled chicken breasts, garlic powder, yellow onion, soy sauce, butter, chicken livers]
3	indian	22213	[water, vegetable oil, wheat, salt]
4	indian	13162	[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]

In total, there are 20 cuisines, including Brazilian, British, French, Spanish and Thai, and over 6,000 ingredients.

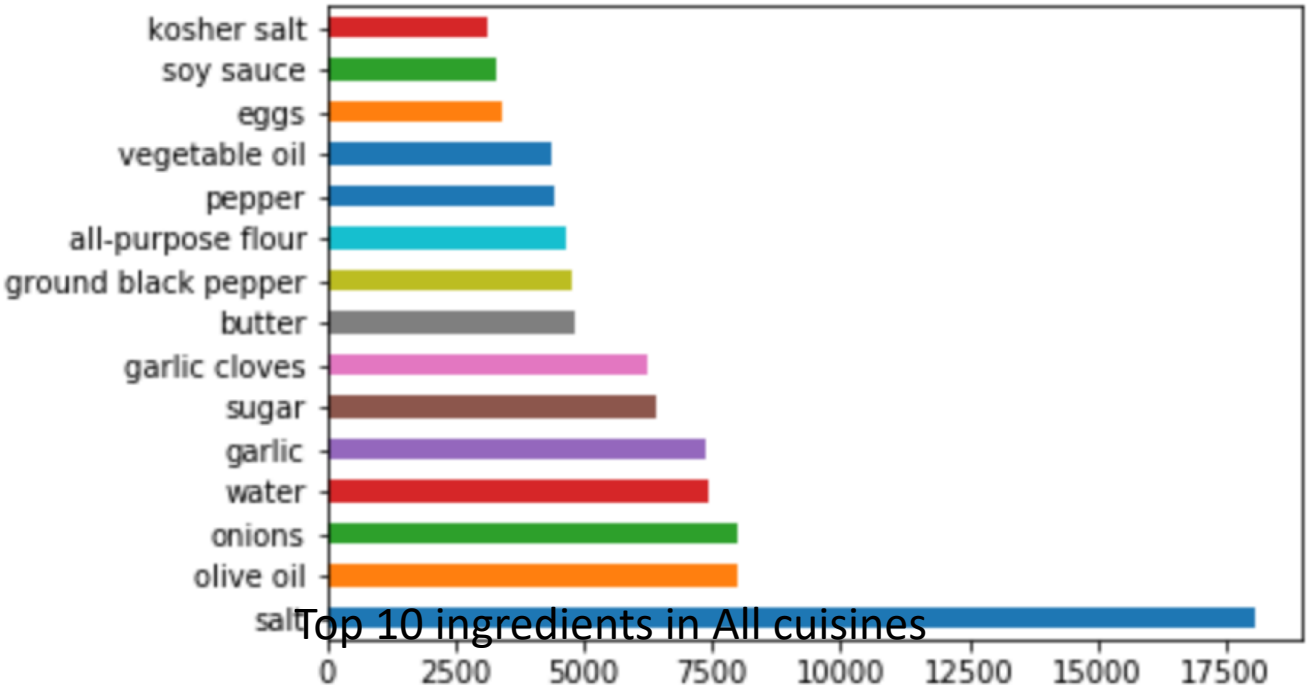
39774 different dishes

What I want to do

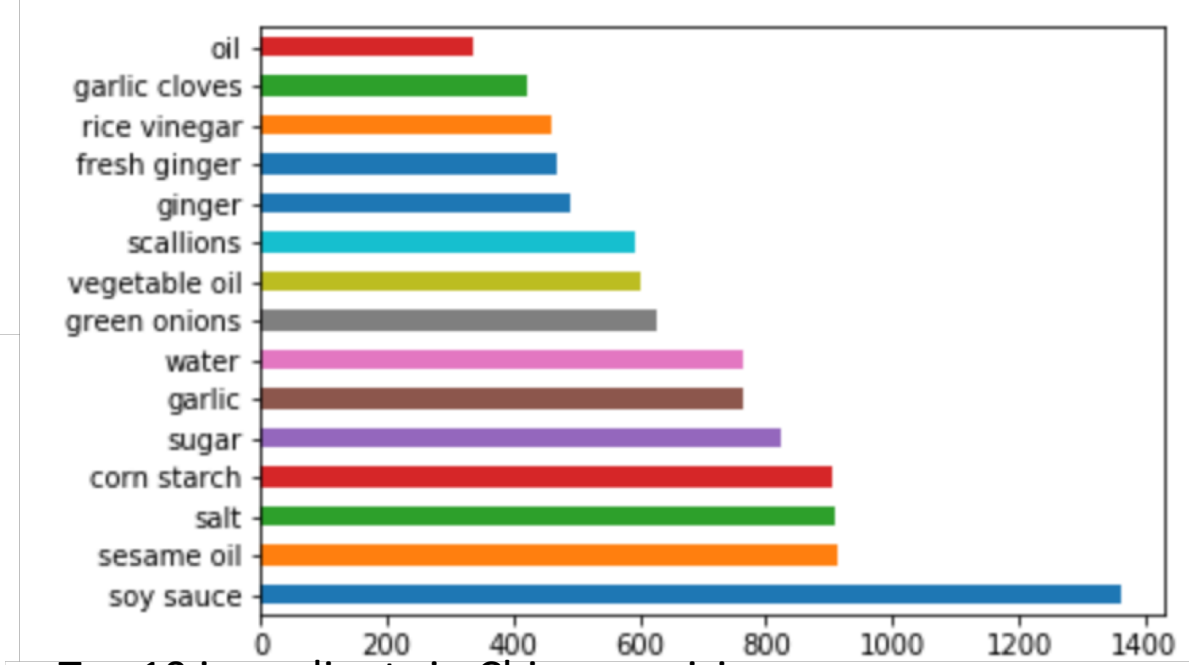
Run a model to accurately predict the type of cuisine.

Using dummy variables cannot reflect the importance of some ingredients.

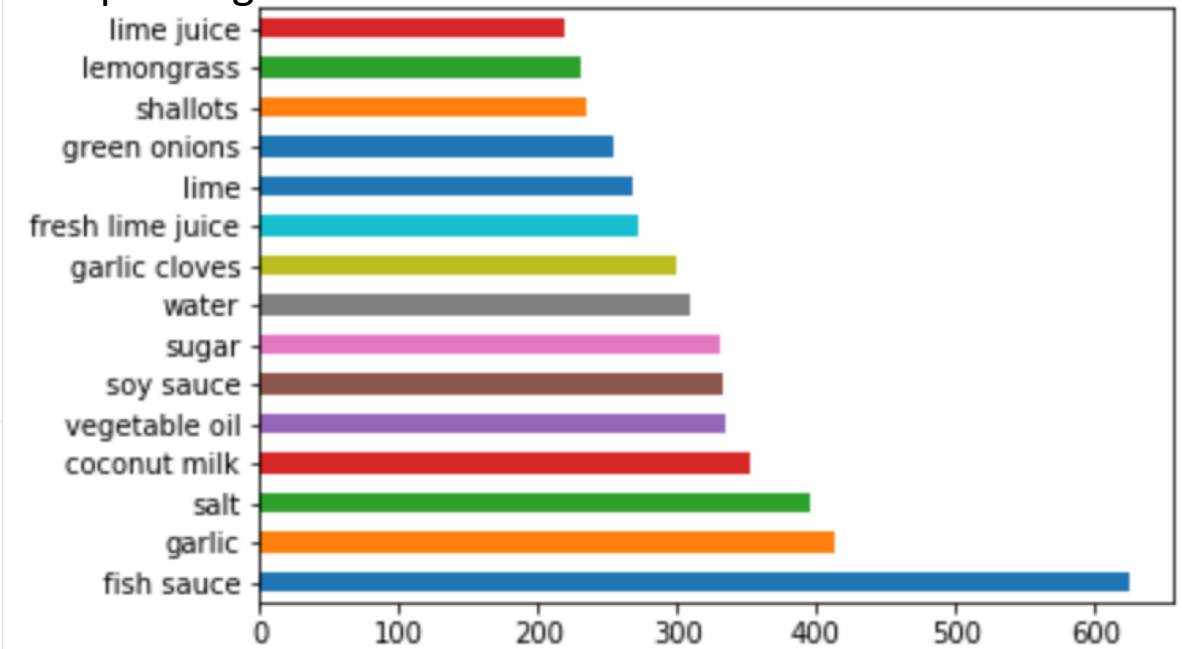
What I want to do



Top 10 ingredients in All cuisines



Top 10 ingredients in Chinese cuisine



Top 10 ingredients in Thai cuisine

What I did

Since some ingredients cannot help us identify what cuisine it belongs to

I used TF-IDF to put weight on different ingredients.

$$\operatorname{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

it gives a higher value when the ingredient occurs rarely.

What I get

After vectorize the ingredients list, I got a 39774 x 2768 matrix

The reduction in size is mainly due to I controlled for tense and plural.

The logistic regression gives a score of 78.7%