Fun With Recipes

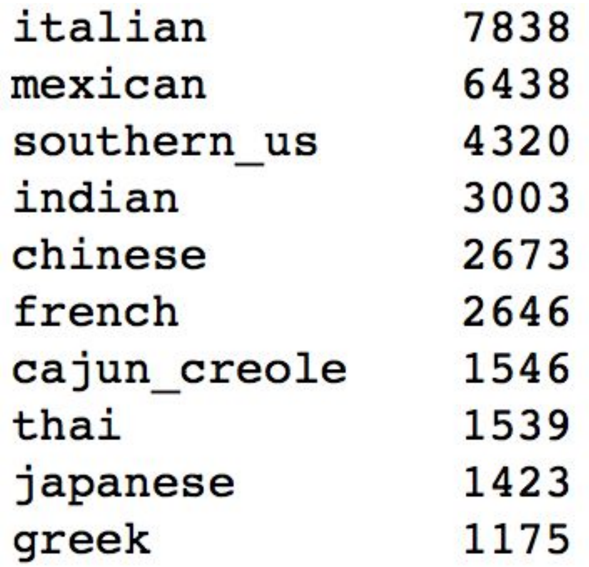
Raj Chakrabarty

Goals

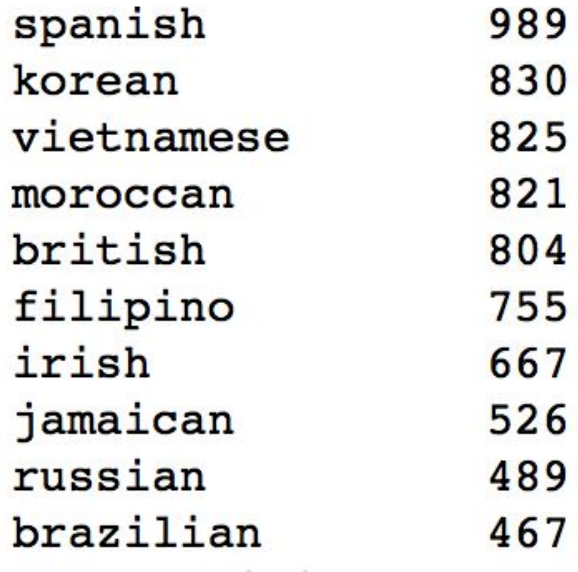
- Develop code for Natural Language Processing of recipe ingredients
- Build a predictive model for predicting the type of cuisine from a list of ingredients
- Learn what ingredients are most representative of each type of cuisine
- Measure which cuisines are most similar to one another

The Data

- 40,000 recipes from the website Yummly, with 428,000 total ingredients
- Each recipe contained a list of ingredients, and was categorized by cuisine
- 20 different cuisines were represented:



italian	7838
mexican	6438
southern_us	4320
indian	3003
chinese	2673
french	2646
cajun_creole	1546
thai	1539
japanese	1423
greek	1175



spanish	989
korean	830
vietnamese	825
moroccan	821
british	804
filipino	755
irish	667
jamaican	526
russian	489
brazilian	467

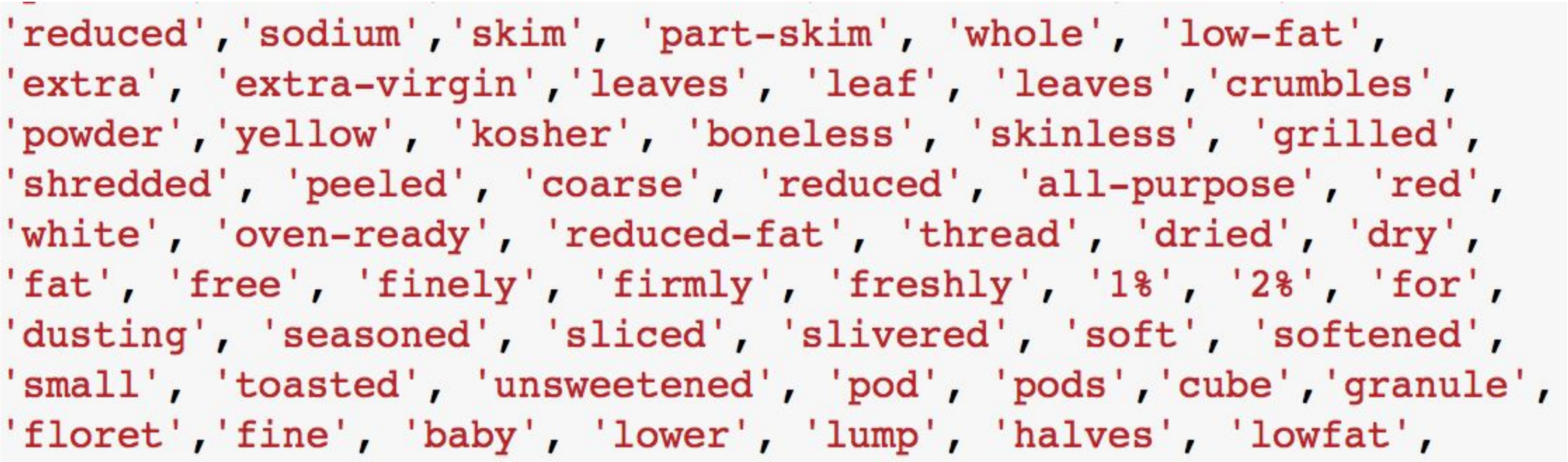
Natural Language Processing

- 1. remove pluralizations - remove ‘s’ from the end of each word, change words ending in ‘oes’ to end in ‘o’
- 2. Create a list of other words to be removed (peeled, fresh, ground, etc.)
- 3. Remove non-alphabetic characters, and words referring to ingredients ie, (20 oz.)
- 4. Save the 1000 most common ingredients to a csv file.
- 5. After inspecting the data, expand the list of words to be removed.
- 7. Standardize alternative spellings (anchovie/anchovy, yoghurt/yogurt)
- 6. Repeat the process until standardized list of one and two word ingredients emerges.

End Result: a standardized list of ingredients

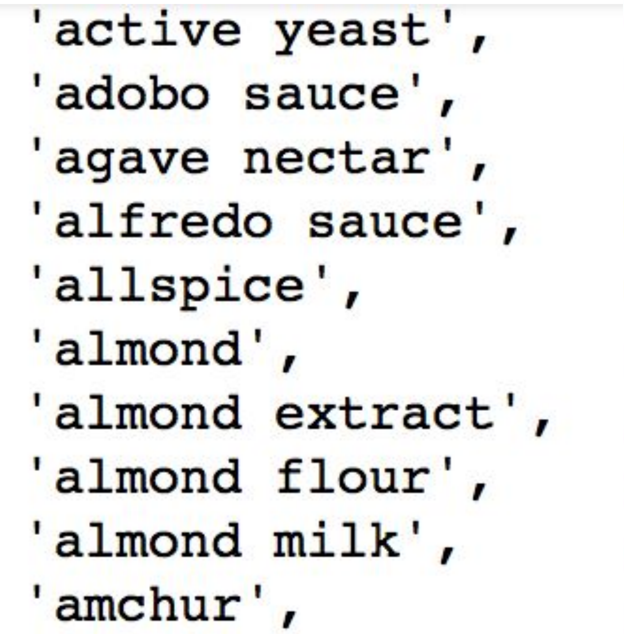
- 428,000 ingredients reduced to a list of roughly 900 standardized ingredient names.
- Ingredient names were one or two words
- Ingredients were unique, but not mutually exclusive: ‘chicken’, ‘chicken bouillon’, and ‘chicken breast’ were all included in the ingredient list

Stop Word Examples

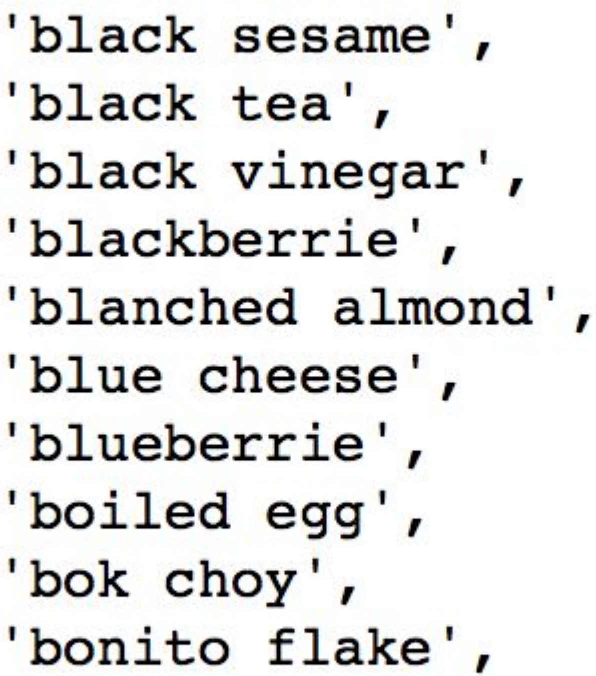


```
'reduced','sodium','skim', 'part-skim', 'whole', 'low-fat',
'extra', 'extra-virgin','leaves', 'leaf', 'leaves','crumbles',
'powder','yellow', 'kosher', 'boneless', 'skinless', 'grilled',
'shredded', 'peeled', 'coarse', 'reduced', 'all-purpose', 'red',
'white', 'oven-ready', 'reduced-fat', 'thread', 'dried', 'dry',
'fat', 'free', 'finely', 'firmly', 'freshly', '1%', '2%', 'for',
'dusting', 'seasoned', 'sliced', 'slivered', 'soft', 'softened',
'small', 'toasted', 'unsweetened', 'pod', 'pods','cube','granule',
'floret','fine', 'baby', 'lower', 'lump', 'halves', 'lowfat',
```

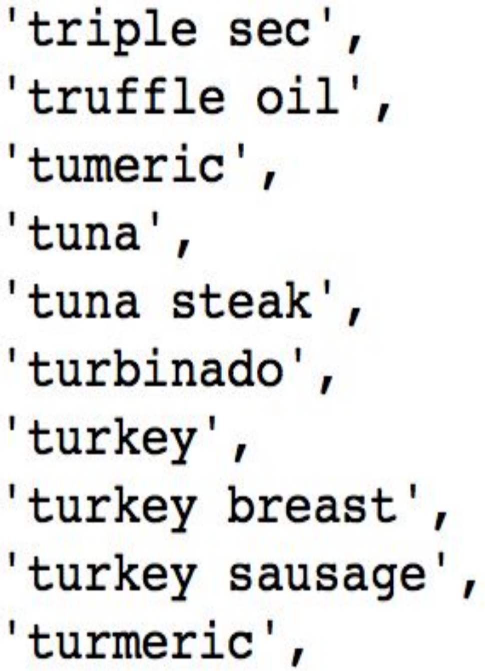
Final Ingredient Examples



```
'active yeast',
'adobo sauce',
'agave nectar',
'alfredo sauce',
'allspice',
'almond',
'almond extract',
'almond flour',
'almond milk',
'amchur',
```



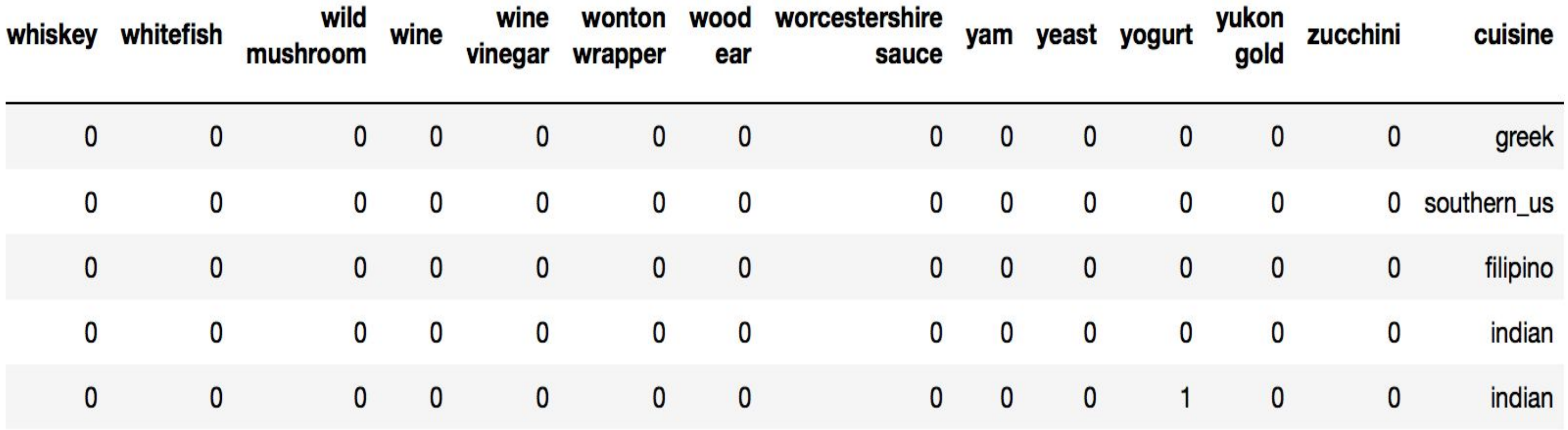
```
'black sesame',
'black tea',
'black vinegar',
'blackberrie',
'blanched almond',
'blue cheese',
'blueberrie',
'boiled egg',
'bok choy',
'bonito flake',
```



```
'triple sec',
'truffle oil',
'tumeric',
'tuna',
'tuna steak',
'turbinado',
'turkey',
'turkey breast',
'turkey sausage',
'turmeric',
```

Count Vectorization

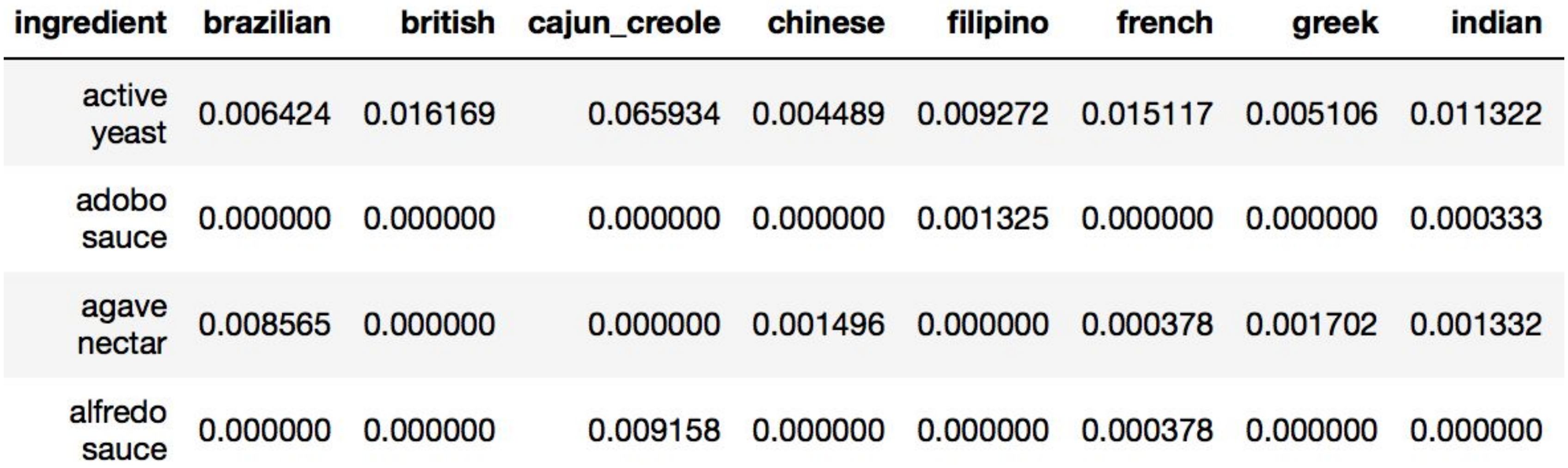
For each recipe, count the occurences of each ingredient.



whiskey	whitefish	wild mushroom	wine	wine vinegar	wonton wrapper	wood ear	worcestershire sauce	yam	yeast	yogurt	yukon gold	zucchini	cuisine
0	0	0	0	0	0	0	0	0	0	0	0	0	greek
0	0	0	0	0	0	0	0	0	0	0	0	0	southern_us
0	0	0	0	0	0	0	0	0	0	0	0	0	filipino
0	0	0	0	0	0	0	0	0	0	0	0	0	indian
0	0	0	0	0	0	0	0	0	0	1	0	0	indian

Finding Representative Ingredients

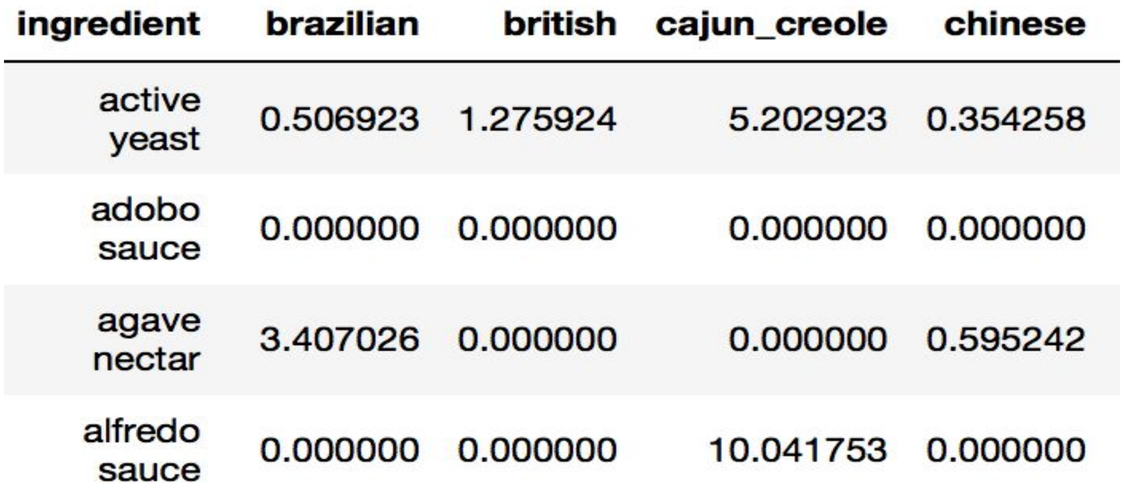
- Aggregate the ingredient counts by cuisine
- Calculate the percentage of of total recipes an ingredient appears in



ingredient	brazilian	british	cajun_creole	chinese	filipino	french	greek	indian
active yeast	0.006424	0.016169	0.065934	0.004489	0.009272	0.015117	0.005106	0.011322
adobo sauce	0.000000	0.000000	0.000000	0.000000	0.001325	0.000000	0.000000	0.000333
agave nectar	0.008565	0.000000	0.000000	0.001496	0.000000	0.000378	0.001702	0.001332
alfredo sauce	0.000000	0.000000	0.009158	0.000000	0.000000	0.000378	0.000000	0.000000

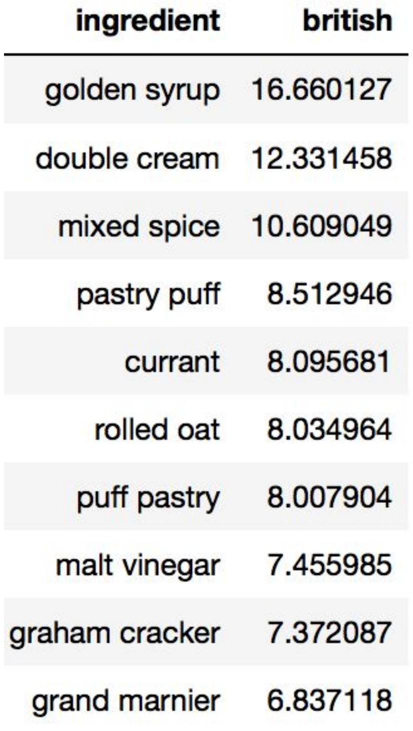
Finding Representative Ingredients

- For each ingredient, calculate the average occurrence across all cuisines, with each cuisine weighted equally
- Prevalence Ratio = average per cuisine / average across all
- Since there were 20 cuisines, this ends up being a number between 1 and 20

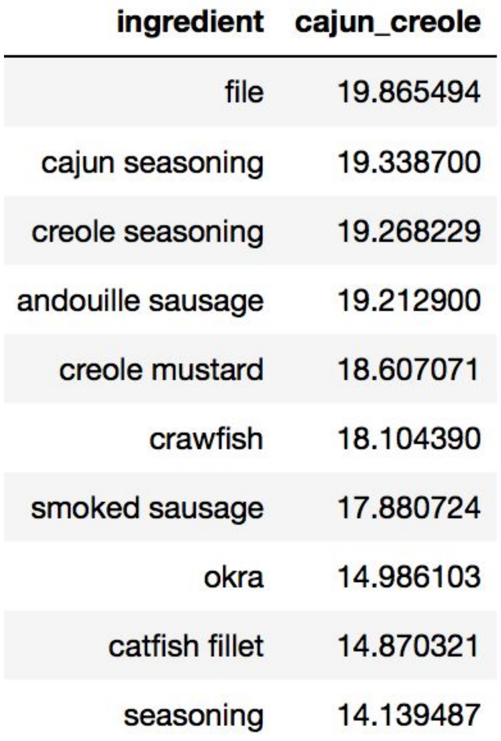


ingredient	brazilian	british	cajun_creole	chinese
active yeast	0.506923	1.275924	5.202923	0.354258
adobo sauce	0.000000	0.000000	0.000000	0.000000
agave nectar	3.407026	0.000000	0.000000	0.595242
alfredo sauce	0.000000	0.000000	10.041753	0.000000

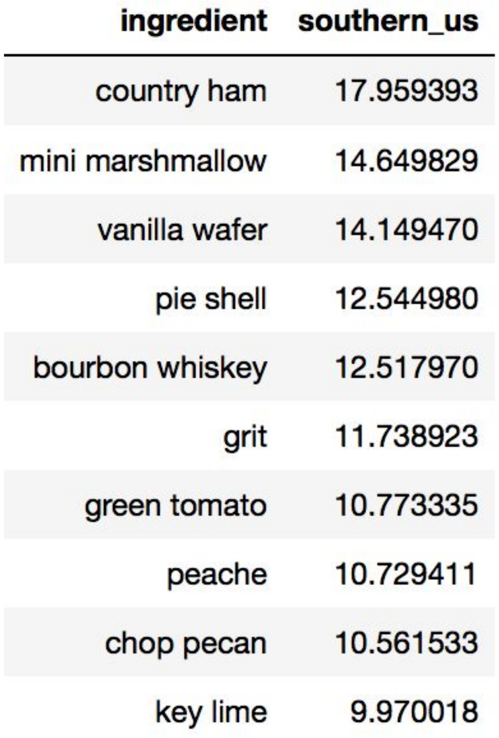
Top 10 Representative Ingredients



ingredient	british
golden syrup	16.660127
double cream	12.331458
mixed spice	10.609049
pastry puff	8.512946
currant	8.095681
rolled oat	8.034964
puff pastry	8.007904
malt vinegar	7.455985
graham cracker	7.372087
grand marnier	6.837118

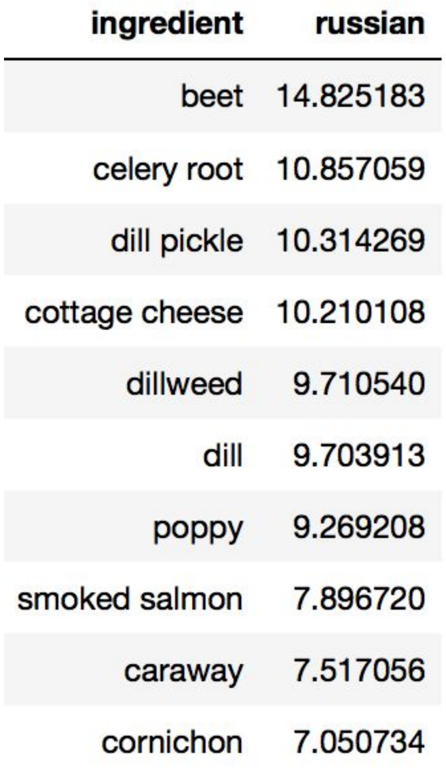


ingredient	cajun_creole
file	19.865494
cajun seasoning	19.338700
creole seasoning	19.268229
andouille sausage	19.212900
creole mustard	18.607071
crawfish	18.104390
smoked sausage	17.880724
okra	14.986103
catfish fillet	14.870321
seasoning	14.139487

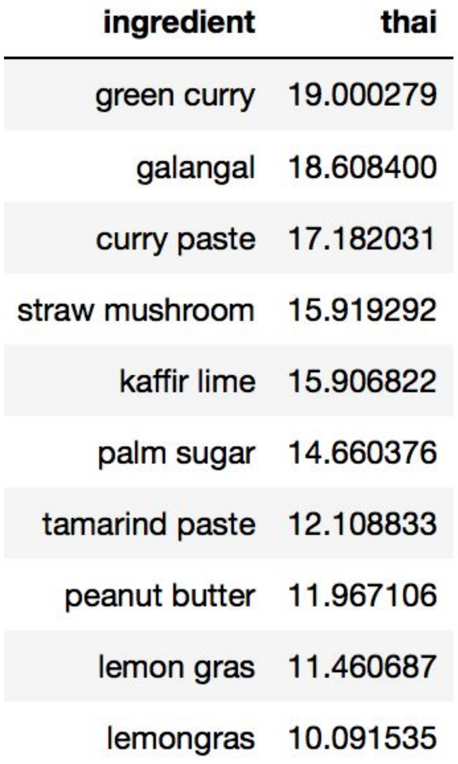


ingredient	southern_us
country ham	17.959393
mini marshmallow	14.649829
vanilla wafer	14.149470
pie shell	12.544980
bourbon whiskey	12.517970
grit	11.738923
green tomato	10.773335
peache	10.729411
chop pecan	10.561533
key lime	9.970018

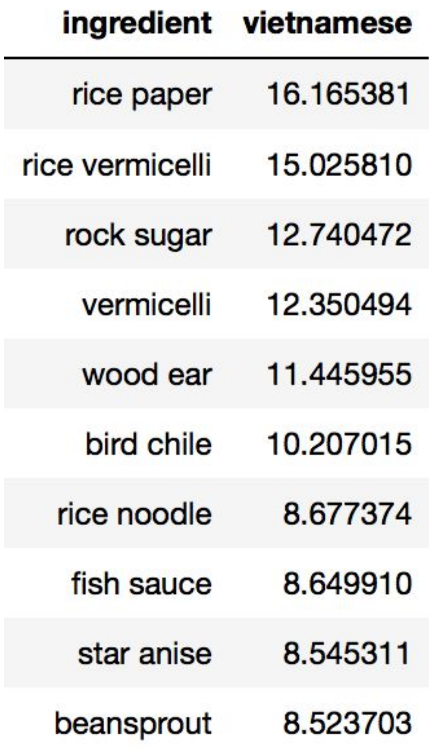
Top 10 Representative Ingredients



ingredient	russian
beet	14.825183
celery root	10.857059
dill pickle	10.314269
cottage cheese	10.210108
dillweed	9.710540
dill	9.703913
poppy	9.269208
smoked salmon	7.896720
caraway	7.517056
cornichon	7.050734



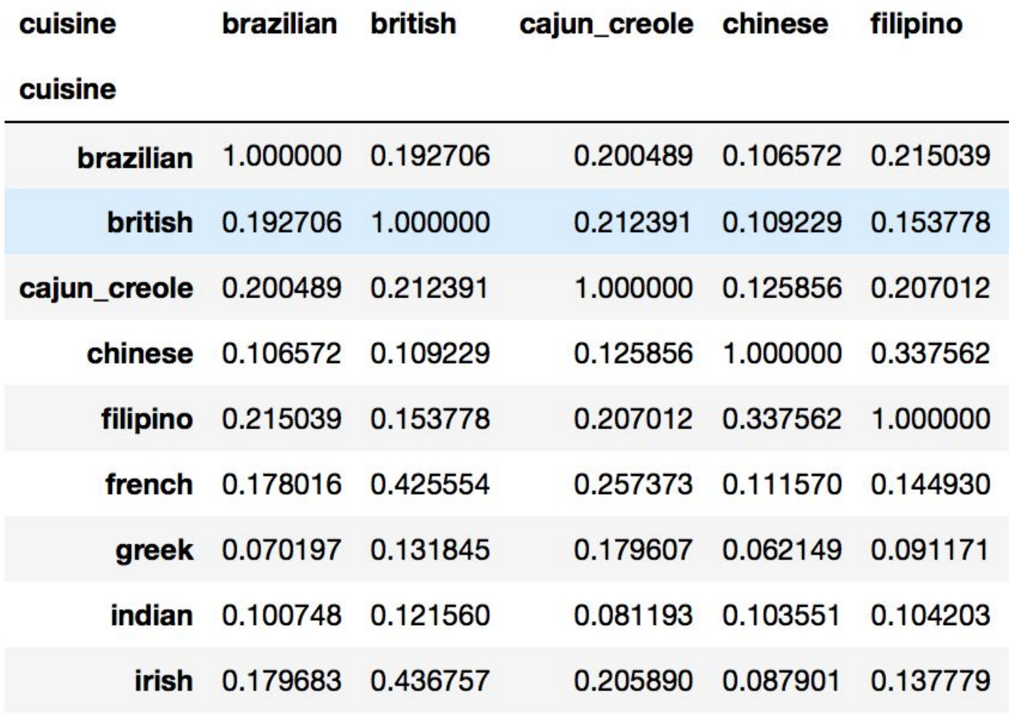
ingredient	thai
green curry	19.000279
galangal	18.608400
curry paste	17.182031
straw mushroom	15.919292
kaffir lime	15.906822
palm sugar	14.660376
tamarind paste	12.108833
peanut butter	11.967106
lemon gras	11.460687
lemongras	10.091535



ingredient	vietnamese
rice paper	16.165381
rice vermicelli	15.025810
rock sugar	12.740472
vermicelli	12.350494
wood ear	11.445955
bird chile	10.207015
rice noodle	8.677374
fish sauce	8.649910
star anise	8.545311
beansprout	8.523703

Calculating Cuisine Similarity

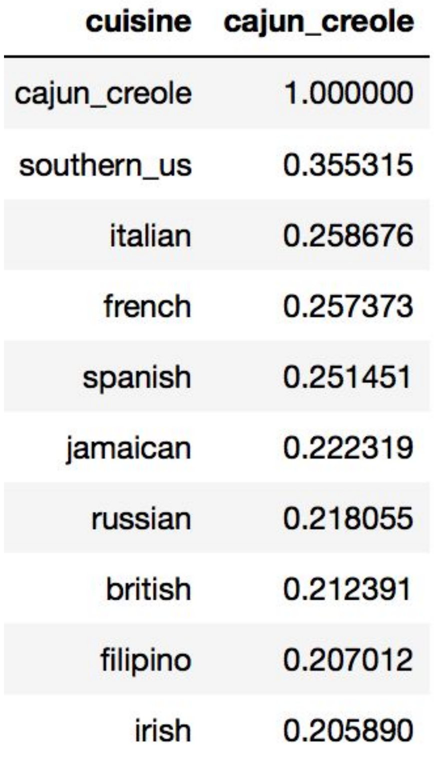
- Use the Prevalence Ratios to calculate Cosine Similarity between cuisines.



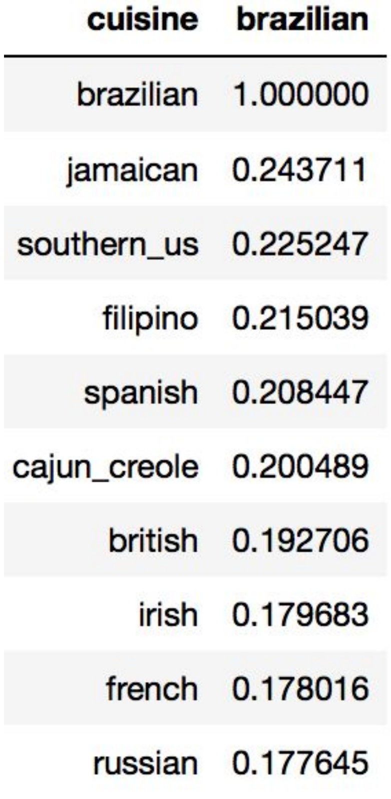
cuisine	brazilian	british	cajun_creole	chinese	filipino
cuisine					
brazilian	1.000000	0.192706	0.200489	0.106572	0.215039
british	0.192706	1.000000	0.212391	0.109229	0.153778
cajun_creole	0.200489	0.212391	1.000000	0.125856	0.207012
chinese	0.106572	0.109229	0.125856	1.000000	0.337562
filipino	0.215039	0.153778	0.207012	0.337562	1.000000
french	0.178016	0.425554	0.257373	0.111570	0.144930
greek	0.070197	0.131845	0.179607	0.062149	0.091171
indian	0.100748	0.121560	0.081193	0.103551	0.104203
irish	0.179683	0.436757	0.205890	0.087901	0.137779

Similarity scores are between
0 and 1

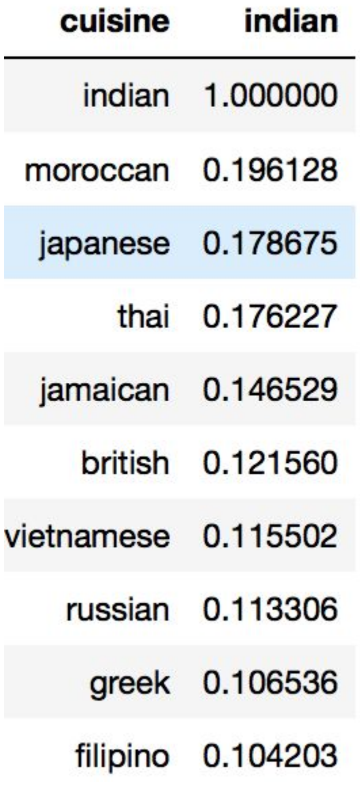
Cuisine Similarities



cuisine	cajun_creole
cajun_creole	1.000000
southern_us	0.355315
italian	0.258676
french	0.257373
spanish	0.251451
jamaican	0.222319
russian	0.218055
british	0.212391
filipino	0.207012
irish	0.205890



cuisine	brazilian
brazilian	1.000000
jamaican	0.243711
southern_us	0.225247
filipino	0.215039
spanish	0.208447
cajun_creole	0.200489
british	0.192706
irish	0.179683
french	0.178016
russian	0.177645



cuisine	indian
indian	1.000000
moroccan	0.196128
japanese	0.178675
thai	0.176227
jamaican	0.146529
british	0.121560
vietnamese	0.115502
russian	0.113306
greek	0.106536
filipino	0.104203

Most Similar Cuisines

● thai - vietnamese:	.51
● british - irish:	.44
● french - british:	.43
● chinese - vietnamese:	.40
● southern us - cajun-creole:	.36

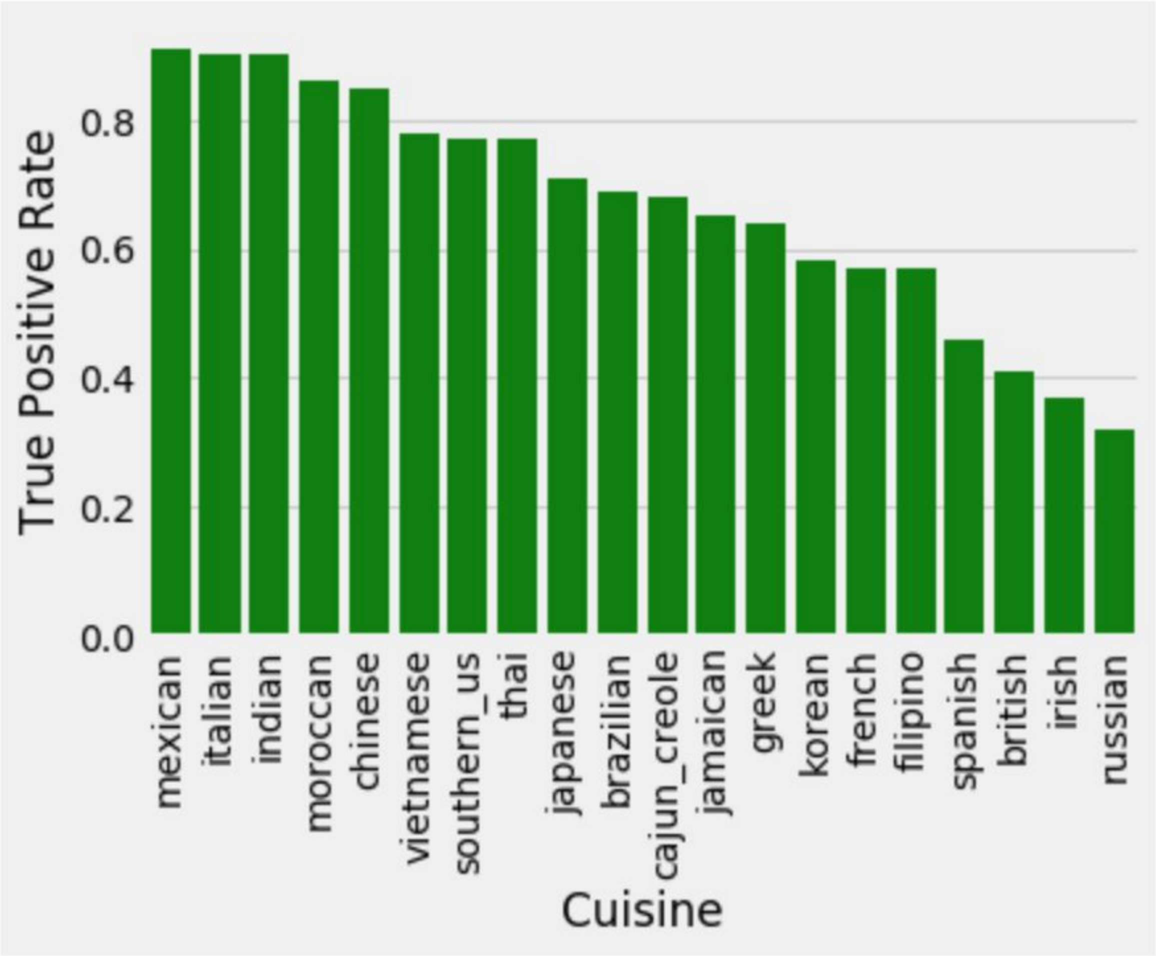
Most Dissimilar Cuisines

● mexican - japanese:	.033
● italian - indian:	.040
● korean - italian:	.047
● british - mexican:	.057

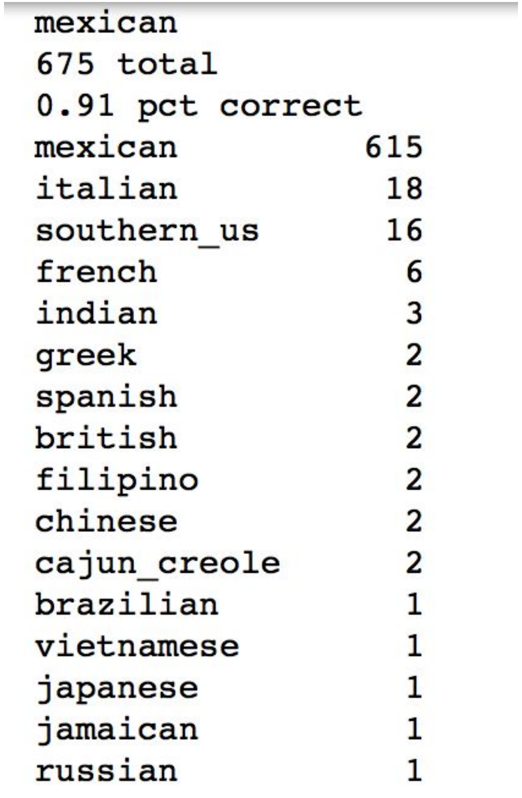
Predictive Modeling

- Used Count Vectorization and TF-IDF, with the custom vocabulary.
- Models: Random Forest, Logistic Regression, K-Nearest Neighbors.
- Best performance was from Random Forest and Logistic Regression.
- Final Model (using Logistic Regression) achieved an accuracy of 78%

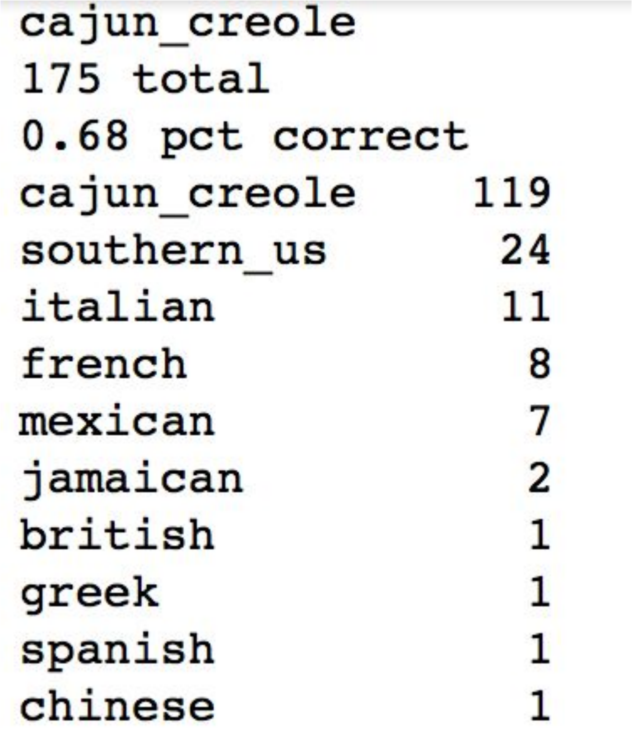
True Positive Rate by Cuisine



Looking at most common mis-predictions



```
mexican
675 total
0.91 pct correct
mexican         615
italian          18
southern_us      16
french            6
indian            3
greek             2
spanish           2
british           2
filipino          2
chinese           2
cajun_creole      2
brazilian         1
vietnamese        1
japanese          1
jamaican          1
russian           1
```



```
cajun_creole
175 total
0.68 pct correct
cajun_creole    119
southern_us      24
italian          11
french            8
mexican           7
jamaican          2
british           1
greek             1
spanish           1
chinese           1
```

Possible Next Steps

- Run against different datasets, and expand ingredient list
- Clustering of ingredients, cuisines
- Hierarchical representation of ingredients (ie, ‘chicken breast’ is a subset of ‘chicken’)

Time to eat!