Recipe Ingredients Inform Cuisine of Origin

Collins Metto[1], Christopher Criscitiello[2], Lili Cai[3], Mingyu Song[3]

Princeton University, Department of [1] Computer Science, [2] Mathematics, [3] Neuroscience

ABSTRACT

- Can recipe ingredients inform their cuisine of origin, and if so, which ingredients are most predictive? Can unsupervised learning pick up latent cuisines?
- The "What's Cooking" dataset has 39.5k recipes from 20 cuisines. Each recipe has 1 to 65 ingredients from 6714 possible ingredients.
- The Kaggle 2015 challenge: use supervised learning to classify these 20 cuisines. We aim to reproduce their leaderboard results within reason. We achieved their baseline result of .77 accuracy.
- We extend the challenge by using unsupervised learning to look at ingredient relationships, cuisine relationships, and if natural clusters emerge based on recipe ingredients. We find key relationships within and across ingredients and cuisines. Topic models were able to recover latent cuisines.

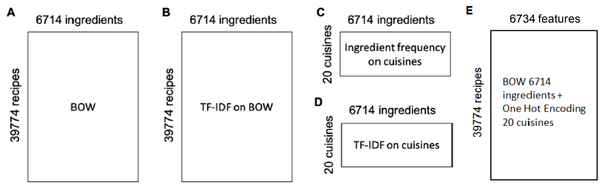
DATA EXPLORATION AND PREPROCESSING

Raw Dataset

# Recipes	39,774
# Cuisines	20
# Ingredients	6,714
# Ingredients/cuisine	1 to 65

	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...

Preprocessed Dataset



Cuisine	# Recipes	Top 5-10 Ingredients based on BOW	Top 5-10 Signature Ingredients based on TF-IDF on cuisines (Representation D)	Mean # Ing/Recip
Italian	7838	salt, olive oil, garlic cloves, grated parmesan cheese, garlic, ground black pepper	lasagna noodles, ricotta cheese, prosciutto, marinara sauce, fresh parmesan cheese	9.9
Mexican	6438	salt, onions, ground cumin, garlic, olive oil, chili powder, jalapeno chilies, avocado	refried beans, enchilada sauce, corn tortillas, tomatillos, salsa	10.9
Southern US	4320	salt, butter, all-purpose flour, sugar, large eggs, baking powder, buttermilk	grits, collard greens, buttermilk, bourbon whiskey, yellow corn meal	9.6
Indian	3003	salt, onions, garam masala, water, ground, turmeric, garlic, cumin seed, ground	garam masala, curry leaves, paneer, ghee, coriander powder, cumin seed	12.7
Chinese	2673	soy sauce, sesame oil, salt, corn starch, sugar, garlic, water, green onions	Shaoxing wine, oyster sauce, sesame oil, hoisin sauce, dark soy sauce	11.9
French	2646	salt, sugar, all-purpose flour, unsalted butter, olive oil, butter, water, large eggs	gruyere cheese, grated Gruyre cheese, chopped fresh thyme, fresh tarragon	9.8
Cajun Creole	1546	salt, onions, garlic, green bell pepper, butter, olive oil, cayenne pepper, cajun	cajun seasoning, andouille sausage, creole seasoning, file powder, crawfish	12.6
Thai	1539	fish sauce, garlic, salt, coconut milk, vegetable oil, soy sauce, sugar, water	fish sauce, Thai red curry paste, red curry paste, kaffir lime leaves, beansprouts	12.5
Japanese	1423	soy sauce, salt, mirin, sugar, water, sake, rice vinegar, vegetable oil, scallions	mirin, sake, dashi, nori, konbu, sushi rice, dried bonito flakes	9.7
Greek	1175	Salt, olive oil, dried oregano, garlic cloves, feta cheese crumbles	feta cheese crumbles, dried oregano, greek seasoning, pitted kalamata olives	10.2
Spanish	989	salt, olive oil, garlic cloves, extra-virgin, olive oil, onions, water, tomatoes	saffron threads, chorizo sausage, spanish chorizo, serrano ham, manchego cheese	10.4
Korean	830	soy sauce, sesame oil, garlic, green onions, sugar, salt, water, sesame seeds	Gochujang base, kimchi, sesame oil, gochugaru, toasted sesame seeds	11.3
Vietnamese	825	fish sauce, sugar, salt, garlic, water, carrots, soy sauce, shallots, garlic cloves	fish sauce, beansprouts, rice paper, rice noodles, thai basil	12.7
Moroccan	821	salt, olive oil, ground cumin, onions, garlic cloves, ground cinnamon, water	couscous, ras el hanout, preserved lemon, saffron threads	12.9
British	804	salt, all-purpose flour, butter, milk, eggs, unsalted butter, sugar, onions	stilton cheese, suet, beef drippings, stilton, golden syrup	9.7
Filipino	755	salt, garlic, onions, water, soy sauce, pepper, oil, sugar, carrots, black pepper	fish sauce, calamansi juice, lumpia wrappers, calamansi, lumpia skins	10.0
Irish	667	salt, all-purpose flour, butter, onions, potatoes, sugar, baking soda, baking powder	Irish whiskey, Guinness Beer, irish cream liqueur, corned beef, irish bacon	9.2
Jamaican	526	salt, onions, water, garlic, ground allspice, pepper, scallions, dried thyme, black	scotch bonnet chile, jamaican jerk season, ackee, callaloo, jerk seasoning	12.2
Russian	489	salt, sugar, onions, all-purpose flour, sour cream, eggs, water, butter	sauerkraut, buckwheat flour, pierogi, dill, fresh dill, farmer cheese	10.2
Brazilian	467	salt, onions, olive oil, lime, water, garlic cloves, garlic, cachaca, sugar, tomatoes	cachaca, aai, manioc flour, palm oil, chocolate sprinkles	9.5

CLASSIFIERS USED

Supervised Learning

Model	Parameters	Input Data	Rationale
Logistic Regression	C=1, L2-penalty	A, A' reduced features	Known for multiclass, high features
Random Forest	N_estimators=100, depth=20, 100	A	Known for multiclass, high features
Gradient Boost	N_estimators=200, depth=20	A	Known for multiclass, high features

Unsupervised Learning

Model	Parameters	Input Data	Rationale
KMeans	n_components = 2, 3	A	Ingredient clustering
SVD	n_components = 2, 3	A	Ingredient clustering
Market Basket	min_support=.01, .005	A, E	Relationship btwn ingredients and cuisine
LDA	n_components = 2,6,15	A, B	Latent cuisines
BMF	n_components = 2,6,15	A	Latent cuisines
PCA	n_components = 2	B	Latent cuisines, distance btwn cuisines
Hierarchical clustering	method='complete'	D	Correlation (distance) btwn cuisines

ACKNOWLEDGEMENTS

We thank Barbara Engelhardt and the COS424 AI staff for a fantastic class, feedback and comments.

TF-IDF Supplementary

$$\mathrm{tf}(t,d) = \frac{1 \text{ if } t \text{ occurs in } d \text{ and 0 otherwise}}{\text{number of words in } d},$$

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

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D),$$

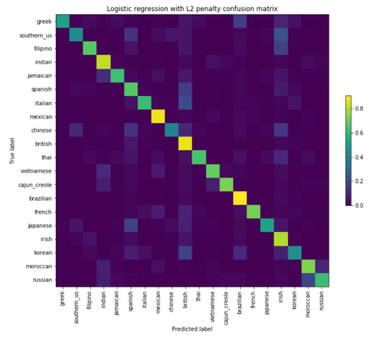
t = term
d = document
D = set of all documents

SUPERVISED LEARNING: Features selected matches signature ingredients, most predictive cuisines are dissimilar to others based on correlation analysis

Features Selected

Top Features	Matching Cuisine
Gochujang base	Korean
Cachaca	Brazilian
Cajun seasoning	Cajun_creole
Coconut milk	Indian
Corn starch	Chinese
Corn tortillas	Mexican
Couscous	Moroccan
Cumin seed	Indian
Feta cheese crumbles	Greek
Fish sauce	Thai
Garam masala	Indian
Grated parmesan cheese	Italian

Log Reg Accuracy: .77 Overall



- Used SVD to extract 1000 top features, accuracy: .77
- On original 6714 features, run time took too long
- Best performing cuisines: Brazilian, Mexican, followed by British, Indian, Irish
- Could be due to large number of recipes + unique ingredients

	precision	recall	f1-score	support
greek	0.78	0.53	0.63	99
southern_us	0.55	0.45	0.49	155
filipino	0.76	0.67	0.71	326
indian	0.79	0.82	0.80	504
jamaican	0.73	0.63	0.68	142
spanish	0.59	0.67	0.63	541
italian	0.77	0.62	0.69	238
mexican	0.87	0.89	0.88	604
chinese	0.64	0.41	0.50	141
british	0.79	0.88	0.83	1542
thai	0.83	0.65	0.73	106
vietnamese	0.78	0.69	0.73	294
cajun_creole	0.84	0.72	0.78	156
brazilian	0.90	0.91	0.90	1280
french	0.84	0.72	0.77	184
japanese	0.69	0.53	0.60	97
irish	0.71	0.80	0.75	880
korean	0.55	0.46	0.50	186
moroccan	0.78	0.74	0.76	311
russian	0.71	0.62	0.66	169
avg / total	0.77	0.77	0.77	7955

UNSUPERVISED LEARNING:

Ingredient Clustering

SVD on BOW (informed by KMeans)

KMeans on BOW partitions by common ingredients

K	Ingredients
2	Salt
3	Salt, onions
4	Salt, onions, olive oil
5	Salt, onions, olive oil, soy sauce

Market Basket Analysis on BOW

Finds relationship btwn ingredients

Antecedents	Consequents	Confidence
Onions, carrots, pepper	Salt	1
Baking powder, white sugar, eggs	All-purpose flour	.98
Baking pwder, white sugar, all-purpose flour	Eggs	.94

Antecedents	Consequents	Lift
Active dry yeast	Warm water	39.1
Baking soda	Baking pwder, buttermilk, salt	29.2
Clove	Cinnamon sticks	22.9
Garlic Powder	Onion powder	20.4
Salt	Dijon mustard	1.0
Olive oil	Garlic, water	1.0

Finds relationship btwn ingredients and cuisines (On BOW + Cuisine OHC)

Antecedents	Consequents	Confidence
Corn tortillas	Mexican	.98
salsa	Mexican	.95
Garam Masala	Indian	.93

Latent Cuisines

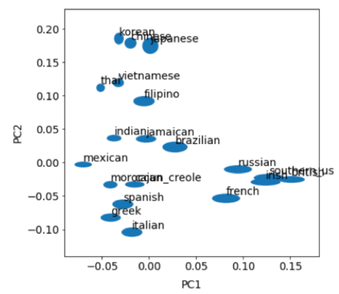
LDA on TF-IDF recovers Latent Cuisines

LDA (on dataset B) (n=6)	Latent Cuisine
Soy sauce, sesame oil, fish sauce, rice vinegar, scallions, green onions, sugar	Korean
Avocado, jalapeno chilies, fresh lime juice, chopped cilantro, purple onion, lime, white onion	Mexican
All-purpose flour, buttermilk, baking powder, milk, warm eggs, baking soda	Southern US
Ground cumin, curry powder, ground coriander, ground cinnamon, chickpeas, ground ginger, olive oil	Indian
extra-virgin olive oil, fresh lemon juice, olive oil, garlic cloves, purple onion, ground cumin, ground black pepper	Italian

BMF on BOW recovers both ingredients and cuisines

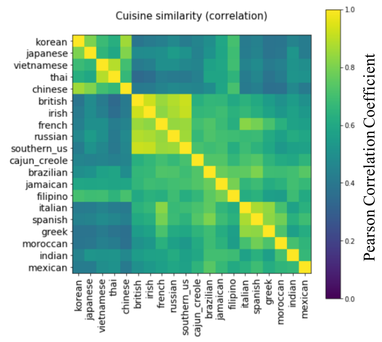
Sugar, all-purpose flour, large eggs, unsalted butter, butter, baking powder, milk	Southern US
Onions, garlic, tomatoes, ground cumin, chili powder, carrots, vegetable oil	Indian
Olive oil, garlic cloves, ground black better, kosher salt, extra-virgin olive oil, grated parmesan cheese, purple onion	Italian
water	-
Soy sauce, sesame oil, green onions, garlic, sugar, vegetable oil, scallions	Asian

PCA on TF-IDF recovers similarity btwn cuisines

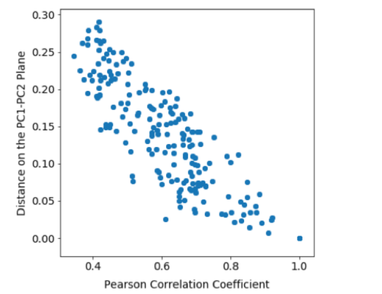


Cuisines Correlations

Hierarchical Clustering on Cuisine TF-IDF finds correlated groups of cuisines



PCA on TF-IDF Ingredients matches HC correlation

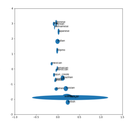


CONCLUSIONS:

- **Supervised and unsupervised learning extracts similar features, which match "top 10 signature ingredients" by cuisine**
- **Supervised learning performed best on cuisines that are most dissimilar to others based correlation analysis**
- **Latent topic models can extract cuisines**
- **KMeans partitions groups based on most common ingredients**
- **PCA on TF-IDF ingredients yields similar cuisine similarity as Hierarchical Clustering directly on TF-IDF cuisines**

Questions for TAs/Prof:

- **Do we need LL, BIC metrics for unsupervised learning, given findings are so different?**
- **Why does standardizing data make PCA results look weird? (bc of binary nature of data?)**
- **How to look at Market Basket Analysis to extract relevant information?**



Works Cited

1. Kaggle "What's Cooking" Dataset. https://www.kaggle.com/c/whats-cooking/data
2. SciKit Learn Python Toolbox