An Examination of Recipes from Around the World

Benjamin W. Sturm
bwsturm@gmail.com
June 1, 2018

The goal of my project was to use data science to gain insights of cuisines from around the world.

Questions I wanted to explore:

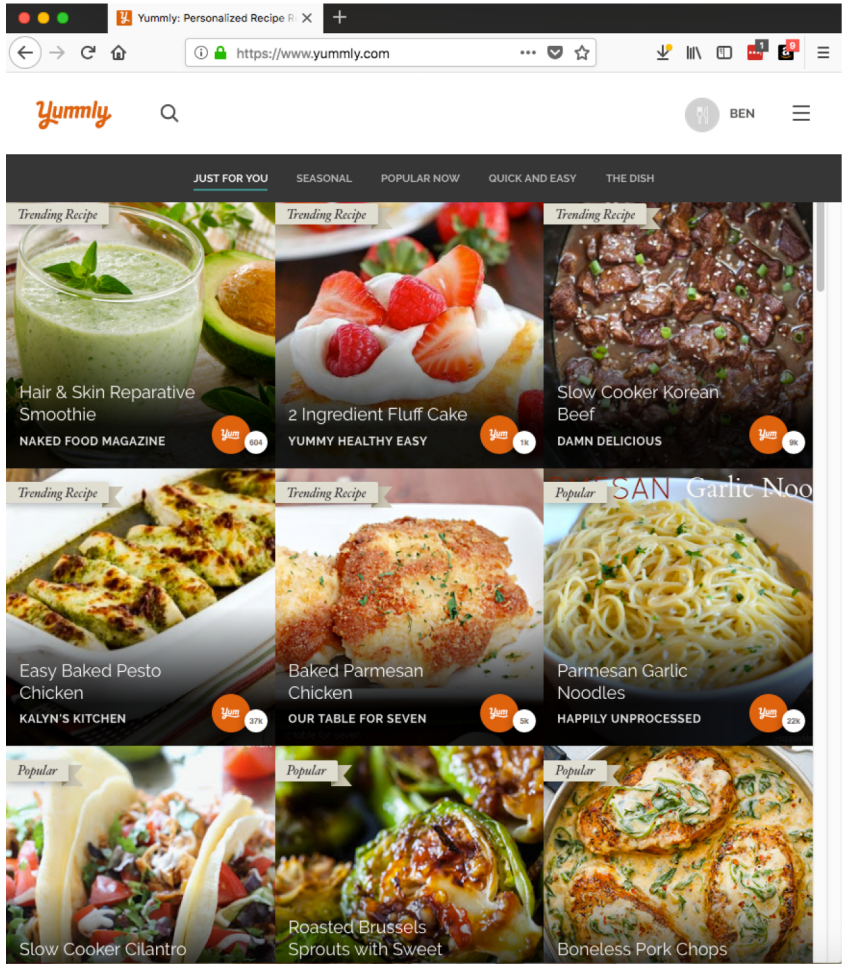
Which cuisines are most similar and which are most different?

Are there any unexpected relationships?

Can we use topic modeling to separate out different cuisines?

In order to explore this topic, I used publicly available recipe data

Yummly.com API



Data Description

Name	Type	Description
Cuisine	String	e.g. italian, indian, mexican
Course	String	e.g. Lunch, Main Dish
Flavors	Dict	e.g. {sour: float, salty: float, sweet: float}
Ingredients	List of Strings	e.g. [bow-tie pasta, bacon slices]
Rating	Int	Rating from 1-5
RecipeName	String	e.g. 'Lemon Chicken Pasta'

Size of data:

Total number of recipes downloaded: 12,492

Total number of cuisines: 25

Number of recipes / cuisine: ~500

Before implementing machine learning, a number of preprocessing steps were necessary

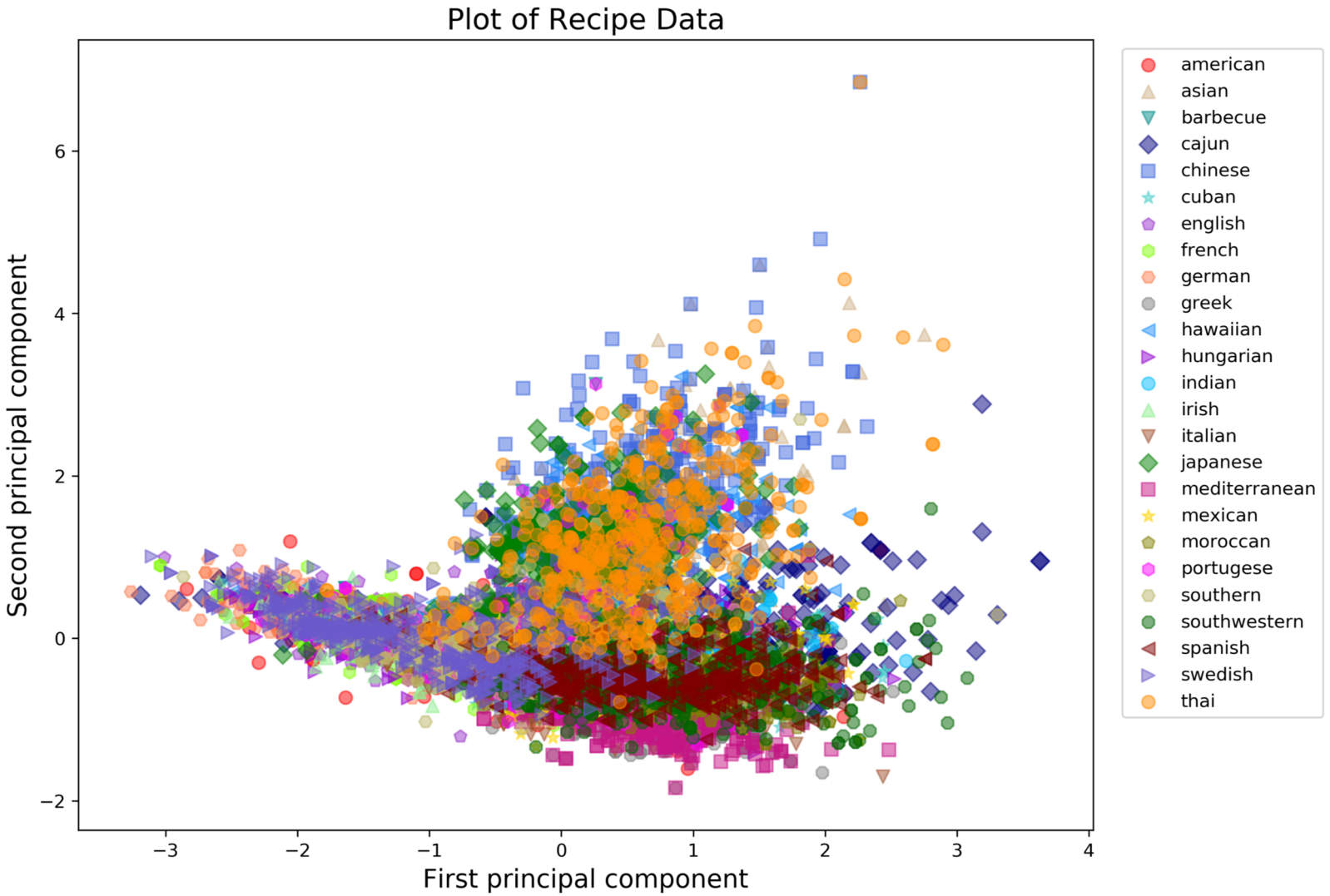
Example of the dataframe

	cuisine	course	ingredients	bitter	meaty	piquant	salty	sour	sweet	rating	recipe_name
9988	japanese	NaN	[pork belly, shoyu, mirin, sake, sugar, scallions, garlic, shallots, ginger, salt]	0.333333	0.833333	0.000000	0.833333	0.166667	0.333333	3	japanese chashu pork belly (for ramen)
9989	japanese	[Condiments and Sauces]	[light brown sugar, mirin, reduced sodium soy sauce]	0.833333	0.166667	0.000000	0.833333	0.000000	0.833333	3	canal house teriyaki sauce
9990	japanese	[Breakfast and Brunch, Lunch]	[fresh spinach, spinach, onions, garlic cloves, large eggs, salt, black pepper, soy sauce, sugar, olive oil]	0.833333	0.166667	0.000000	0.666667	0.833333	0.166667	4	spinach tamagoyaki (spinach packed omelette)
9991	japanese	[Main Dishes]	[pork shoulder, soy sauce, mirin, sake, sugar, garlic, green onions, ginger, shallots]	NaN	NaN	NaN	NaN	NaN	NaN	4	slow braised japanese chashu pork
9992	japanese	[Side Dishes]	[gai lan, cooking oil, fresh ginger, garlic, hot pepper, miso paste, water, toasted sesame oil, soy sauce]	0.500000	0.166667	0.166667	0.333333	0.833333	0.166667	5	chinese broccoli with garlicky ginger miso

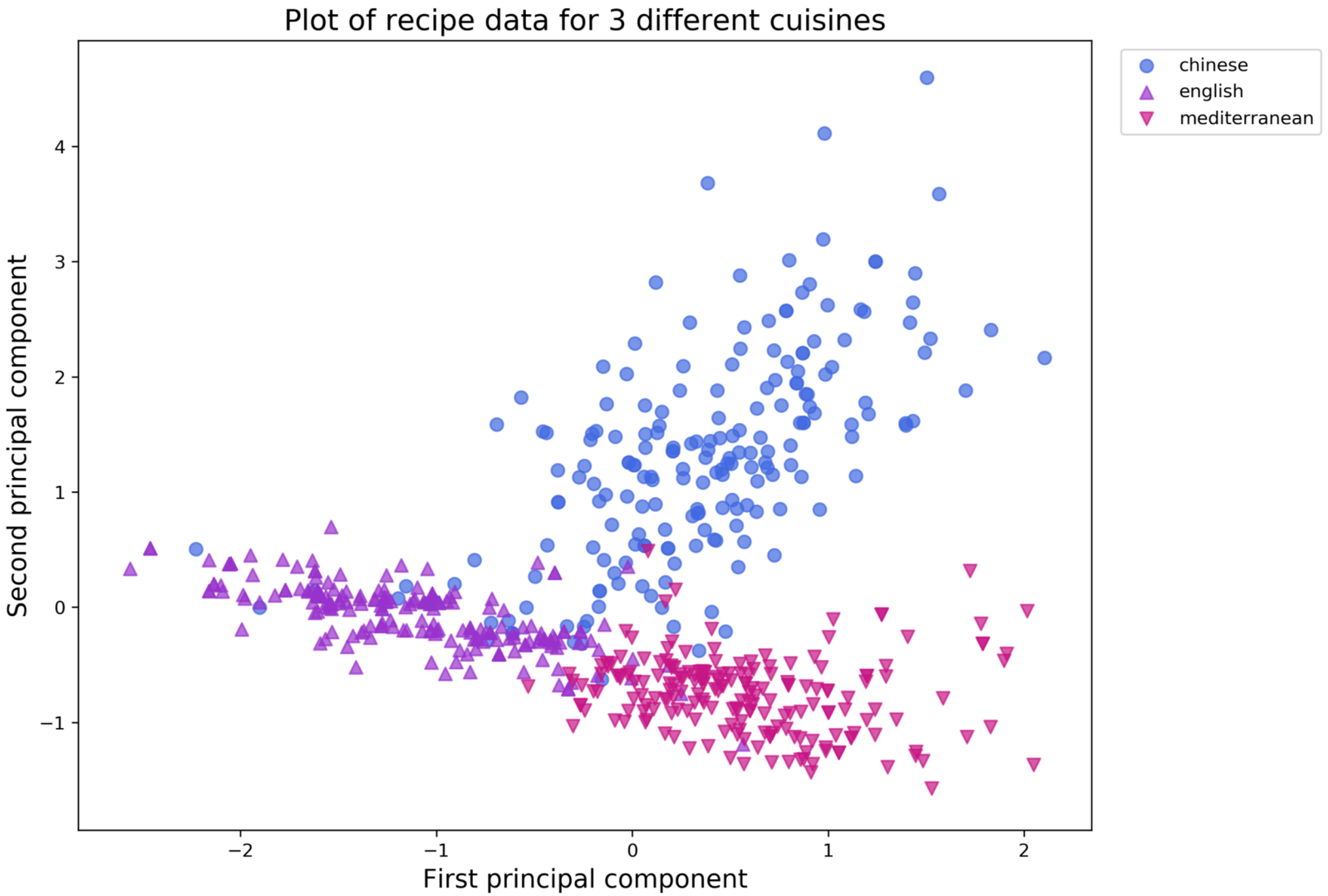
Summary of preprocessing steps:

- **Hyphenating certain ingredients (e.g. olive-oil, corn-starch)**
- **Removing stop words and other frequently occurring words (e.g. salt, pepper, water)**
- **Stemming by dropping plural forms of words and other suffixes**
- **Bag-of-words processing**

Applying PCA to the recipe data results in many overlapping points

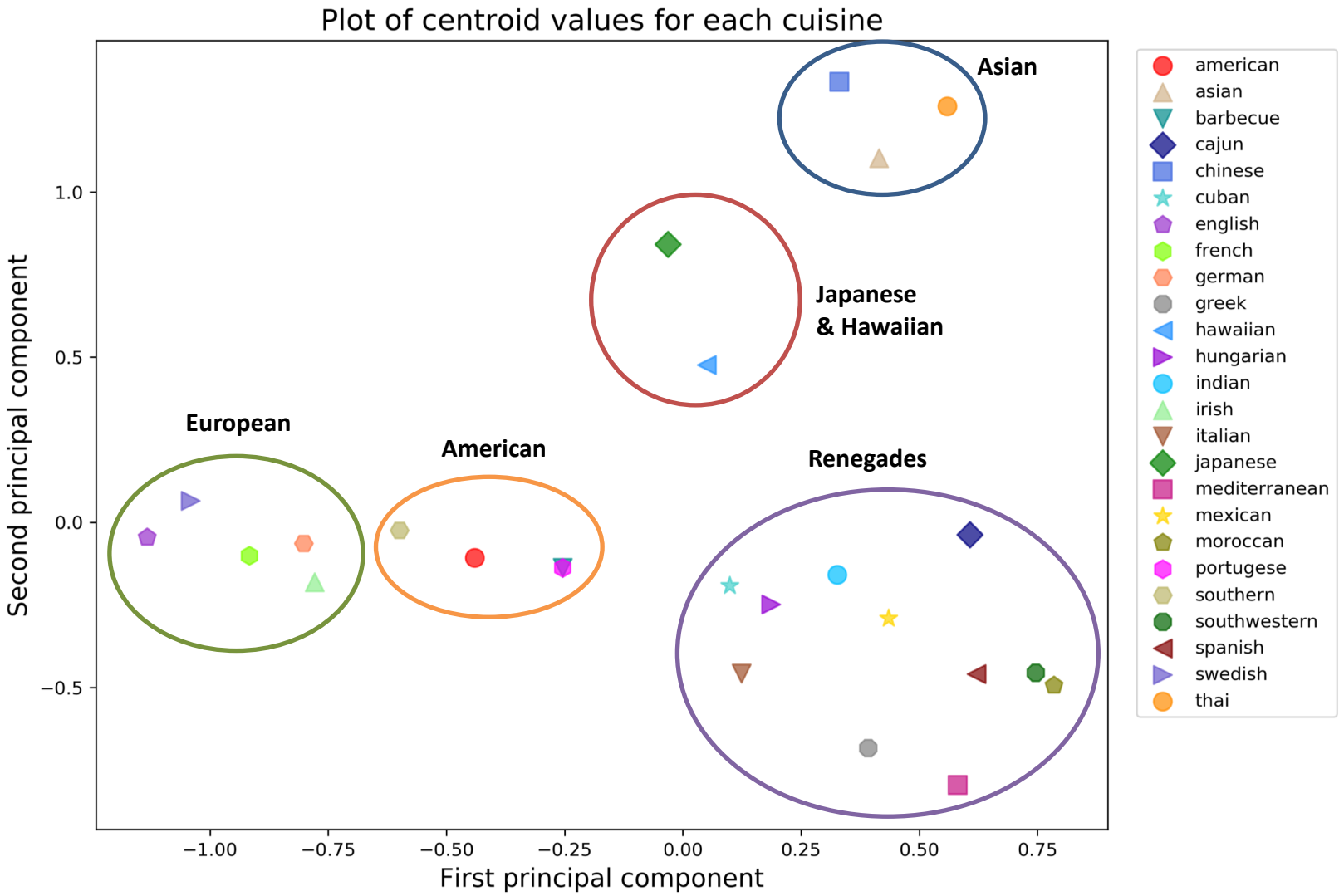


A much clearer separation is observed when we focus on 3 different cuisines



Cocktails

Now observing the centroids for each of the cuisines, we can see some patterns



Results from LDA topic modeling showed that ingredients were grouped partially based on cuisine, but also based on category

topic 0	topic 1	topic 2	topic 3	topic 4	topic 5	topic 6	topic 7	topic 8	topic 9	topic 10	topic 11	topic 12
sauc	cream	flour	dri	potato	chicken	vinegar	olive-oil	cumin	tomato	pineappl	chili	lime
soy	beef	butter	oregano	onion	broth	appl	wine	garlic	feta	ham	cilantro	mint
sugar	egg	all-purpos	garlic	olive-oil	sodium	cider	garlic	ginger	olive-oil	mix	onion	ice
garlic	onion	sugar	powder	bacon	low	sugar	white	coriand	chees	chees	powder	rum
ginger	sour	egg	olive-oil	kosher	onion	peanut	parsley	turmer	cucumb	pickl	lime	juic
corn-starch	flour	milk	onion	butter	thigh	fillet	leaf	onion	onion	instant	cumin	sugar
rice	butter	baking-powd	thyme	egg	paprika	rice	vinegar	olive-oil	oliv	bread	garlic	soda
vinegar	bread	baking-soda	paprika	garlic	rice	butter	dri	tomato	garlic	butter	bean	leav
onion	parsley	unsalt	flake	sweet	breast	roast	flat	paprika	crumbl	roll	jalapeno	white
sesam	broth	larg	pork	larg	olive-oil	salmon	onion	lemon	lemon	soup	tomato	whiskey

topic 13	topic 14	topic 15	topic 16	topic 17	topic 18	topic 19	topic 20	topic 21	topic 22	topic 23	topic 24
mustard	chees	cinnamon	sugar	seed	sauc	chees	lemon	chicken	onion	sauc	coconut
dijon	cheddar	almond	cream	beef	soy	parmesan	juic	breast	season	chees	lime
mayonnais	cream	appl	egg	onion	onion	grate	yogurt	broccoli	bell	tortilla	sauc
syrup	shred	butter	vanilla	tomato	rice	mozzarella	orang	halv	tomato	corn	fish
mapl	tortilla	slice	butter	carrot	sesam	sauc	greek	sauc	garlic	lettuc	milk
vinegar	sour	sugar	extract	leav	garlic	olive-oil	plain	floret	shrimp	shred	cilantro
boil	flour	golden	heavi	garlic	ginger	garlic	zest	cook	sauc	onion	curri
balsam	cook	sauerkraut	milk	caraway	carrot	basil	dill	prepar	rice	jack	past
low-fat	fat	raisin	chocol	paprika	scallion	shred	clove	breadcrumb	celeri	bean	sugar
squash	half	nutmeg	powder	cabbag	veget	egg	cucumb	style	sausag	enchilada	garlic

Desserts

Cocktails

Sauces

Italian

Thai

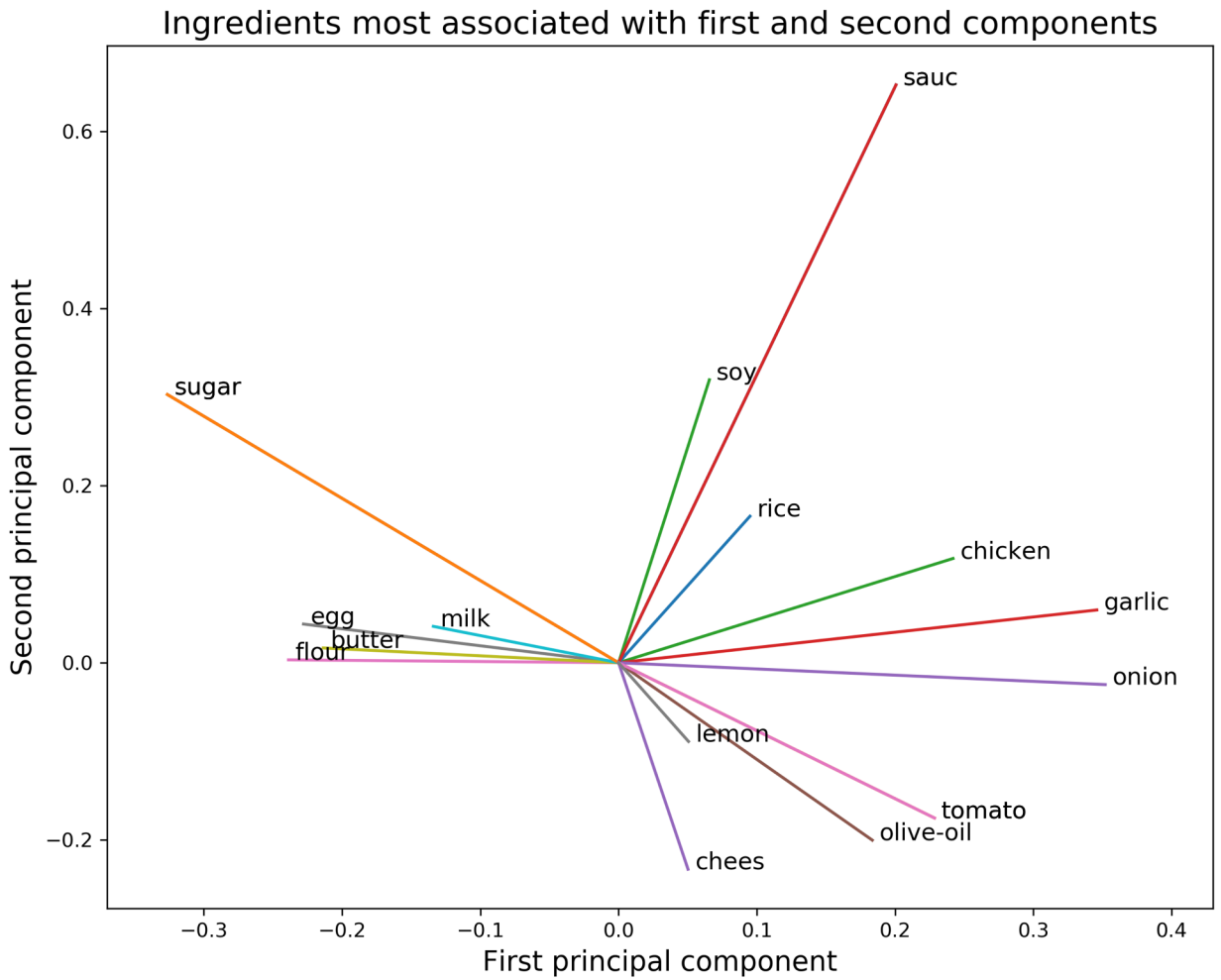
Conclusions

PCA was applied to the recipe data in order to help us better visualize the relationship between different cuisines.

We were able to observe clear separations between different types of cuisines when plotting the centroid values along the first two principal components

Topic modeling using LDA was only partially effective at grouping ingredients based on cuisine

Appendix: A plot demonstrating the ingredients most strongly associated with the first two principal components



Appendix: The goal of my project was to use data science to gain insights of cuisines from around the world.

Questions I wanted to explore:

Which cuisines are most similar and which are most different?

Are there any unexpected relationships?

Can we use topic modeling to separate out different cuisines?