Recipe2Cuisine

Celia Eddy

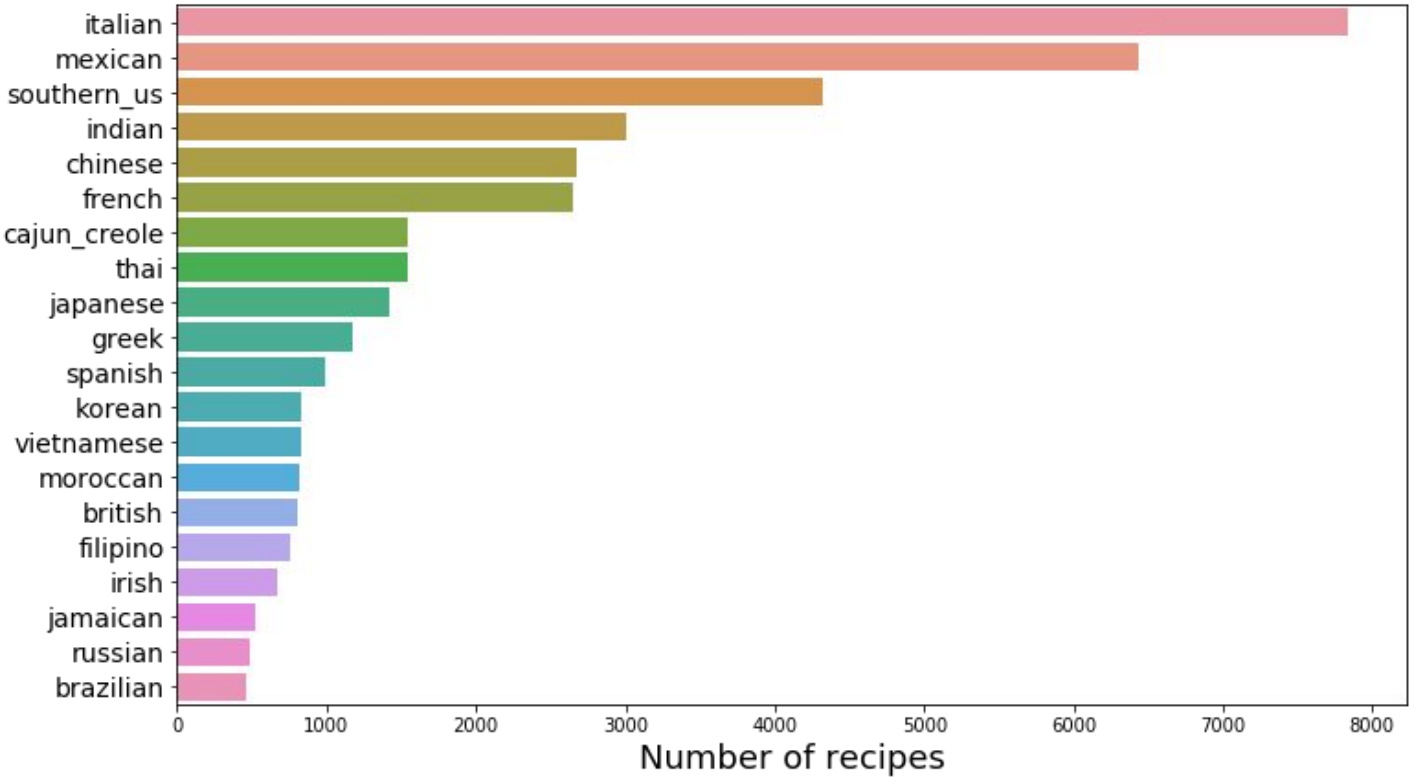
Context

- Food publication - goal for users to search for recipes by cuisine type
  - E.g. search “Italian” and get a list of Italian recipes
- Most recipes are unlabeled
- Small subset of 10,000 recipes labeled by cuisine
- Can we leverage this data to make predictions of cuisine based on recipe ingredients?

Business Questions

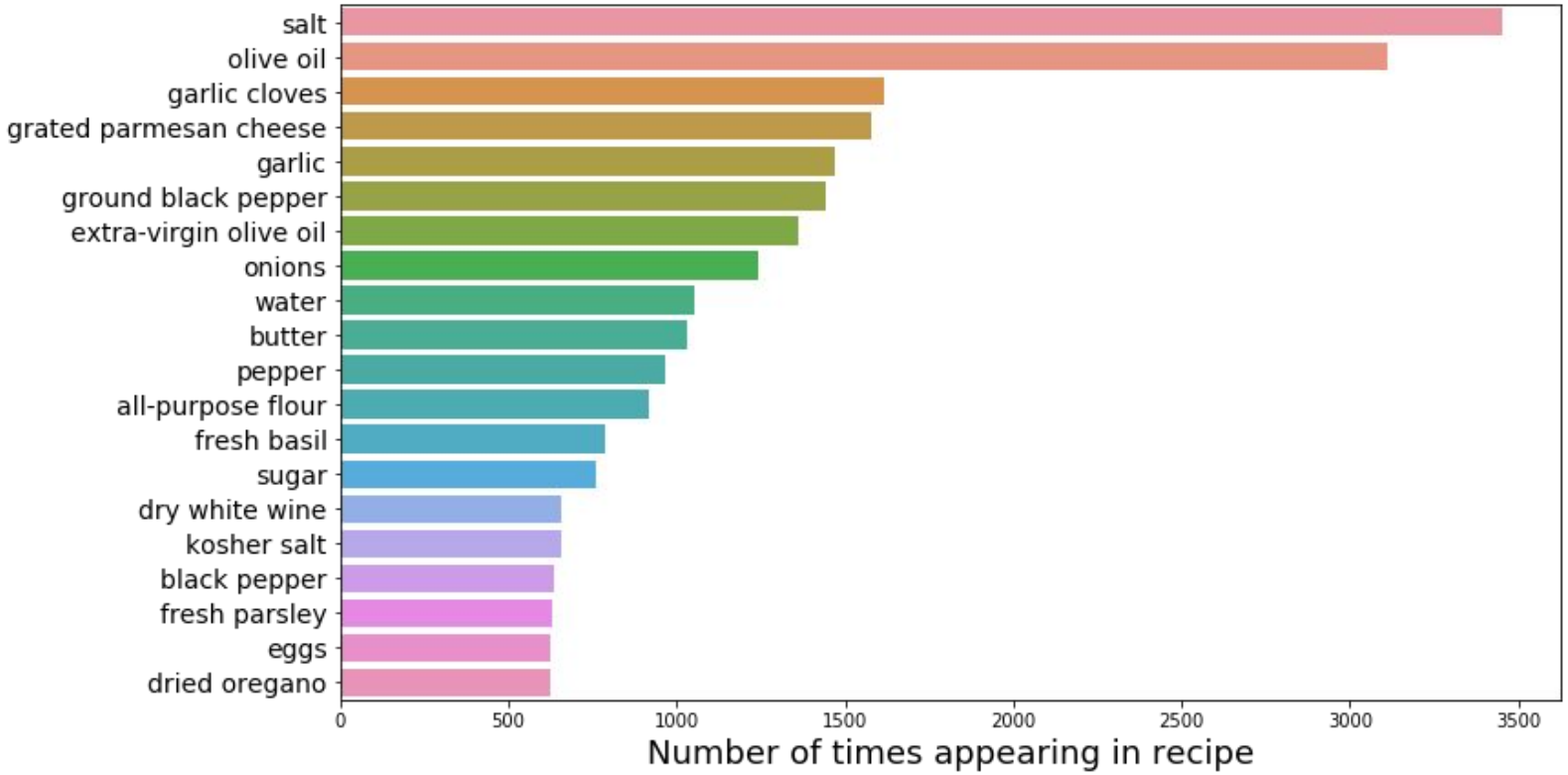
- Design a method to predict the cuisine of a recipe given its ingredients
  - Make it robust enough to understand similarities/substitutions between ingredients
- What are the main ingredients that characterize major cuisines?
- Write a guideline for an outsourced team to hand-label the remaining corpus

Cuisines

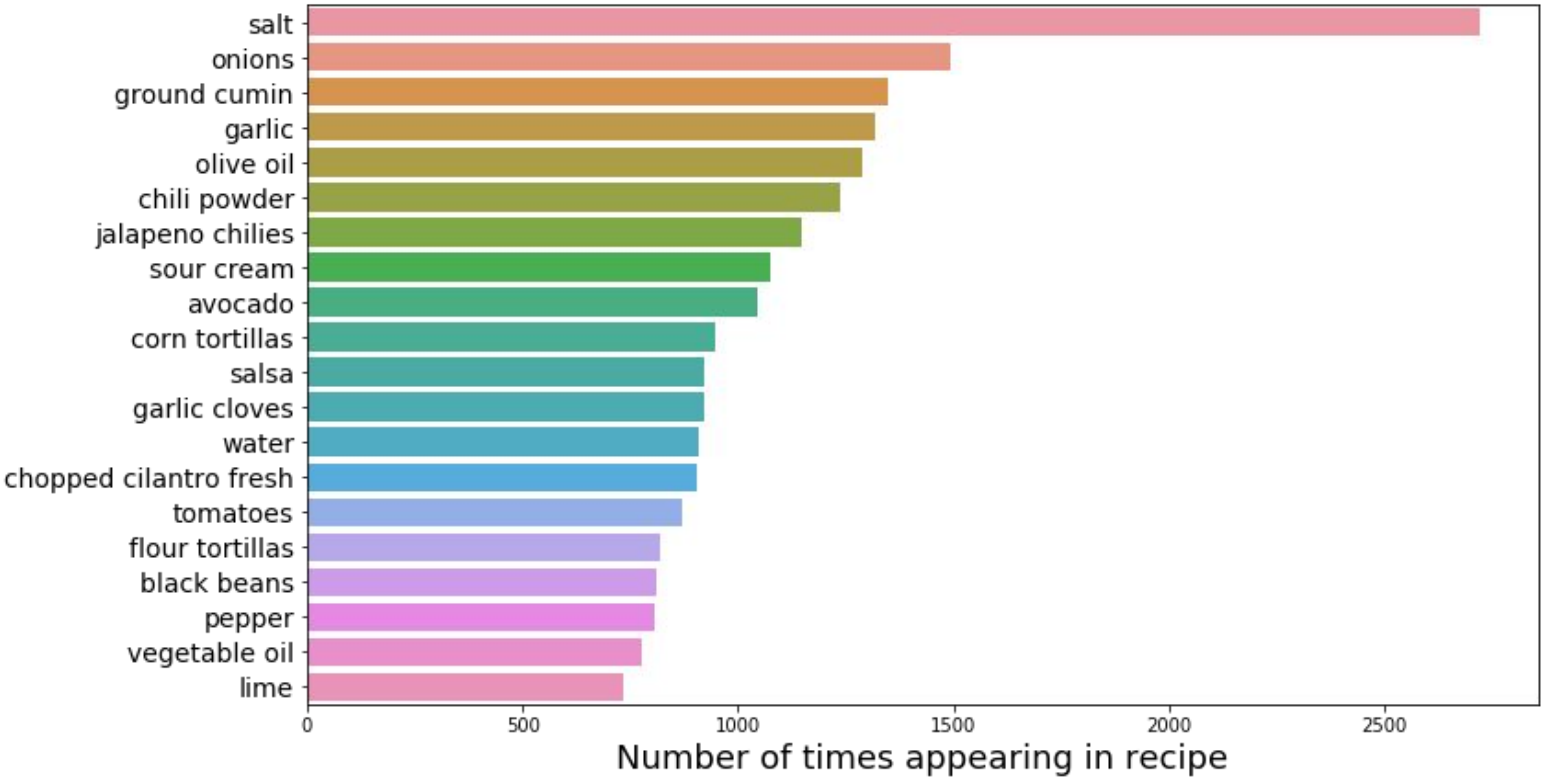

italian
mexican
southern_us
indian
chinese
french
cajun_creole
thai
japanese
greek
spanish
korean
vietnamese
moroccan
british
filipino
irish
jamaican
russian
brazilian
0
1000
2000
3000
4000
5000
6000
7000
8000
Number of recipes

What are the ingredients that characterize major cuisines?

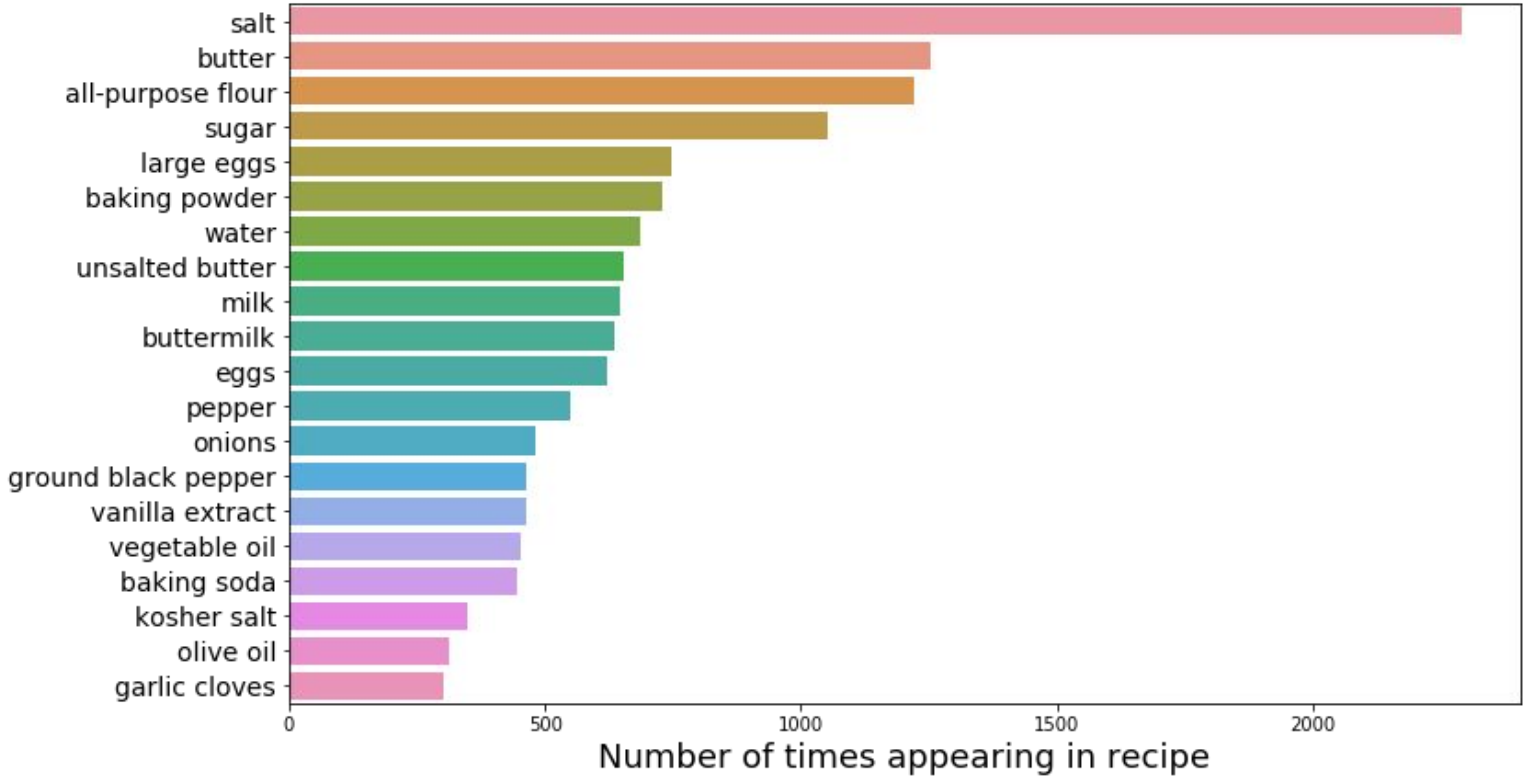
Ingredients: Italian



Ingredients: Mexican



Ingredients: Southern



Water, salt, garlic, olive oil...

Calculate Tf-Idf matrix

Each “document” consists of all the ingredients from all the recipes within one cuisine

→ 20 “cuisine documents”

Water, salt, garlic, olive oil...

Calculate Tf-Idf matrix, find top terms for each cuisine

Italian: 'lasagna', 'parmigiano', 'part', 'arborio', 'prosciutto', 'marinara', 'romano', 'pecorino', 'pesto', 'provolone'

Mexican: 'taco', 'enchilada', 'mexican', 'refried', 'tortilla', 'tomatillos', 'poblano', 'guacamole', 'queso', 'cotija'

Southern: 'grits', 'whiskey', 'bourbon', 'collard', 'eyed', 'quickcooking', 'pecan', 'cajun', 'biscuits', 'barbecue'

The ingredients characterizing each cuisine can be found from the top terms in each “cuisine document’s” Tf-Idf matrix

Predicting the cuisine of a recipe

Features

- Tf-Idf matrix (each document is a recipe)

- How to make robust for similarities/substitutions between ingredients?
  - Train word2vec model on ingredients

Most similar ingredients

Garlic:

```
[('minced garlic', 0.8619658946990967),
 ('garlic cloves', 0.8069359064102173),
 ('chopped garlic', 0.7982699871063232),
 ('crushed garlic', 0.7048181295394897),
 ('large garlic cloves', 0.6887021064758301),
 ('garlic puree', 0.5056877136230469),
 ('clove garlic, fine chop', 0.504534125328064),
 ('garlic powder', 0.46947526931762695),
 ('roasted garlic', 0.4611283540725708),
 ('squash', 0.4175143241882324)]
```

Yellow corn meal:

```
[('cornmeal', 0.8410243988037109),
 ('white cornmeal', 0.7736024856567383),
 ('stone-ground cornmeal', 0.6534618139266968),
 ('saltines', 0.607382595539093),
 ('self-rising cornmeal', 0.6028713583946228),
 ('quickcooking grits', 0.5902442932128906),
 ('catfish fillets', 0.5886045098304749),
 ('all purpose unbleached flour', 0.5649878978729248),
 ('Bisquick Original All-Purpose Baking Mix', 0.5631225109100342),
 ('cornbread', 0.5535959005355835)]
```

Features

- Tf-Idf matrix (each document is a recipe)

- How to make robust for similarities/substitutions between ingredients?
  - Train word2vec model on ingredients

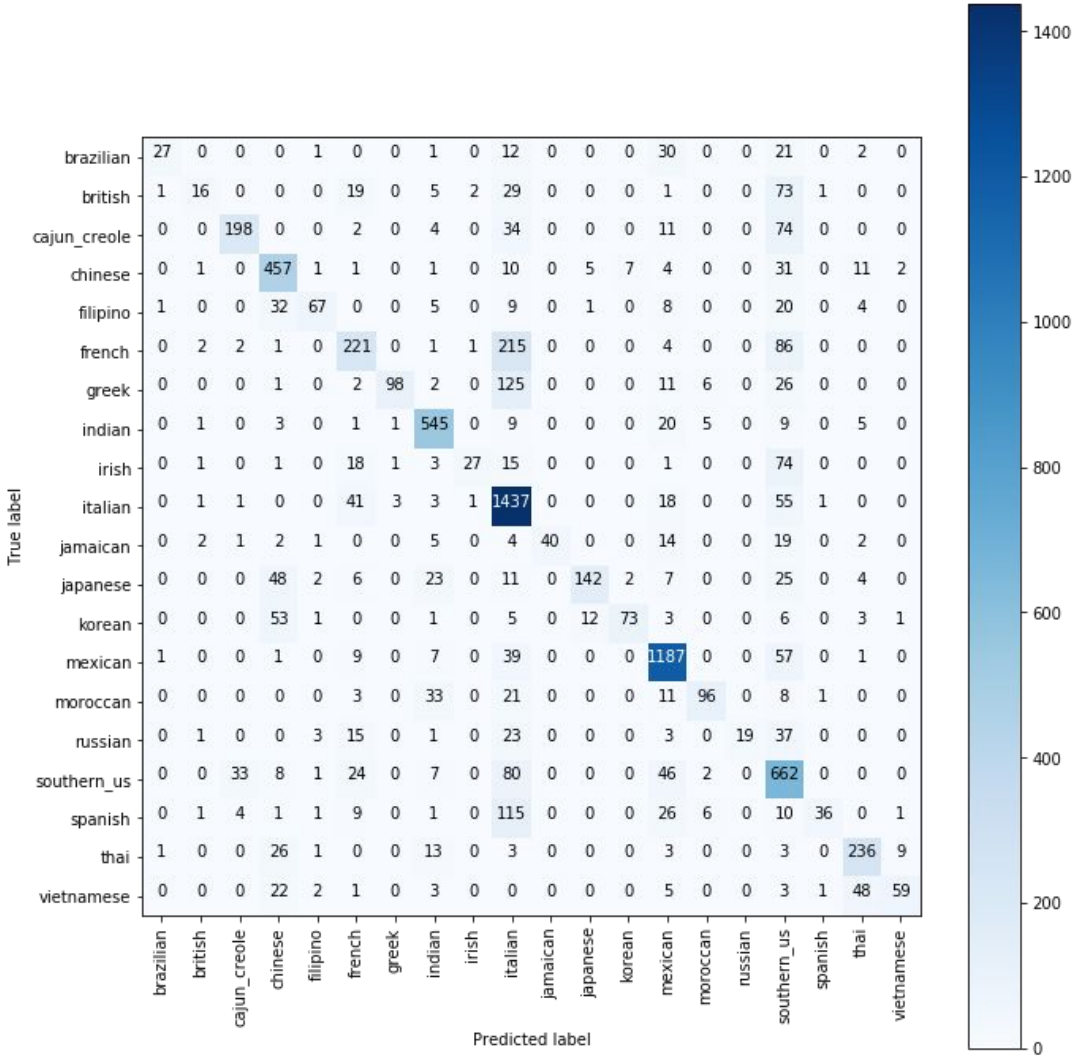
    → average vector embedding for each recipe

Modeling

Random Forest Classifier

Precision = 75%

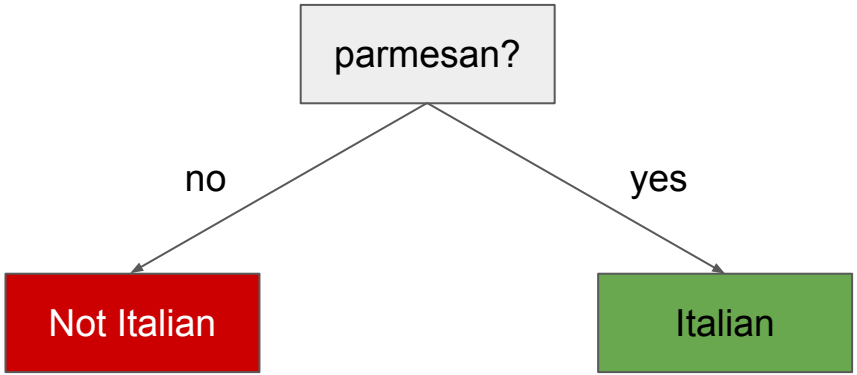
Recall = 71%



Can accurately predict the cuisine of a recipe 75% of the time

Guide for hand-labeling recipes

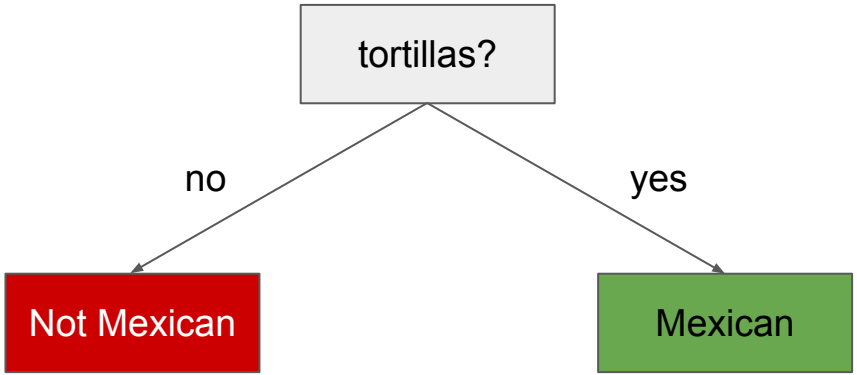
Decision tree for each cuisine



85% Accuracy!

With a decision tree of depth 8, can predict Italian cuisine with 87% accuracy

Decision tree for each cuisine



89% Accuracy!

With a decision tree of depth 8, can predict Mexican cuisine with 92% accuracy

Simple decision trees could form the basis of a guide for hand-labeling recipes

(at least for popular cuisines)