Recipe Recommender

Praveena Suresh
NLP and Unsupervised Learning

Project Goal

The goal of this project is to build a recipe recommendation system based on a recipe that is chosen from the recipes dataset.

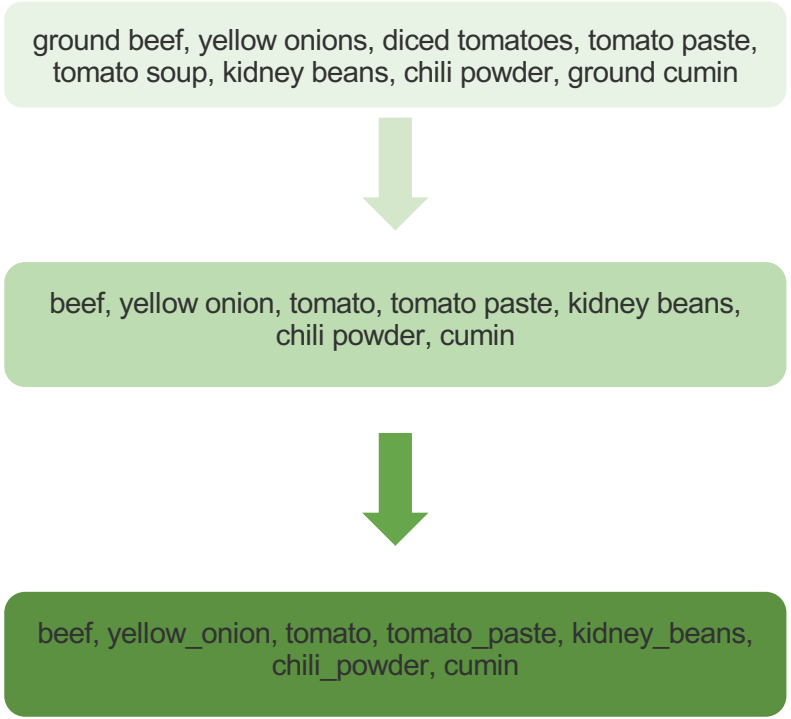
Data

The dataset is obtained from Kaggle[1] which includes 2 csv files. One is the interactions file containing recipe IDs rated by user IDs and the other is raw recipes csv file containing 230,000 recipes with names, ID, ingredients, description, steps etc.

Preprocessing

Custom Tokenization

Convert plural to singular and use underscore to keep the full ingredient names to get a better topic model



Workflow

Preprocessed Data

Pandas
Numpy
NLTK

Vectorize

Custom
Tokenization
and Vectorize

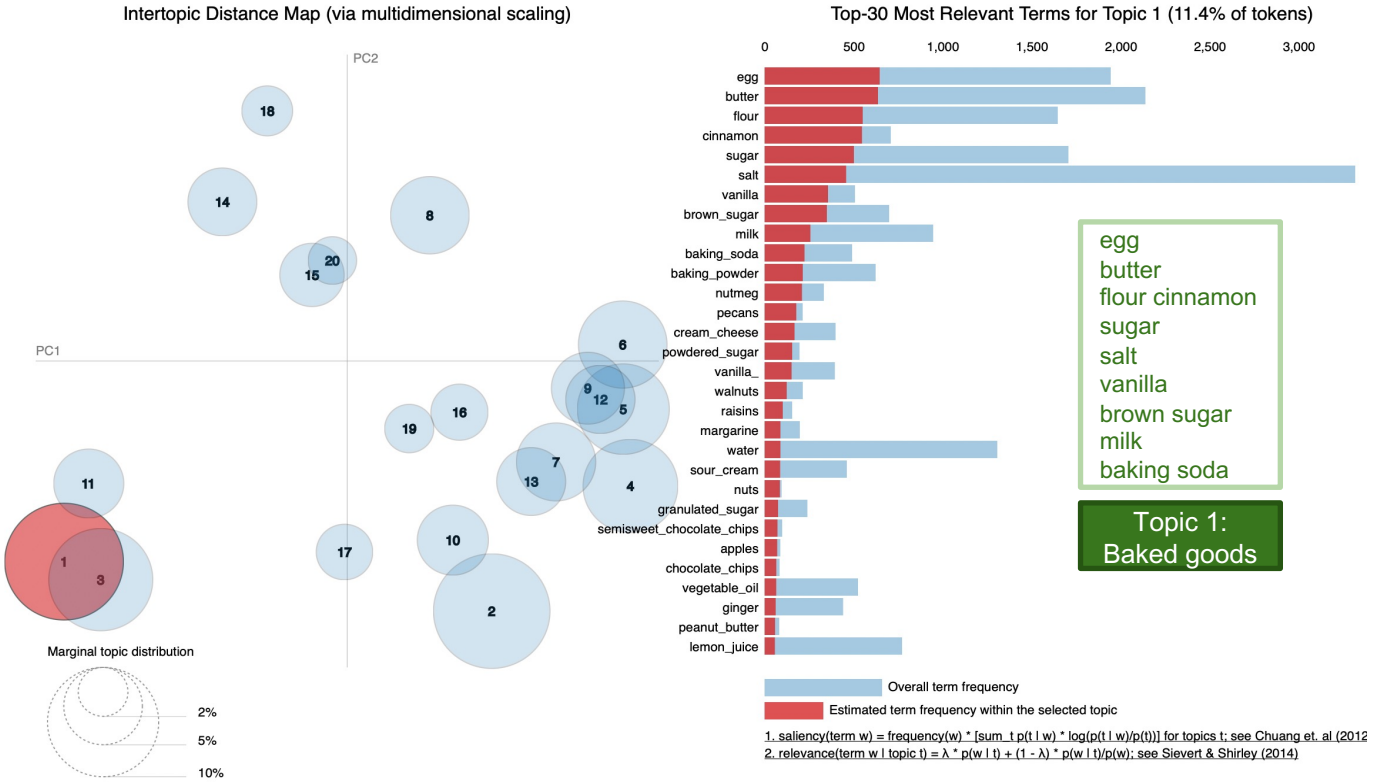
Topic Modeling

NMF
LDA

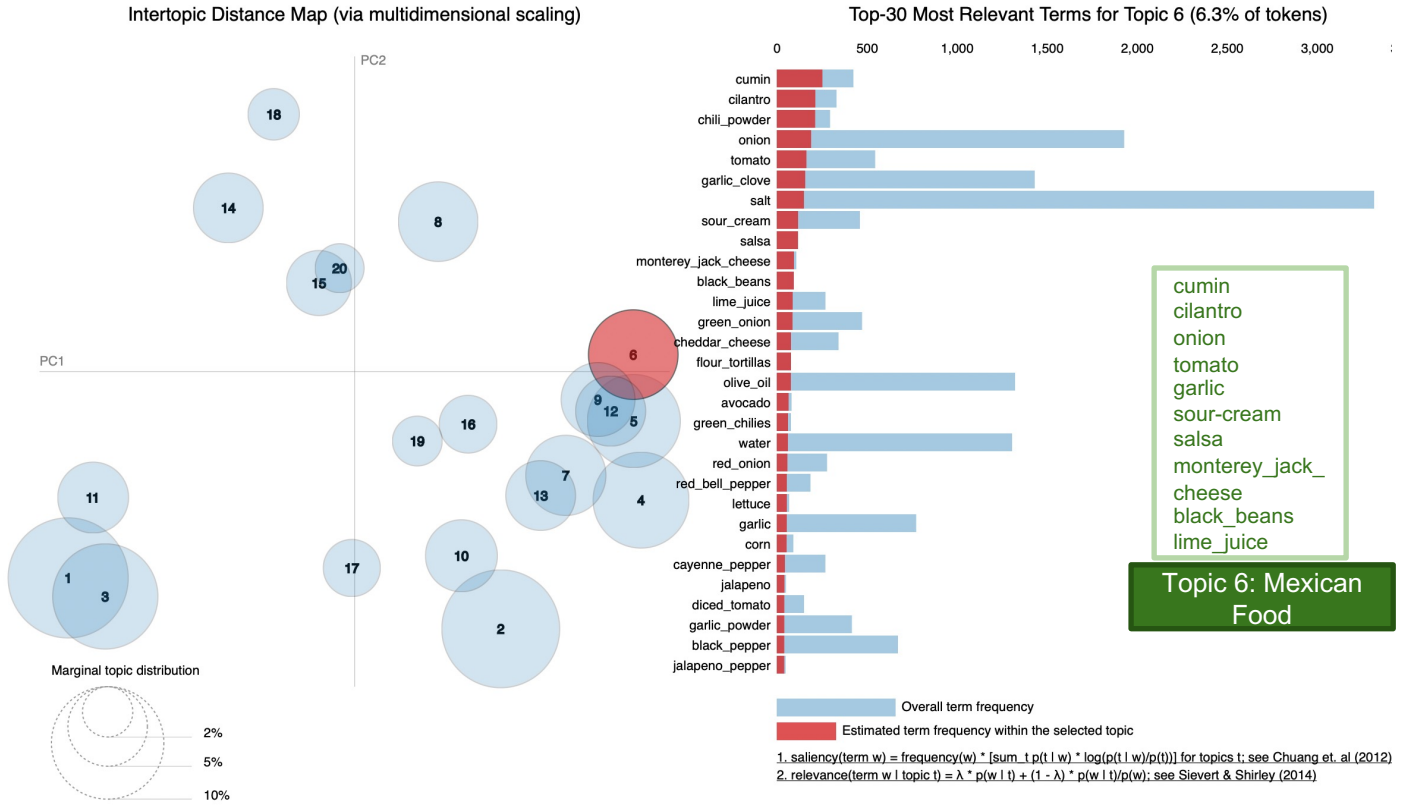
Recommender

Cosine
Similarity

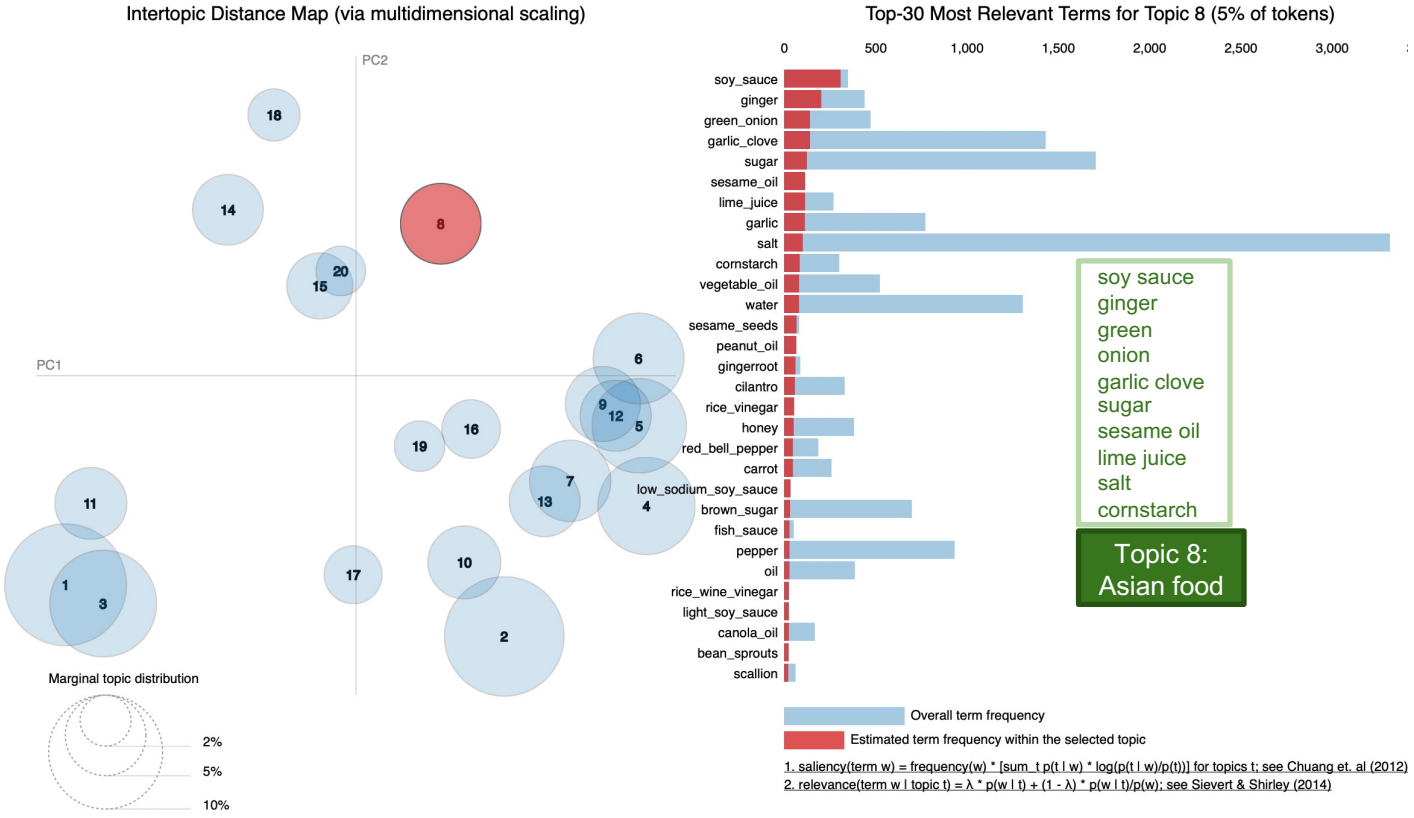
Interactive Topic Visualization with pyLDavis



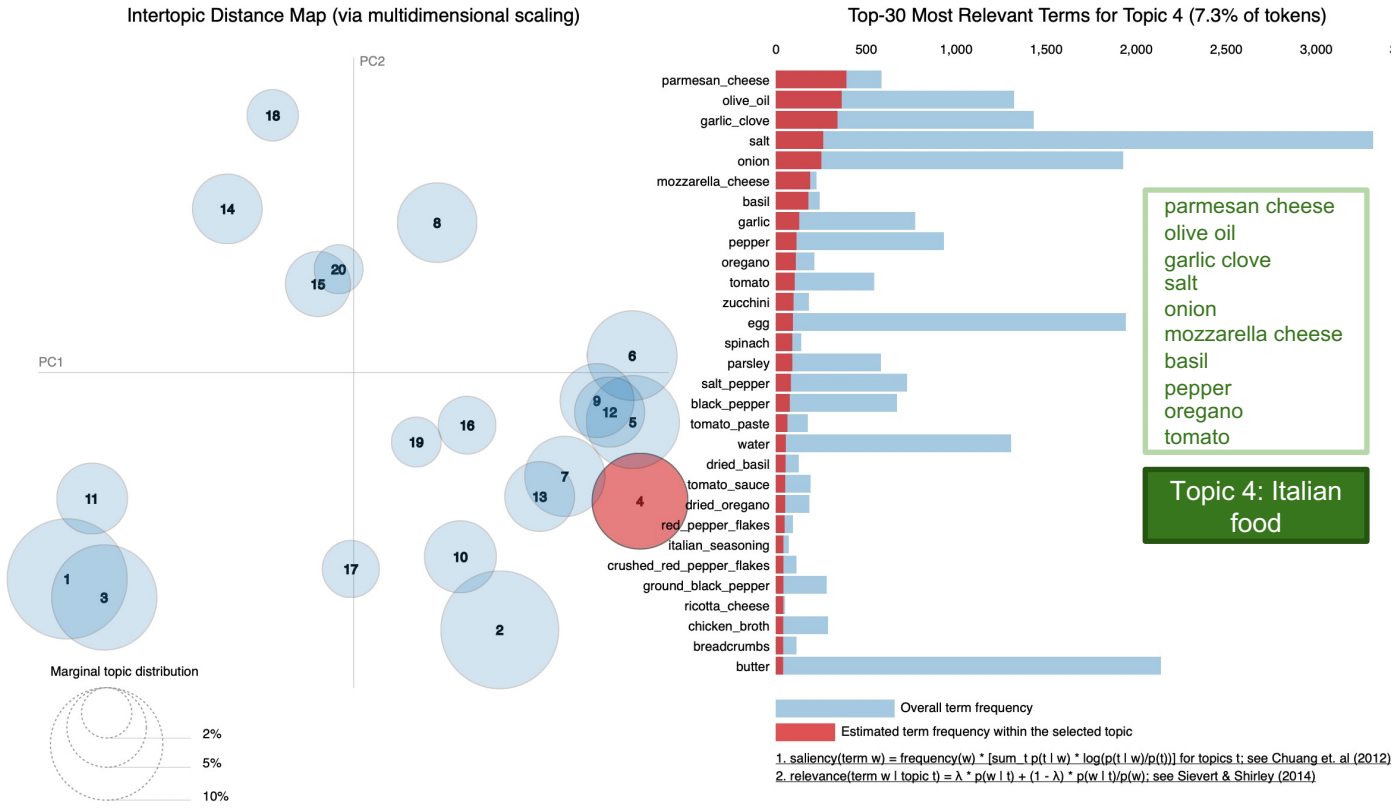
Interactive Topic Visualization with pyLDavis



Interactive Topic Visualization with pyLDavis



Interactive Topic Visualization with pyLDavis



Recommender

baby green salad

asian pear salad
blackened steak salad
roasted beet salad with
raspberry balsamic vinaigrette

Recommender

classic hummus

lemon dill hummus light
detoxifying hummus creamy
roasted garlic hummus

Future Work

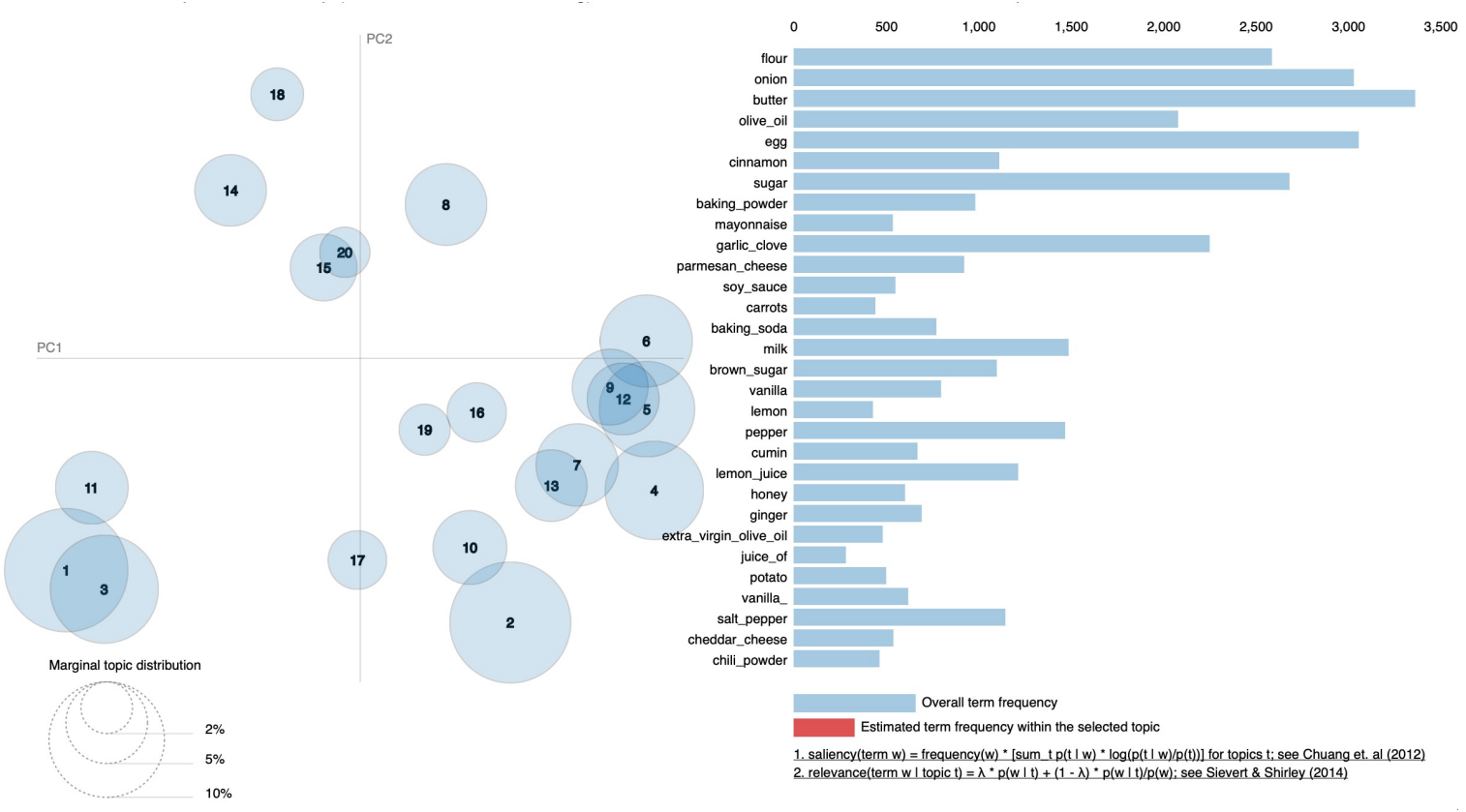
1. Improve tokenization by more NLP pre-processing.
2. Better model selection by tuning hyperparameters
3. Build a recommender system using collaborative filtering

Appendix

1. Food.com Recipes and Interactions Kaggle dataset
https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions

2. Evaluate Topic Models: Latent Dirichlet Allocation (LDA)
https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Appendix


PC2
PC1
18
14
8
20
15
6
9
12
5
19
16
7
13
4
11
1
3
17
10
2
Marginal topic distribution
2%
5%
10%
0
500
1,000
1,500
2,000
2,500
3,000
3,500
flour
onion
butter
olive_oil
egg
cinnamon
sugar
baking_powder
mayonnaise
garlic_clove
parmesan_cheese
soy_sauce
carrots
baking_soda
milk
brown_sugar
vanilla
lemon
pepper
cumin
lemon_juice
honey
ginger
extra_virgin_olive_oil
juice_of
potato
vanilla_
salt_pepper
cheddar_cheese
chili_powder
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Thanks!

Do you have any questions?