Application of Data Mining for Food Recommendation

Akshit Jain
Naga Santhosh Kartheek Karnati
Samar Dikshit

Objective

- People usually consume 3 meals per day and have 2 important questions to answer repeatedly which is tiring and monotanous.
  - *“What do we make, and what ingredients do we need to make it?”*

- **Our solution:** a system that recommends recipes and ingredients on the fly.

The Dataset

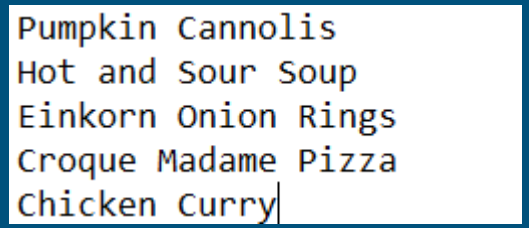
- Source: http://www.ub.edu/cvub/recipes5k/
- 4,826 recipes to prepare 101 dishes using 3,213 *raw* ingredients.
- The dataset also contained images for each recipe grouped by dishes.

```
whole milk ricotta cheese,pumpkin pur\u00e9e,Dixie Crystals Confectioners Powdered Sugar,pumpkin pie spice,cannoli shells
black fungus,enokitake,shiitake,lily buds,bamboo shoots,red wine vinegar,white vinegar,soy sauce,corn starch,chicken broth,firm tofu,
yellow onion,flour,baking powder,seasoning salt,large eggs,milk,gluten free panko breadcrumbs,coconut oil,salt
shredded cheddar cheese,grated Gruy\u00e8re cheese,sour cream,dijon mustard,worcestershire sauce,salt,freshly ground black pepper,fla
oil,potatoes,carrots,onions,garlic,chicken,fish sauce,coconut milk,water,green bell pepper,red bell pepper,curry powder,salt
```

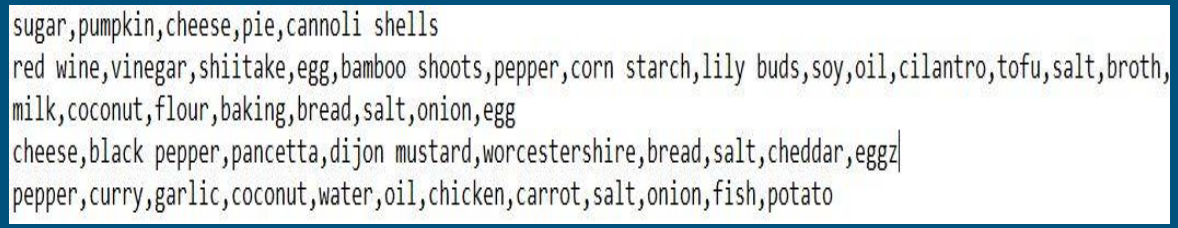
ingredients_Recipes5k.txt - The Original List of Ingredients

Preprocessing the Data

- Removed overly descriptive words (eg. unsalted, sliced).
- Removed/replaced all unicode characters and duplicate ingredients.



classes_Recipes5k.txt
(Recipe names)



cleaned_ingredients.txt
(Ingredients in the recipes)

Preprocessed Dataset for Food Analysis

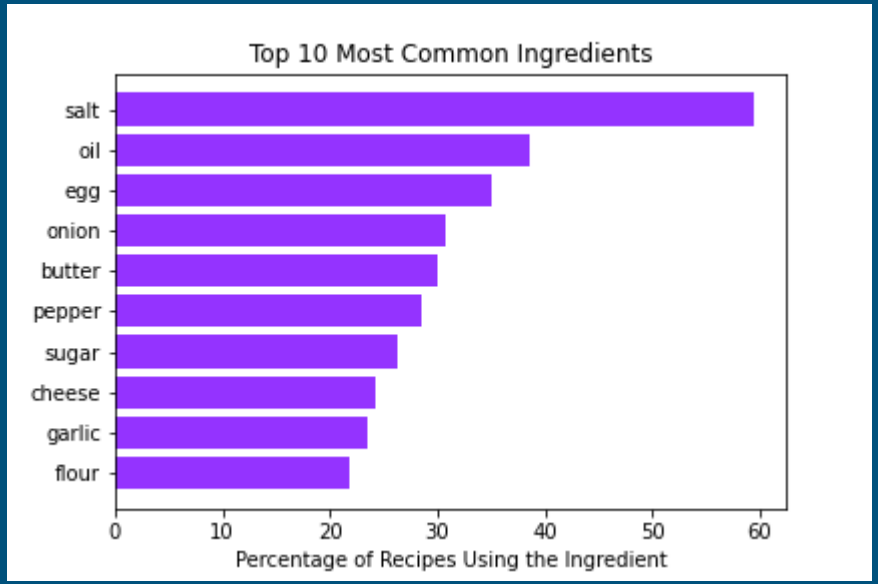
Removed recipes which don't map to any specific dish, resulting in **4,330 unique recipes, 668 unique ingredients and 101 unique dishes.**

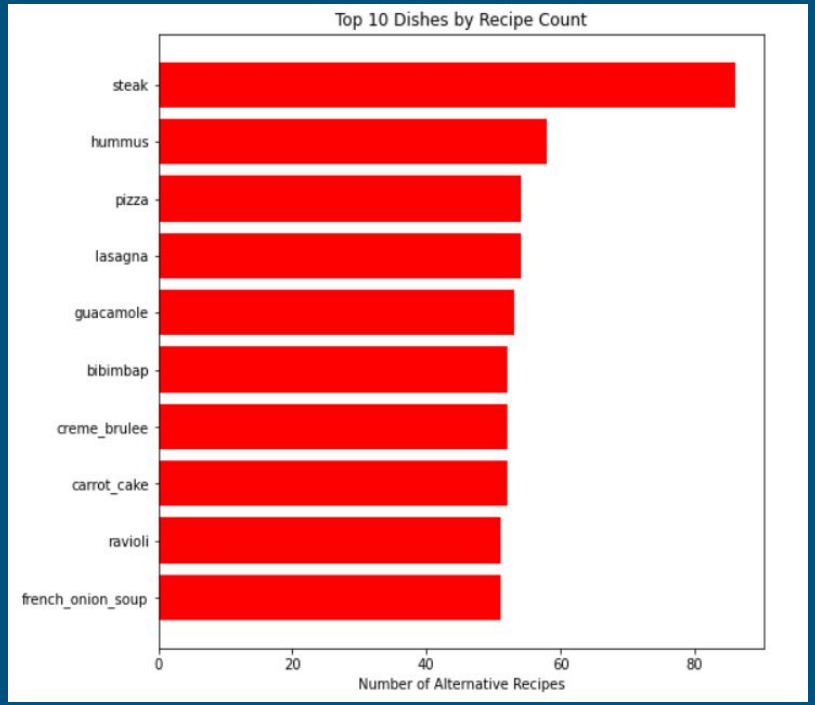
	recipe_name	ingredients	dish
0	Freezer Breakfast Burritos_3	pepper,salsa,garlic,black	breakfast_burrito
1	Mobile-Style Oysters	pepper,parsley,garlic,cheese,black	oysters
2	Cheesecake	sugar,egg,cheese,corn	cheesecake
3	Fried Ravioli_6	ravioli,water,oil,bread,marinara,egg	ravioli
4	Poutine_10	gravy,pepper,cheese,butter,oil,flour,beef,salt...	poutine
...	...	...	...
4325	Death By Oreo Cupcakes	sugar,cheese,oreo	cupcakes
4326	Thai Green Chicken Curry	basil,coconut,gin,chicken,zucchini,mushroom,sp...	chicken_curry
4327	Mussels in White Wine	pepper,parsley,garlic,oil,white	mussels
4328	Low Carb Crab Cakes	pepper,parsley,mayonnaise,crab,oil,lemon,dijon	crab_cakes
4329	Tart Vanilla Frozen Yogurt	plain	frozen_yogurt

4330 rows × 4 columns

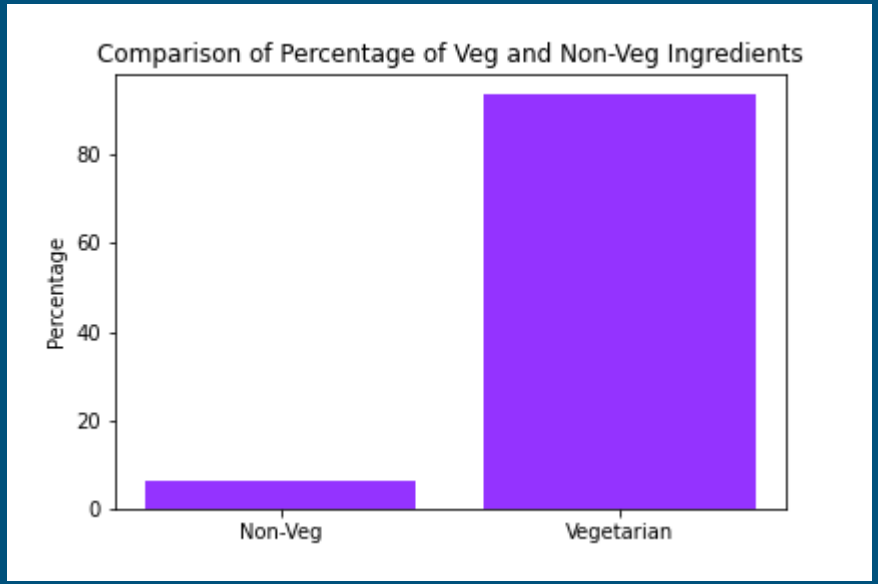
Final dataset

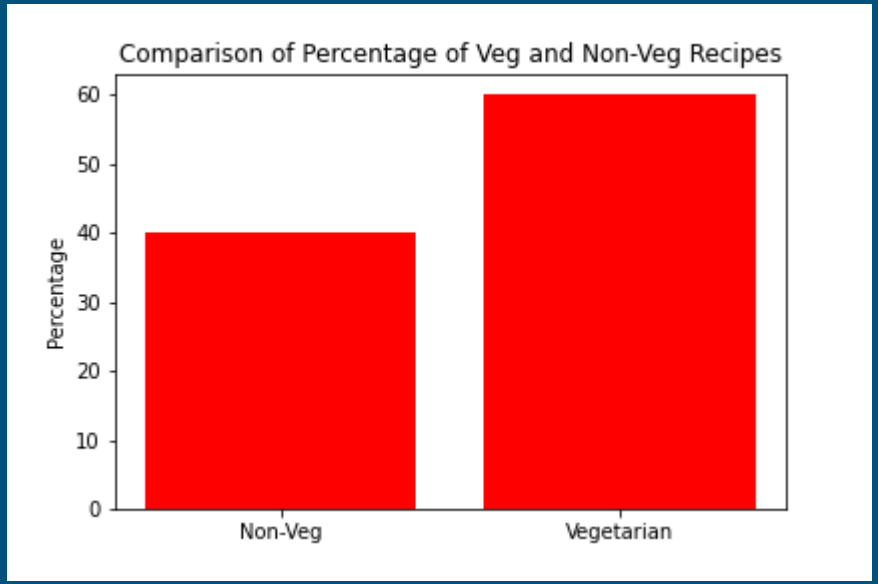
Exploratory Data Analysis - Summary plots



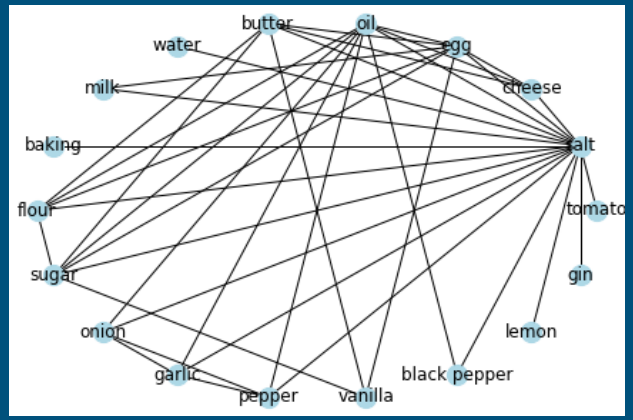


Exploratory Data Analysis - Summary plots





Graphs and Association Analysis



Pairs that co-occur the most:

1. Salt and oil
2. Salt and pepper
3. Salt and egg
4. Salt and butter
5. Salt and onion

Market Basket Analysis backed up our network analysis

	items <fctr>	support <dbl>	transIdenticalToItemsets <dbl>	count <int>
[1]	{oil,salt}	0.2679237	0.0026937422	1293
[2]	{pepper,salt}	0.2206797	0.0060091173	1065
[3]	{egg,salt}	0.2186075	0.0024865313	1055
[4]	{butter,salt}	0.2028595	0.0012432656	979
[5]	{onion,salt}	0.2009946	0.0035225860	970
[6]	{flour,salt}	0.1732283	0.0002072109	836
[7]	{garlic,salt}	0.1647327	0.0014504766	795
[8]	{egg,sugar}	0.1603813	0.0103605470	774
[9]	{salt,sugar}	0.1556154	0.0010360547	751
[10]	{black pepper,salt}	0.1508496	0.0060091173	728

Association Rule Mining

Found a few ingredients that occur frequently individually, but do not occur together as much as we expect them to. This is backed up by the low lift values (<1) for these pairs.

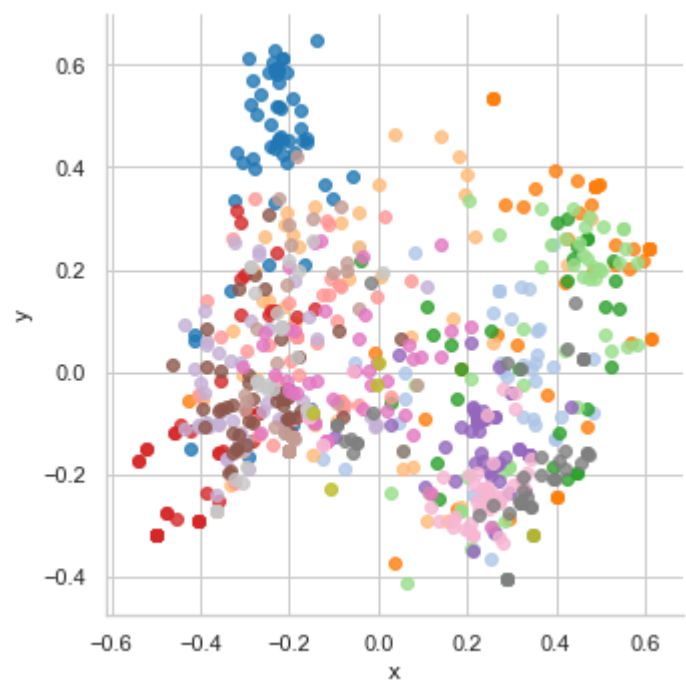
lhs <fctr>	<fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	coverage <dbl>	lift <dbl>	count <int>
{}	=>	{oil}	0.3852051	0.3852051	1.0000000	1.0000000	1859
{}	=>	{salt}	0.5938666	0.5938666	1.0000000	1.0000000	2866
{}	=>	{soy}	0.1000829	0.1000829	1.0000000	1.0000000	483
{}	=>	{tomato}	0.1301285	0.1301285	1.0000000	1.0000000	628
{salt}	=>	{sugar}	0.1556154	0.2620377	0.5938666	0.9957432	751
{sugar}	=>	{salt}	0.1556154	0.5913386	0.2631579	0.9957432	751
{egg}	=>	{oil}	0.1340655	0.3826138	0.3503937	0.9932729	647
{oil}	=>	{egg}	0.1340655	0.3480366	0.3852051	0.9932729	647
{salt}	=>	{cheese}	0.1396602	0.2351710	0.5938666	0.9675491	674
{cheese}	=>	{salt}	0.1396602	0.5745951	0.2430584	0.9675491	674

Vectorization

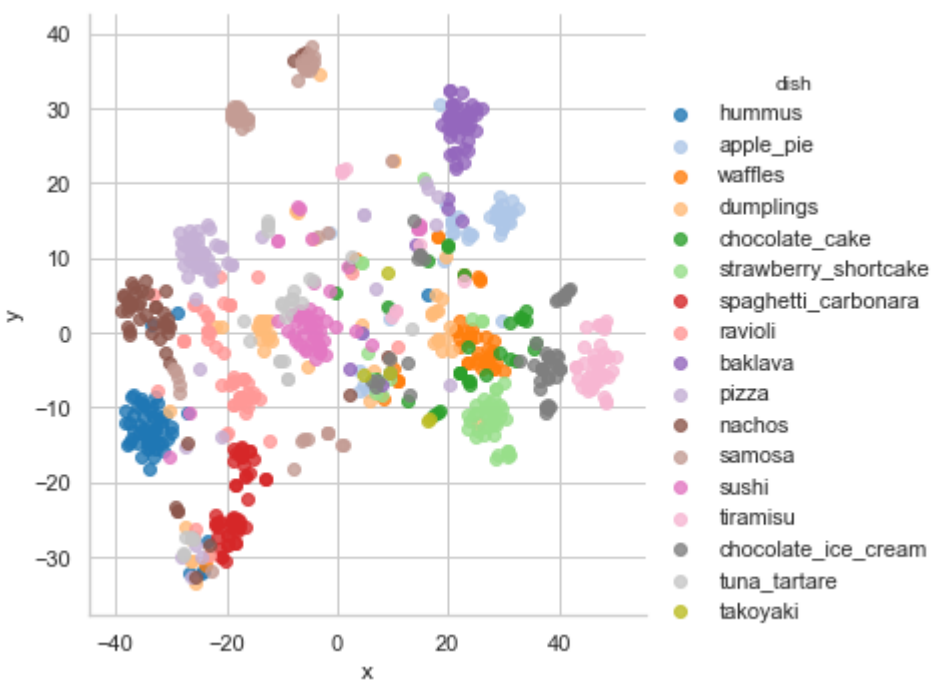
- Represented a recipe with a one hot encoded vector of its ingredients.
- Established a vocabulary of ingredients and encoded each recipe in a **668-dimensional vector of ingredients.**

```
Ingredients represented by Bag of Words for each Recipe


0       {'tahini': 1, 'beans': 1, 'garlic': 1, 'water'...
1       {'pepper': 1, 'cheese': 1, 'kalamata': 1, 'gre...
2            {'garlic': 1, 'cheese': 1, 'stouffer''s': 1}
3       {'chili': 1, 'parsley': 1, 'garlic': 1, 'chees...
4                       {'pepper': 1, 'salt': 1, 'egg': 1}
                                ...
4325                              {'garlic': 1, 'black': 1}
4326    {'beans': 1, 'salsa': 1, 'flour': 1, 'corn': 1...
4327    {'sugar': 1, 'milk': 1, 'fat': 1, 'nut': 1, 'f...
4328    {'oil': 1, 'saffron': 1, 'shallot': 1, 'white'...
4329    {'mayonnaise': 1, 'cheese': 1, 'butter': 1, 'c...
Name: bow, Length: 4330, dtype: object


Shape of X and y after one hot encoding of the ingredients


(4330, 668)
[[0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 ...
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]
 [0. 0. 0. ... 0. 0. 0.]]


(4330,)
0       59
1       50
2       60
3       48
4       67
       ..
4325    42
4326    58
4327    33
4328    65
4329    27
Length: 4330, dtype: int8
```

Ingredients-based Recipe Clustering

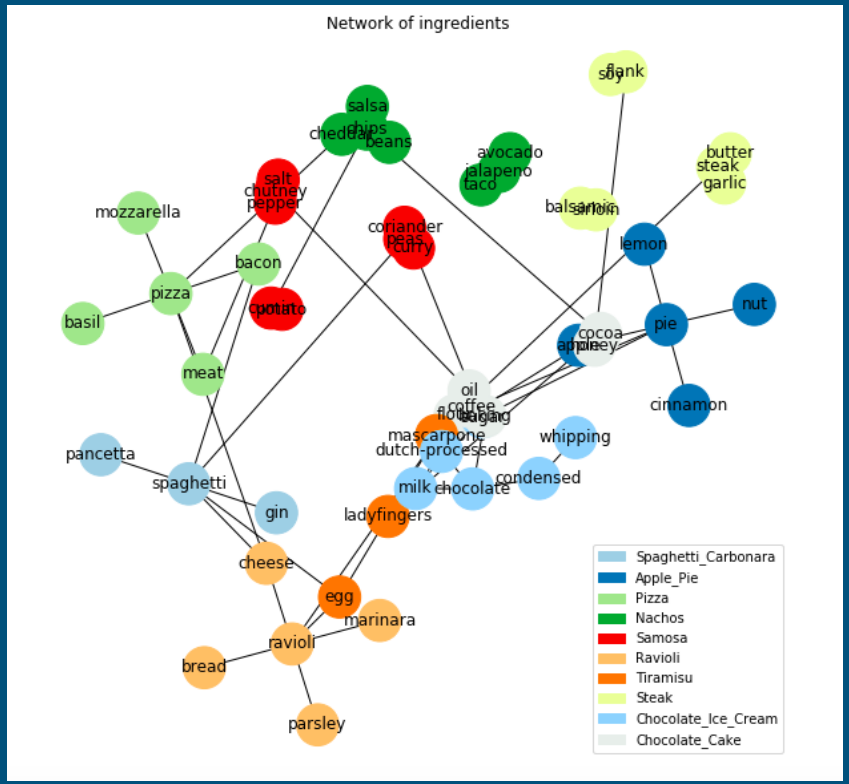


PCA



t-SNE

Ingredient Interactions across Dishes



Cosine Similarity: Doc2Vec vs. One Hot Encoding

- Doc2Vec creates a vectorized representation of a document by taking into consideration their contexts.
  - Our intuition is that it will help **capture deeper relations between ingredients** given the context in which they are used.

- Compare the recipes by summing the one hot encoding vectors of their ingredients.

Recommendations: Doc2Vec

Dark Chocolate Mousse

Doc2Vec
Cosine Similarity

Best Eggs Benedict
100%

Fish and Chips!
100%

Death by Chocolate Cake
100%

Panna Cotta_3
100%

Excellent Failure!

Recommendations: One Hot Encoding

Dark Chocolate Mousse

One Hot Encoding
Cosine Similarity

Chocolate Mousse_5
100%

Classic Chocolate Mousse_3
89.44%

Caffe Mocha Creme Brulee
86.44%

Two Ingredient Chocolate Mousse
86.60%

Conclusion & Future Work

- Using the knowledge of inter and intra-class variability of recipes through EDA, we built a system to recommend similar recipes.

- The Doc2Vec model gave poor results, while the one-hot encoded model performed well for ingredients-based recommendation.

- Going forward, we plan to integrate our recommendation system into a digital cookbook.