

# Recipe Data Analysis

Rithic Kumar Nandakumar,  
Student ID 20998160  
rknandak@uwaterloo.ca

Prepared for CS 651 Course Project  
University of Waterloo, Winter 2024

## Contents

<b>1</b>	<b>Goal</b>	<b>2</b>
<b>2</b>	<b>Methodology</b>	<b>3</b>
2.1	Question 1: Clustering . . . . .	3
2.2	Question 2: Frequent itemsets . . . . .	3
2.3	Question 3: Recommendation System . . . . .	4
<b>3</b>	<b>Results</b>	<b>4</b>
3.1	Question 1: Clustering . . . . .	4
3.2	Question 2: Frequent itemsets . . . . .	6
3.3	Question 3: Recommendation System . . . . .	7
<b>4</b>	<b>Future Work</b>	<b>10</b>

## 1 Goal

The Recipe dataset from Kaggle contains over 2 million recipes from all over the world. The dataset contains the dish name, ingredients used, and directions. The author also provides a NER column that applies Named Entity Recognition (NER) on the ingredients and gives the base name of the ingredients used without the brands and quantity used.

	title	ingredients	directions	link	source	NER	site
0	No-Bake Nut Cookies	["1 c. firmly packed brown sugar", "1/2 c. eva...	["In a heavy 2-quart saucepan, mix brown sugar...	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=44874">www.cookbooks.com/Recipe-Details.aspx?id=44874</a>	Gathered	["bite size shredded rice biscuits", "vanilla"...	www.cookbooks.com
1	Jewell Ball'S Chicken	["1 small jar chipped beef, cut up", "4 boned ...	["Place chipped beef on bottom of baking dish....	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=699419">www.cookbooks.com/Recipe-Details.aspx?id=699419</a>	Gathered	["cream of mushroom soup", "beef", "sour cream"...	www.cookbooks.com
2	Creamy Corn	["2 (16 oz.) pkg. frozen corn", "1 (8 oz.) pkg...	["In a slow cooker, combine all ingredients. C...	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=10570">www.cookbooks.com/Recipe-Details.aspx?id=10570</a>	Gathered	["frozen corn", "pepper", "cream cheese", "gar..."	www.cookbooks.com
3	Chicken Funny	["1 large whole chicken", "2 (10 1/2 oz.) cans...	["Boil and debone chicken.", "Put bite size pi...	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=897570">www.cookbooks.com/Recipe-Details.aspx?id=897570</a>	Gathered	["chicken gravy", "cream of mushroom soup", "c..."	www.cookbooks.com
4	Reeses Cups(Candy)	["1 c. peanut butter", "3/4 c. graham cracker ...	["Combine first four ingredients and press in ...	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=659239">www.cookbooks.com/Recipe-Details.aspx?id=659239</a>	Gathered	["graham cracker crumbs", "powdered sugar", "p..."	www.cookbooks.com

Figure 1: Dataset

We plan to explore the following questions:

- **Question 1:** Can we employ clustering techniques to categorize recipes based on their ingredient profiles and identify underlying patterns?
- **Question 2:** What are the most common ingredients used together?
- **Question 3:** Now given the clusters, and a list of ingredients the user has and their preferred dishes, can we recommend dishes that are similar to the user's taste and make use of the ingredients the user has?

Since the decompressed dataset is over 2GB, we can use Apache Spark MLlib to analyse the dataset.

## 2 Methodology

### 2.1 Question 1: Clustering

To cluster the recipes based on ingredients, the first step is to convert the ingredients to a list of vectors that can be used as input for any clustering algorithm. We use word2vec word embedding model to convert the ingredients to vectors. Word2vec model produces similar vectors for similar words, thereby grouping similar ingredients together.

Since the ingredients list have a median length of 8, the vector size was chosen as 10 to ensure adequate representation of the ingredients. Spark MLlib has inbuilt support for word2vec model.

The vectors are used as inputs for K-Means Clustering algorithm. We identify the best value of cluster size by varying  $k$  from 2 to 10 and choosing the number of clusters that have the best silhouette coefficient.

The silhouette coefficient  $s(i)$  for a data point  $i$  is calculated as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Where:

- $a(i)$  is the mean distance between  $i$  and all other points in the same cluster
- $b(i)$  is the mean distance between  $i$  and all points in the next closest cluster

The coefficient ranges from  $-1$  to  $1$ , with values close to  $1$  indicating well-clustered data points, and values close to  $-1$  suggesting poorly matched data points.

### 2.2 Question 2: Frequent itemsets

To explore the most common ingredients used together, we can employ Frequent itemset mining (FIM) algorithms to discover sets of items that co-occur together. FIM algorithms can identify all itemsets above the specified support threshold whereas Pointwise Mutual Information (PMI) considers only pairs of items.

FP-Growth [1] by Han et al is used to mine the frequent itemsets. Spark's FP-Growth implementation takes the following hyperparameters:

- minSupport: Minimum support for an itemset to be considered frequent
- minConfidence: Minimum confidence for generating Association Rule

Since the dataset is large, the minimum support is chosen as 0.01 and the minimum confidence is chosen as 0.01.

## 2.3 Question 3: Recommendation System

To build a recommendation system, we get the list of ingredients available and dishes the user likes. The most frequent cluster from the dishes liked by the user is chosen as the reference cluster. We employ Jaccard index to calculate the similarity score between the ingredients available and the ingredients of the dishes among the reference cluster to get the list of the most relevant dishes.

The Jaccard index  $J(A, B)$  is a measure of similarity between two sets  $A$  and  $B$ , defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where:

- $|A \cap B|$  is the cardinality of the intersection of  $A$  and  $B$
- $|A \cup B|$  is the cardinality of the union of  $A$  and  $B$

The Jaccard index ranges from 0 to 1, with 0 indicating no similarity between the sets, and 1 indicating that the sets are identical.

## 3 Results

### 3.1 Question 1: Clustering

From the analysis of the silhouette coefficients,  $k = 2$  gives the best clusters.

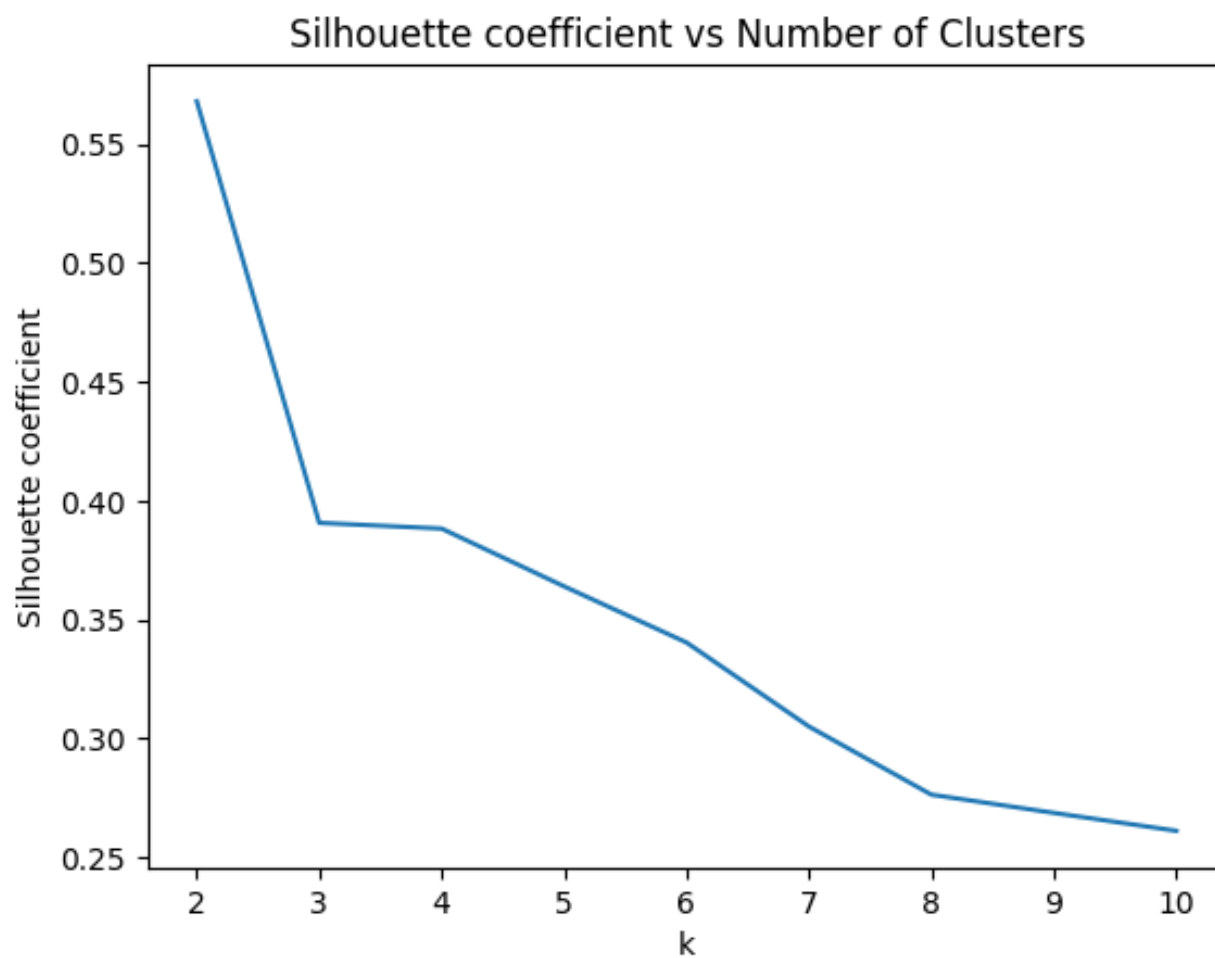


Figure 2: Silhouette analysis

Table 1: Example dishes from the clusters

Cluster 0	Cluster 1
Jewell Ball'S Chicken	No-Bake Nut Cookies
Creamy Corn	Reeses Cups(Candy)
Chicken Funny	Rhubarb Coffee Cake
Cheeseburger Potato Soup	Millionaire Pie
Scalloped Corn	Double Cherry Delight
Nolan's Pepper Steak	Buckeye Candy
Quick Barbecue Wings	Pink Stuff(Frozen Dessert)
Taco Salad Chip Dip	Fresh Strawberry Pie
Broccoli Salad	Easy German Chocolate Cake
Cuddy Farms Marinated Turkey	Strawberry Whatever

Doing exploratory analysis of the 2 clusters reveals that cluster 1 contains cookies, cakes, candies and breads - dishes that are commonly classified as bakery items. Cluster 0 contains the remaining dishes.

### 3.2 Question 2: Frequent itemsets

The following table lists the top 10 association rules listed in descending order of confidence

Table 2: Association Rules and Confidence

antecedent	consequent	confidence
soda,eggs	flour	0.910157233835858
soda,salt	flour	0.8864585658259733
soda	flour	0.8652257869390224
shortening,salt	flour	0.8495803909528196
baking soda,cinnamon	flour	0.843564783217788
baking powder,salt	flour	0.8310132101676302
egg,sugar,salt	flour	0.8296724470134875
shortening	flour	0.8259009645264135
sugar,vanilla,flour,salt	eggs	0.8255105038930513

We can observe the top association rules come from bakery recipes as they use a common set of base ingredients.

### 3.3 Question 3: Recommendation System

A webapp is created using Streamlit. The app takes the ingredients available and list of dishes the user likes. Dishes with highest Jaccard index values are returned amongst the most frequent cluster of the user's taste



project\_streamlit - Streamlit x New Tab x | +

localhost:8501

Waterloo Free PDF, Video, L... Open Peeps, Hand... Mail - Rithic Kuma...

Deploy

## Recipe Recommendation System

Choose Ingredients available

Choose an option

Choose your favourite dishes

Choose an option

Submit

Ingredients cannot be empty

Figure 3: Home page

**Recipe Recommendation System**

Choose Ingredients available

pepper x eggs x oregano x

Choose your favourite dishes

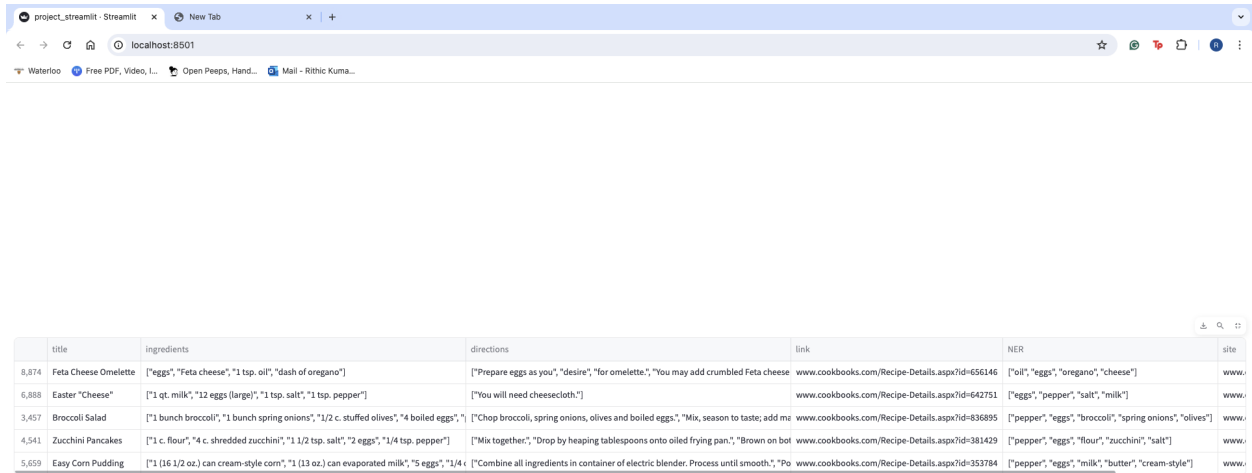
Creamy Corn x French Toast x Lasagna x

Submit

Selected Options:

	title	ingredients
8,874	Feta Cheese Omelette	["eggs", "Feta cheese", "1 tsp. oil", "dash of oregano"]
6,888	Easter "Cheese"	["1 qt. milk", "12 eggs (large)", "1 tsp. salt", "1 tsp. pepper"]
3,457	Broccoli Salad	["1 bunch broccoli", "1 bunch spring onions", "1/2 c. stuffed olives", "4 boiled eggs", "1
4,541	Zucchini Pancakes	["1 c. flour", "4 c. shredded zucchini", "1 1/2 tsp. salt", "2 eggs", "1/4 tsp. pepper"]
5,659	Easy Corn Pudding	["1 (16 1/2 oz.) can cream-style corn", "1 (13 oz.) can evaporated milk", "5 eggs", "1/4 c

Figure 4: Input from the user



	title	ingredients	directions	link	NER	site
8,874	Feta Cheese Omelette	["eggs", "Feta cheese", "1 tsp. oil", "dash of oregano"]	["Prepare eggs as you", "desire", "for omelette.", "You may add crumbled Feta cheese"]	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=656146">www.cookbooks.com/Recipe-Details.aspx?id=656146</a>	["oil", "eggs", "oregano", "cheese"]	www.
6,888	Easter Cheese	["1 qt. milk", "12 eggs (large)", "1 tsp. salt", "1 tsp. pepper"]	["You will need cheesecloth;"]	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=642751">www.cookbooks.com/Recipe-Details.aspx?id=642751</a>	["eggs", "pepper", "salt", "milk"]	www.
3,457	Broccoli Salad	["1 bunch broccoli", "1 bunch spring onions", "1/2 c. stuffed olives", "4 boiled eggs", "]	["Chop broccoli, spring onions, olives and boiled eggs", "Mix, season to taste; add ma"]	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=836895">www.cookbooks.com/Recipe-Details.aspx?id=836895</a>	["pepper", "eggs", "broccoli", "spring onions", "olives"]	www.
4,541	Zucchini Pancakes	["1 c. flour", "4 c. shredded zucchini", "1 1/2 tsp. salt", "2 eggs", "1/4 tsp. pepper"]	["Mix together", "Drop by heaping tablespoons onto oiled frying pan.", "Brown on bot"]	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=381429">www.cookbooks.com/Recipe-Details.aspx?id=381429</a>	["pepper", "eggs", "flour", "zucchini", "salt"]	www.
5,659	Easy Corn Pudding	["1 (16 1/2 oz.) can cream-style corn", "1 (13 oz.) can evaporated milk", "5 eggs", "1/4 t"]	["Combine all ingredients in container of electric blender. Process until smooth.", "Po"]	<a href="http://www.cookbooks.com/Recipe-Details.aspx?id=353784">www.cookbooks.com/Recipe-Details.aspx?id=353784</a>	["pepper", "eggs", "milk", "butter", "cream-style"]	www.

Figure 5: Results of the recommendation system

The results from the recommendation system can also be saved as a CSV file for later access.

## 4 Future Work

We have identified two major divisions in the recipes, namely bakery and non-bakery dishes by clustering the ingredients used. For future work, the quantity of ingredients can be normalised based on the quantity of other ingredients in the recipe to pave the way for uniform representation. Other word embedding models can also be utilised to ensure better representation.

## References

- [1] Jiawei Han, Jian Pei, and Yiwen Yin. “Mining frequent patterns without candidate generation”. In: *SIGMOD Rec.* 29.2 (May 2000), pp. 1–12. ISSN: 0163-5808. DOI: 10.1145/335191.335372. URL: <https://doi.org/10.1145/335191.335372>.