Recipe Data Analysis

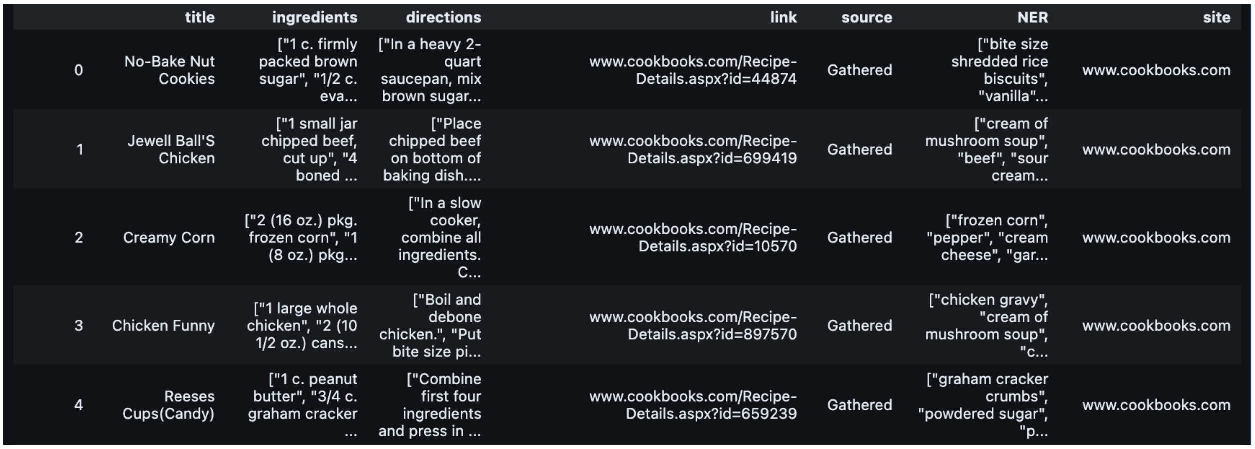
Rithic Kumar Nandakumar,
Student ID 20998160
rknandak@uwaterloo.ca

Prepared for CS 651 Data-Intensive Distributed Computing Course Project
University of Waterloo, Winter 2024

Contents

1 Goal

The Recipe dataset [1]from Kaggle contains over 2 million recipes from all over the world. The dataset contains the dish name, ingredients used and cooking instructions denoted in ”directions” column. 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...	www.cookbooks.com/Recipe-Details.aspx?id=44874	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....	www.cookbooks.com/Recipe-Details.aspx?id=699419	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...	www.cookbooks.com/Recipe-Details.aspx?id=10570	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...	www.cookbooks.com/Recipe-Details.aspx?id=897570	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 ...	www.cookbooks.com/Recipe-Details.aspx?id=659239	Gathered	["graham cracker crumbs", "powdered sugar", "p...	www.cookbooks.com

Figure 1: Dataset

This project aims to explore the possibilities opened by the below mentioned questions:

- **Question 1:** Can clustering techniques be employed to categorize recipes based on their ingredient profiles and identify underlying patterns?
- **Question 2:** What are the very common ingredients used together?
- **Question 3:** Given the clusters, and a list of ingredients the user has and their preferred dishes, can dishes be recommended to the user based on the user’s taste?

Since the decompressed dataset is over 2GB, Apache Spark MLlib will be used 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. In this project, word2vec [2] word embedding model has been used 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. The best value of cluster size has been identified by varying k from 2 to 10 and choosing the number of clusters that have the best silhouette coefficient.

The silhouette coefficient [3]$s(i)$ 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, Frequent itemset mining (FIM) algorithms has been employed to discover sets of items that co-occur together. FIM algorithms can identify all item sets above the specified support threshold whereas Pointwise Mutual Information (PMI) considers only pairs of items.

FP-Growth [4] 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, the list of ingredients available and dishes the user likes, is made use of. The most frequent cluster from the dishes liked by the user is chosen as the reference cluster. Jaccard index is used 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.

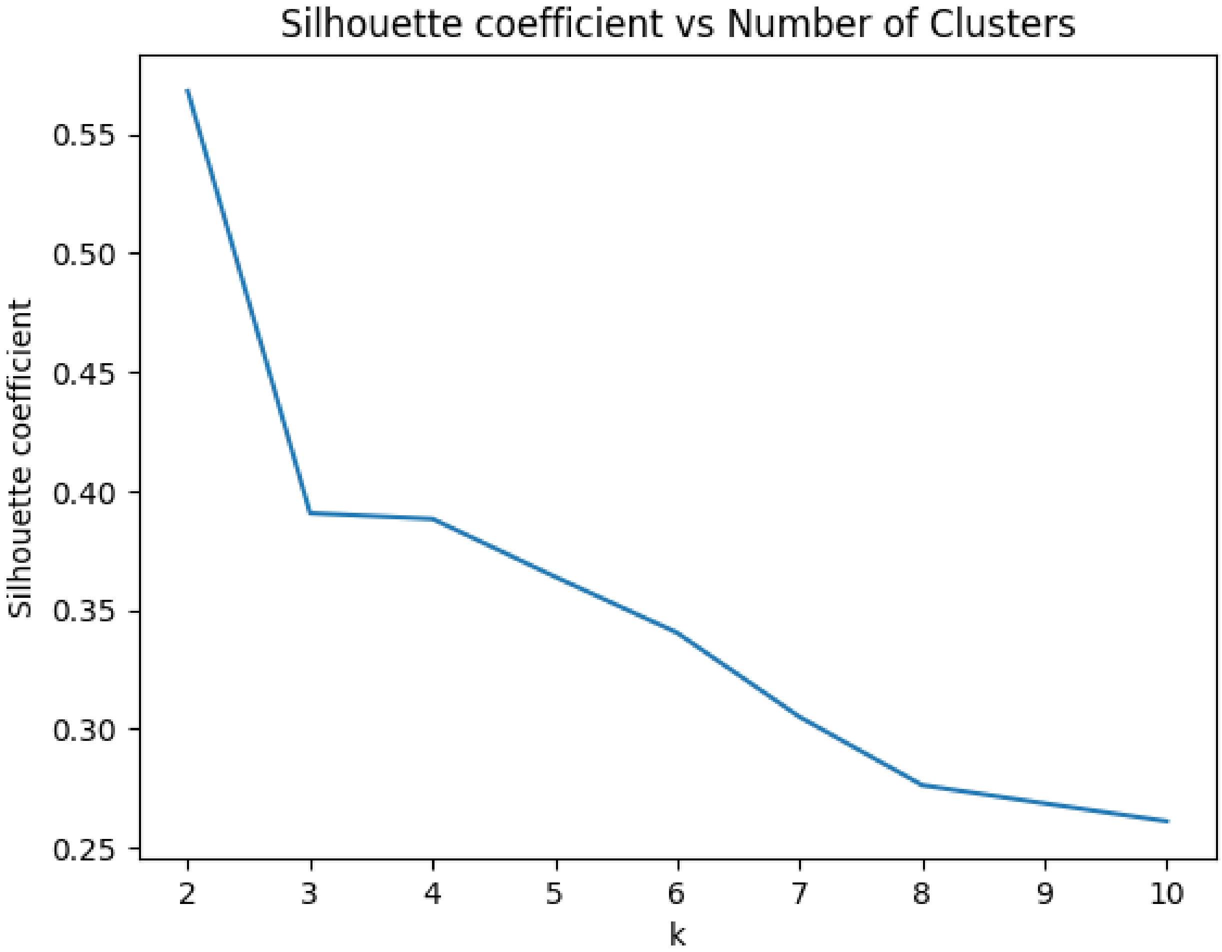


Figure 2: Silhouette analysis

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

Exploratory analyses of the 2 clusters reveal 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

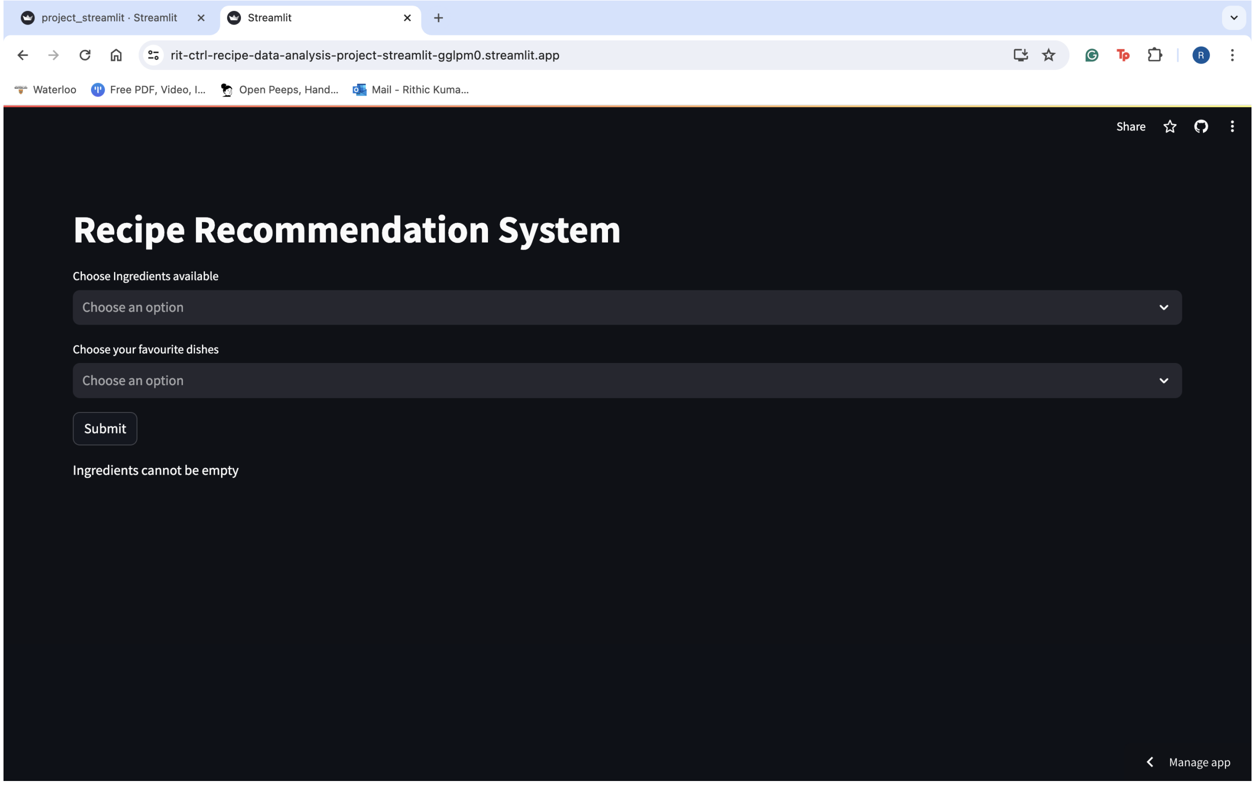
Table 2: Association Rules and Confidence

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

It is observed that 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 [5]. A subset of the original data is used to match the bandwidth available for hosting a Streamlit app. The app takes the ingredients available and a list of dishes the user likes. Dishes with the highest Jaccard index values are returned amongst the most frequent cluster of the user's taste.


project_streamlit · Streamlit
Streamlit
rit-ctrl-recipe-data-analysis-project-streamlit-gglpm0.streamlit.app
Waterloo
Free PDF, Video, I...
Open Peeps, Hand...
Mail - Rithic Kuma...
Share
Manage app

Recipe Recommendation System

Choose Ingredients available

Choose an option

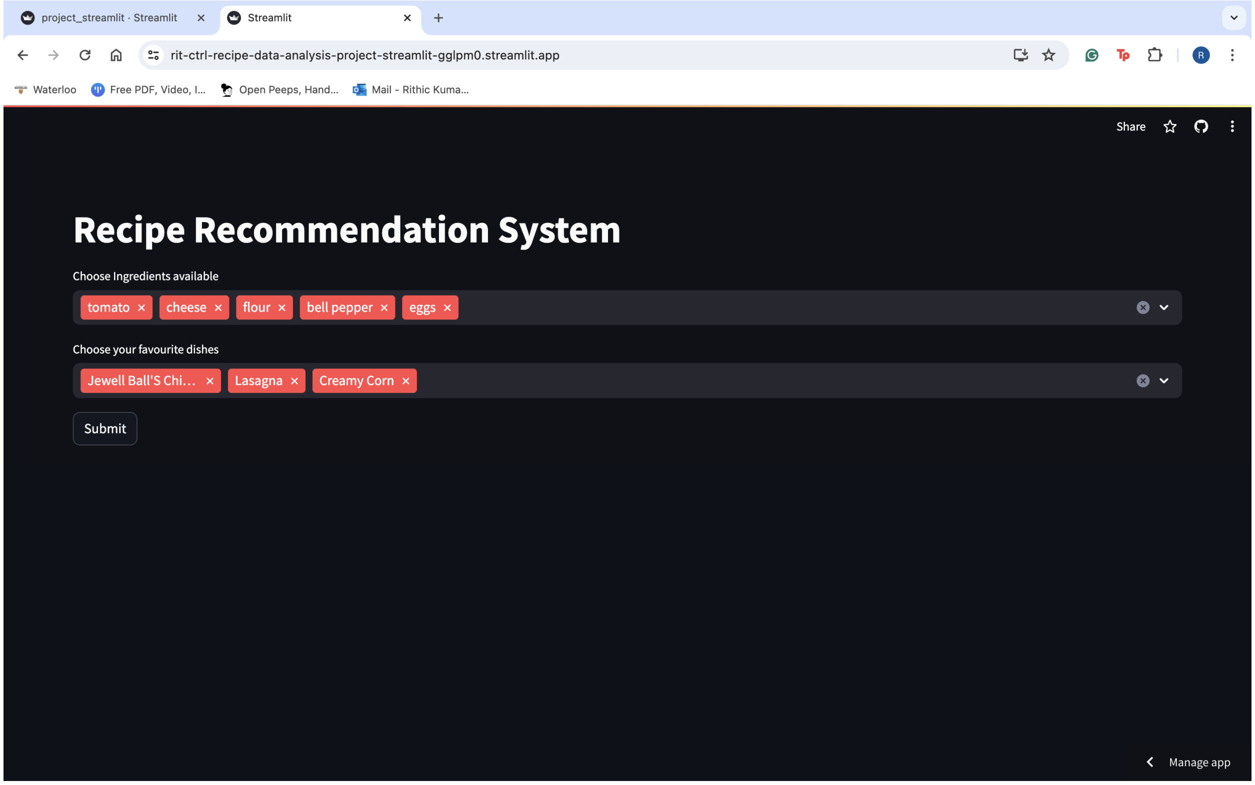
Choose your favourite dishes

Choose an option

Submit

Ingredients cannot be empty

Figure 3: Home page



project_streamlit · Streamlit
Streamlit
rit-ctrl-recipe-data-analysis-project-streamlit-gglpm0.streamlit.app
Waterloo
Free PDF, Video, I...
Open Peeps, Hand...
Mail - Rithic Kuma...
Share
Manage app

Recipe Recommendation System

Choose Ingredients available

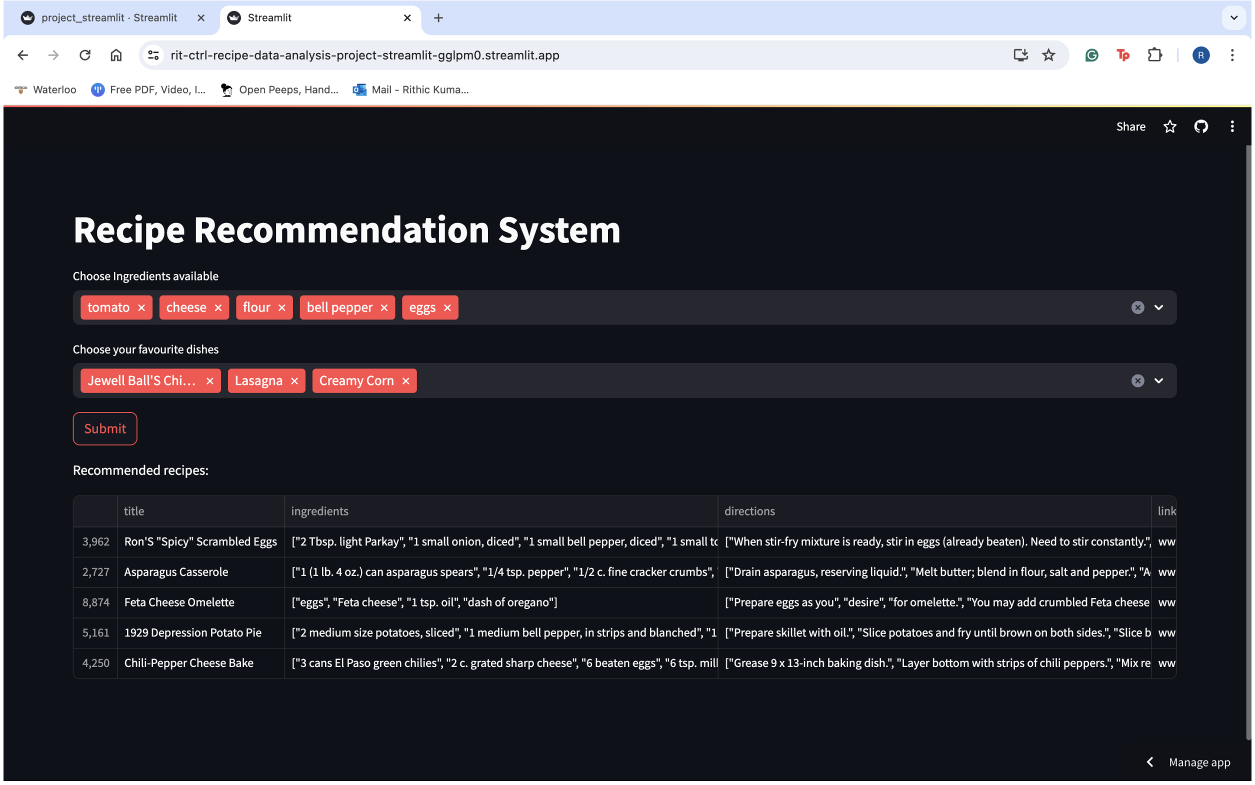
tomato × cheese × flour × bell pepper × eggs ×

Choose your favourite dishes

Jewell Ball'S Chi... × Lasagna × Creamy Corn ×

Submit

Figure 4: Input from the user


project_streamlit · Streamlit
Streamlit
rit-ctrl-recipe-data-analysis-project-streamlit-gglpm0.streamlit.app
Waterloo
Free PDF, Video, I...
Open Peeps, Hand...
Mail - Rithic Kuma...
Share
Manage app

Recipe Recommendation System

Choose Ingredients available

tomato × cheese × flour × bell pepper × eggs ×

Choose your favourite dishes

Jewell Ball'S Chi... × Lasagna × Creamy Corn ×

Submit

Recommended recipes:

	title	ingredients	directions	link
3,962	Ron'S "Spicy" Scrambled Eggs	["2 Tbsp. light Parkay", "1 small onion, diced", "1 small bell pepper, diced", "1 small to	["When stir-fry mixture is ready, stir in eggs (already beaten). Need to stir constantly.",	ww
2,727	Asparagus Casserole	["1 (1 lb. 4 oz.) can asparagus spears", "1/4 tsp. pepper", "1/2 c. fine cracker crumbs",	["Drain asparagus, reserving liquid.", "Melt butter; blend in flour, salt and pepper.", "A	ww
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	ww
5,161	1929 Depression Potato Pie	["2 medium size potatoes, sliced", "1 medium bell pepper, in strips and blanched", "1	["Prepare skillet with oil.", "Slice potatoes and fry until brown on both sides.", "Slice b	ww
4,250	Chili-Pepper Cheese Bake	["3 cans El Paso green chilies", "2 c. grated sharp cheese", "6 beaten eggs", "6 tsp. mill	["Grease 9 x 13-inch baking dish.", "Layer bottom with strips of chili peppers.", "Mix re	ww

Figure 5: Results of the recommendation system

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

The demo of the webapp can be accessed using this link - https://rit-ctrl-recipe-data-analysis-project.streamlit.app/

4 Future Work

Two major divisions in the recipes, namely bakery and non-bakery dishes have been identified 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] *Recipe Dataset (over 2M) Food — kaggle.com.* https://www.kaggle.com/datasets/wilmerarltstrmberg/recipe-dataset-over-2m. [Accessed 14-04-2024].

[2] Tomas Mikolov et al. *Efficient Estimation of Word Representations in Vector Space.* 2013. arXiv: 1301.3781 [cs.CL].

[3] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257.

[4] 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.

[5] *Streamlit &x2022; A faster way to build and share data apps — streamlit.io.* https://streamlit.io/. [Accessed 14-04-2024].