Dishdelight

# Problem Statement

In 2022, National Zero Waste Council conducted research in Canada and the following were the results:

|  |  |
| --- | --- |
| “63% of the food Canadians thrown away could have been eaten” | “For the average Canadian household that amounts to *140 kilograms of wasted food per year – at a cost of more than $1,300 per year!*” |

Food waste is a huge problem that goes subtlety unnoticed in our household as the amount of food thrown out each week may be minimal but accumulates to noticeable effects.

Our goal of this project is to create a recipe recommender based on ingredients leftover in the fridge or provide users the ability to make new and interesting recipes with similar ingredients. This will allow households to reduce food wasted.

# Industry Background

In 2022, the grocery industry boasted a staggering $114.2 billion in revenue, underscoring its vast reach in every household's consumption.

One of the biggest players in the recipe recommendation is Yummly. Yummly provides a personalized recipe recommendation and grocery delivery services. It has in-depth filtering from dietary restrictions to user’s cooking skill level. As of 2015, Yummly is valued at $100 million, which is only a small fraction of the grocery industry.

# Data Overview

There were three datasets that were sourced from Kaggle, which were web-scrapped from Food.com. After merging, preprocessing and cleaning the datasets, the recipe dataset totaled 230,000 rows and 11 columns, and the review dataset totaled 765,000 rows and 3 columns. For further information regarding the details of the data columns, please see the Jupyter notebook.

**Data source:**

[**https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews?select=reviews.csv**](https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews?select=reviews.csv)

[**https://www.kaggle.com/datasets/shuyangli94/foodcom-recipes-with-search-terms-and-tags**](https://www.kaggle.com/datasets/shuyangli94/foodcom-recipes-with-search-terms-and-tags)

[**https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions?select=RAW\_recipes.csv**](https://www.kaggle.com/datasets/shuyangli94/food-com-recipes-and-user-interactions?select=RAW_recipes.csv)

# Exploratory Data Analysis

A graph with blue lines and numbers

Description automatically generated A graph with blue dots and numbers

Description automatically generated

A graph of food items

Description automatically generatedMost recipes were added and reviewed around 2006-2008. There is a huge drop-off in recipes added in the later years. This may indicate that Food.com was being used less frequently in latter years. This is a takeaway for future state regarding creating a more diverse recipe database.

The recipe category frequency is the dessert category. This category accounts for 11% of all recipes in the dataset. This is something to note when recommending recipes as desserts may be a higher liklihood of being recommended.

Most recipes have a rating score of 5. This

# Models

A list of food items

Description automatically generatedMost popular recipes:

I ran three different models and here is the summary:

|  |  |  |
| --- | --- | --- |
| **Recommender System** | **Algorithm** | **Summary** |
| Collaborative Filtering Recommender | Funk Singular Value Decomposition | Evaluation metric – Fraction of Concordant Pairs score = 56% |
| Content-based Filtering Recommender | Word2Vec 🡪 Mean Embeddings | Cosine Similarity Score = 70% - 80%  Jaccard Similarity Score = 10% |

Content-based recommender results:

A screenshot of a recipe

Description automatically generated

A screenshot of a computer

Description automatically generatedLast model I combined both the collaborative and content-based recommender into a hybrid model. Below is a sample test result from the hybrid model:

# Model Results

From our collaborative recommender model, we are only attaining a 56% score. This is a poor score, but still better than recommending the most popular items. We will need to utilize other models such as Stochastic Gradient Descent and Alternating Least Square to determine if these models will provide a better score.

The content-based recommender model ran well with the cosine similarity score. One thing to note regarding the regarding the score range is that we ran multiple tests for input ingredients, and they all fell within the 70% to 80% range. The Jaccard similarity score was very poor and will not be used.

As you can see from the score from the hybrid model, the score decreased dramatically. This is due to the content-based recommender system running on a refined dataset of only 100 recipes. Currently, the hyper parameter for the threshold of recipes will need to be manually tuned and adjusted as input ingredient with beef is receiving 0.

# Conclusion

Out of all our models, the content-based recommender is providing the best results. Although the user’s historical preference is not considered, by providing a large enough list (10-20 recipes), the user has the ability to pick the one that he or she likes the most. If we do look at the recipes recommended though, they are very plain and bland and do not seem like recipes that most people would like. We will need to refine our collaborative and hybrid model to attain better filtering results to help us create a more refined and personalized recommendation list for individual users.

**Limitations**

Our initial goal was to create a recipe recommender that allowed the user to utilize leftover ingredients in the fridge to create a dish that they would like. Since we are using cosine similarity for our content-based recommender, we can see from the results that certain recipe ingredients do not match the inputs. We tried to compensate for this using the Jaccard similarity score which compares the same items between lists, but the score came out very poorly at 10%.

Our hybrid model requires manual hyperparameter tuning in regards to how large the refined dataset should be. In addition, we are only testing on one user when we are running our test trials. We require a better method to determine this hyperparameter as manually tuning this hyperparameter is not the best method.

**How can we quantify the amount of food waste reduced from our project?**

Currently our models won’t be able to quantify this question. The value that we can put on this can only be determined through some sort of customer feedback loop once we implement the model. The issue is that someone could be using the recommender just to suggest a recipe. To counter this issue, prior to entering any sort of input ingredients and recommending any recipe, we can have an initial questionnaire regarding the use of the recommender. This may be a turn-off to some customers as it’s an additional step, but hopefully the environmental impact behind the use case will entice customers.