Food Pairing Recommendations

Bhavya Narnoli
bhavya21316@iiitd.ac.in

Nikita Rajesh Verma
nikita21546@iiitd.ac.in

Jeremiah Rokhum
jeremiah21533@iiitd.ac.in

***Abstract*—This study explores the intersection of culinary arts and computational techniques to analyze and visualize ingredient combinations for optimal food pairings. By leveraging natural language processing (NLP) and machine learning (ML), the project curates a dataset of 34,000 recipes sourced from online platforms. Using techniques like cosine similarity, and clustering algorithms such as KNN, autoencoders, and DBSCAN, the research identifies frequent ingredient pairings and underlying patterns in culinary data. Visualizations, including co-occurrence graphs, heatmaps, and word clouds, provide intuitive insights into the relationships between ingredients. Additionally, a recommendation tool hosted on a user-friendly website enables ingredient-based pairing suggestions, enhancing culinary creativity and decision-making. This work demonstrates the power of data-driven approaches in the culinary domain, benefiting chefs, food enthusiasts, and researchers.**

***Index Terms*—Food Pairing, Natural Language Processing (NLP), Machine Learning (ML), Recipe Dataset, Cosine Similarity, K-Nearest Neighbors (KNN), Autoencoders, DBSCAN, Data Visualization, Co-occurrence Graphs, Heatmaps, Ingredient Recommendation System.**

I. Introduction

Food pairing plays a pivotal role in culinary science, shaping the taste, texture, and overall appeal of a dish. With the advent of large-scale recipe datasets and advanced computational methods, data-driven approaches are transforming how we explore ingredient compatibility. This project harnesses natural language processing (NLP) and machine learning (ML) to analyze thousands of recipes, revealing hidden patterns and relationships among ingredients.

The key objective is to identify optimal ingredient pairings by extracting frequent combinations, clustering similar ingredients, and uncovering latent patterns in recipe data. Using a curated dataset sourced from platforms like AllRecipes, we employ a range of algorithms including Apriori, cosine similarity, K-Nearest Neighbors (KNN), autoencoders, DBSCAN, and t-SNE (t-distributed Stochastic Neighbor Embedding). These methods allow for dimensionality reduction and the identification of both frequent and subtle ingredient relationships.

The results are visualized through a variety of tools, including co-occurrence graphs, heatmaps, word clouds, and t-SNE plots. The t-SNE visualization, in particular, highlights clusters of ingredients in high-dimensional space, offering insights into potential substitutes and closely related ingredients. These visualizations, along with an interactive ingredient recommendation tool hosted on a website, empower users to explore ingredient combinations and optimize their culinary creations.

This project bridges computational techniques and culinary arts, providing chefs, food enthusiasts, and researchers with innovative tools to discover and create harmonious and creative dishes.

II. Literature Review

The **KitcheNette** study presents a data-driven method using Siamese neural networks to predict and rank ingredient pairings, trained on 300,000 pairing scores from 1 million recipes. By leveraging co-occurrence data and normalized pointwise mutual information, it outperforms traditional methods and baseline models in metrics like RMSE and correlation. The model discovers novel pairings, offers practical recommendations, and surpasses databases like FlavorDB, showcasing its potential to innovate culinary pairings for both experts and casual users.[1]

The paper titled **"Alternative-ingredient Recommendation Based on Co-occurrence Relation on Recipe Database"** presents a method for recommending alternative ingredients in recipes by analyzing co-occurrence relations within a recipe database. Two algorithms are proposed: one using a Naive Bayes filtering approach to prioritize compatibility between alternative and remaining ingredients in a recipe, and the other leveraging mutual information to assess the functional similarity between an exchange-ingredient and a candidate alternative. Through cooking demonstrations and subjective evaluations, the authors confirm the effectiveness of both methods in recommending acceptable alternatives, with future work focusing on enhancing ingredient modeling and expanding the scope to include nutritional and allergen considerations. The study emphasizes data-driven approaches for practical ingredient substitution without relying on predefined ontologies. [2]

The paper **"RecipeDB: A Resource for Exploring Recipes"** introduces a structured database of over 118,000 recipes from 74 countries, integrating culinary attributes, nutritional data, flavor profiles, and health associations. It enables scientific exploration of recipes and their impacts on nutrition and health by linking them to resources like USDA data, FlavorDB, and DietRx. RecipeDB uses advanced annotation methods to standardize recipes and offers powerful search functionalities for queries based on cuisine, ingredients, dietary styles, or nutrients. This resource advances computational gastronomy, facilitating data-driven studies on global culinary diversity and dietary patterns. [3]

III. Dataset

The dataset for this project comprises approximately 34,000 recipes sourced from the AllRecipes website, stored in a structured CSV format. Each recipe entry includes the recipe

name, URL, ingredient phrases, preparation instructions, and preparation time. Using named entity recognition (NER) and preprocessing techniques, individual ingredients were accurately extracted from the ingredient phrases, ensuring a clean and structured representation of the data. This process helped to eliminate noise and standardize the ingredient data for analysis.

The data collection pipeline relied on web scraping using Python libraries such as BeautifulSoup and requests. Recipe categories were identified and parsed, URLs of individual recipes were extracted, and structured recipe details were retrieved from embedded JSON schemas within the webpages. To address potential challenges, retry logic was implemented for handling HTTP errors, and optional proxy rotation ensured uninterrupted data scraping.

The resulting dataset is comprehensive, organized, and consistent, making it suitable for a wide range of NLP and machine learning tasks. It enables the identification of frequent ingredient pairings, the discovery of culinary patterns, and the clustering of similar ingredients. Serving as the backbone of this project, the dataset facilitates insightful analysis and visualization of ingredient combinations, contributing to a deeper understanding of optimal food pairings.

IV. Methodology

The methodology for this project involves several stages, from data collection to the application of machine learning models for ingredient pairing analysis and visualization. The process is divided into three main steps: data collection, data analysis, and visualization, followed by the development of a recommendation tool.

A. Data Collection

The dataset was collected using a web scraping approach with Python libraries such as BeautifulSoup and requests. Initially, recipe categories were identified by scraping the AllRecipes website. For each category, individual recipe URLs were extracted. The data from these recipes, including ingredient phrases, preparation instructions, and preparation times, were then retrieved from embedded JSON data within the webpages. Named entity recognition (NER) and preprocessing techniques were applied to extract and standardize individual ingredients from the ingredient phrases, ensuring consistency and accuracy in the dataset.

B. Data Analysis

Various machine learning techniques were employed to analyze ingredient pairings and uncover hidden patterns within the data. The following methods were used:

- **Association Rule Mining (Apriori)**: This technique was applied to identify frequently occurring ingredient combinations by analyzing co-occurrence patterns across the recipes. It helped in discovering common ingredient pairings based on their frequency in the dataset.
- **Cosine Similarity**: This method was used to measure the similarity between ingredient pairings based on their vector representations. It helped in finding ingredients that frequently appear together in recipes and measuring their compatibility.
- **K-Nearest Neighbors (KNN)**: KNN was utilized to find ingredient pairings by identifying similar recipes or ingredient clusters in high-dimensional space.
- **Autoencoders**: An autoencoder model was used to reduce the dimensionality of the dataset and capture latent features of ingredient pairings, enabling the discovery of complex relationships among ingredients.
- **DBSCAN**: This clustering algorithm was applied to group ingredients based on density, helping uncover non-linear and non-uniform ingredient pairing patterns.
- **t-SNE**: The t-distributed Stochastic Neighbor Embedding (t-SNE) technique was employed for dimensionality reduction, allowing the visualization of ingredient clusters in a lower-dimensional space. This helped identify similar ingredients and potential substitutes based on their distribution in the high-dimensional space.

C. Visualization

The results of the data analysis were visualized using several techniques to offer intuitive insights into ingredient pairings:

- **Co-occurrence Graphs**: These graphs were used to visualize relationships among the most common ingredients based on their frequency of co-occurrence in recipes.
- **Heatmaps**: Heatmaps were generated to represent the co-occurrence frequencies of ingredient pairs, helping identify strong connections between ingredients.
- **Word Clouds**: Word clouds were created to illustrate the frequency of ingredients across recipes, with the size of each word reflecting its occurrence.
- **t-SNE Plots**: The t-SNE algorithm was used to visualize ingredient clusters, helping identify groups of similar ingredients and potential substitutes.

D. Recommendation Tool

An interactive web-based tool was developed to allow users to input a list of ingredients and receive pairing suggestions based on the analysis. This tool leverages the insights derived from the machine learning models and visualizations, offering users personalized ingredient pairings. The website also displays the visualizations, enabling users to explore the data interactively and gain deeper insights into the relationships between ingredients.

In summary, this methodology integrates web scraping, machine learning, and data visualization to analyze and recommend optimal food pairings, providing valuable insights into ingredient combinations and enhancing culinary creativity.

V. Visualization and Plots

To provide an intuitive understanding of ingredient relationships and patterns, several visualization techniques were employed. These visualizations highlight ingredient pairings, co-occurrence frequencies, and clustering patterns, offering insights into optimal food pairings.

A. Ingredient Pairing Visualization

The ingredient pairing visualization is a co-occurrence graph that displays the relationships among the top n most common ingredients. Ingredients are represented as nodes, and lines (edges) connecting the nodes indicate the frequency of their co-occurrence in recipes. Thicker lines represent stronger associations, allowing users to quickly identify popular and complementary ingredient pairings. This graph is particularly useful for exploring high-level relationships between frequently used ingredients.

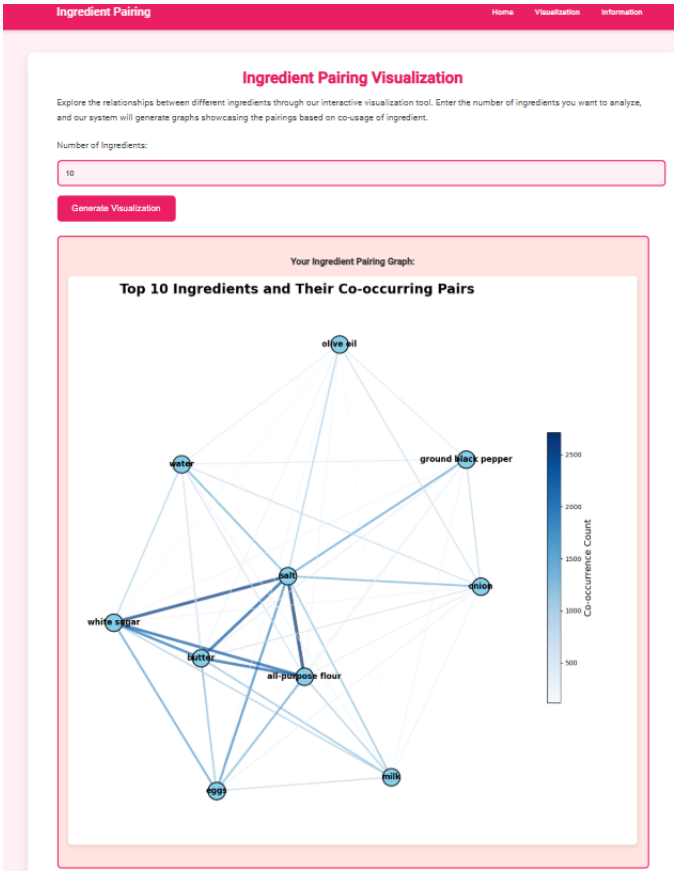

Fig. 1. Ingredient Pairing Visualization: A co-occurrence graph showing ingredient relationships and pairing frequencies.

B. Ingredient Co-occurrence Heatmap

The co-occurrence heatmap represents the frequency with which pairs of ingredients appear together in recipes. The heatmap is a grid where each cell corresponds to a pair of ingredients, with the intensity of the color indicating the frequency of their co-occurrence. This visualization provides a comprehensive view of ingredient relationships and helps identify strong and weak associations at a glance.

C. Ingredient Wordmap

The ingredient wordmap is a word cloud that visualizes the frequency of individual ingredients across the dataset. The size of each word corresponds to the frequency of its occurrence

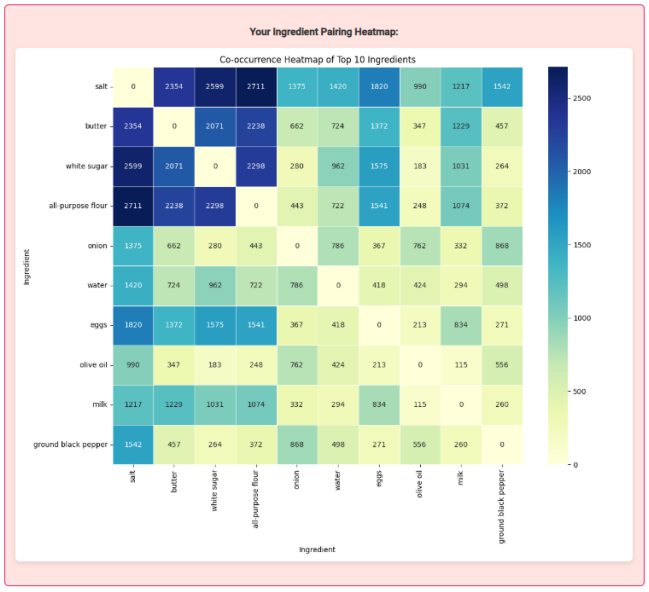

Fig. 2. Ingredient Co-occurrence Heatmap: A heatmap representing pairing frequencies of ingredients across recipes.

in recipes, with more common ingredients appearing larger. This visualization emphasizes key ingredients such as ”salt,” ”pepper,” and ”sugar” while also showcasing other frequently used components in cooking.

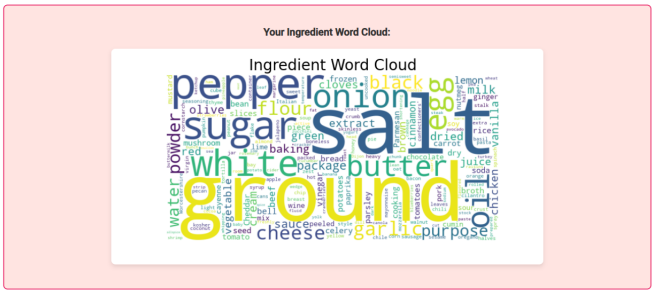

Fig. 3. Ingredient Wordmap: A word cloud showcasing the most frequently used ingredients in the dataset.

D. t-SNE Plot

The t-SNE (t-distributed Stochastic Neighbor Embedding) plot visualizes clusters of similar ingredients in a two-dimensional space. Ingredients that are frequently used together or share similar roles in recipes are grouped closer together, forming distinct clusters. This clustering highlights ingredient categories and potential substitutes, providing a deeper understanding of the relationships within the dataset. For instance, clusters may reveal groups of spices, dairy products, or baking ingredients, helping users identify alternative options within a recipe.

E. Interactive Visualizations

All these visualizations are integrated into an interactive web-based tool, designed to enhance user engagement and

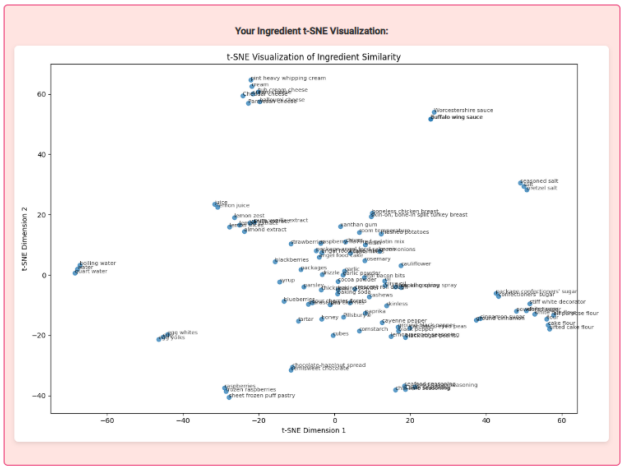

Fig. 4. t-SNE Plot: A clustering of similar ingredients, revealing categories and potential substitutes.

exploration. The tool allows users to:

- Input a specific number of top ingredients (n) to generate tailored visualizations for co-occurrence graphs, word clouds, heatmaps, and t-SNE plots.
- Dynamically filter and zoom into graphs to focus on specific ingredients or clusters.
- Access algorithm-based ingredient pairing recommendations for personalized insights.
- View an information page detailing the recipe data and methodologies used for analysis.

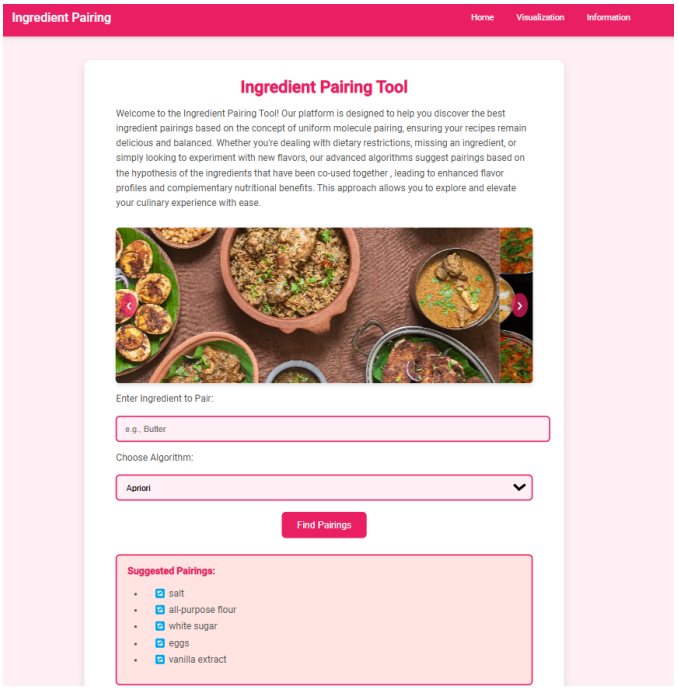

Fig. 5. Home Page: Interface for choosing algorithms and generating visualizations for ingredient relationships.

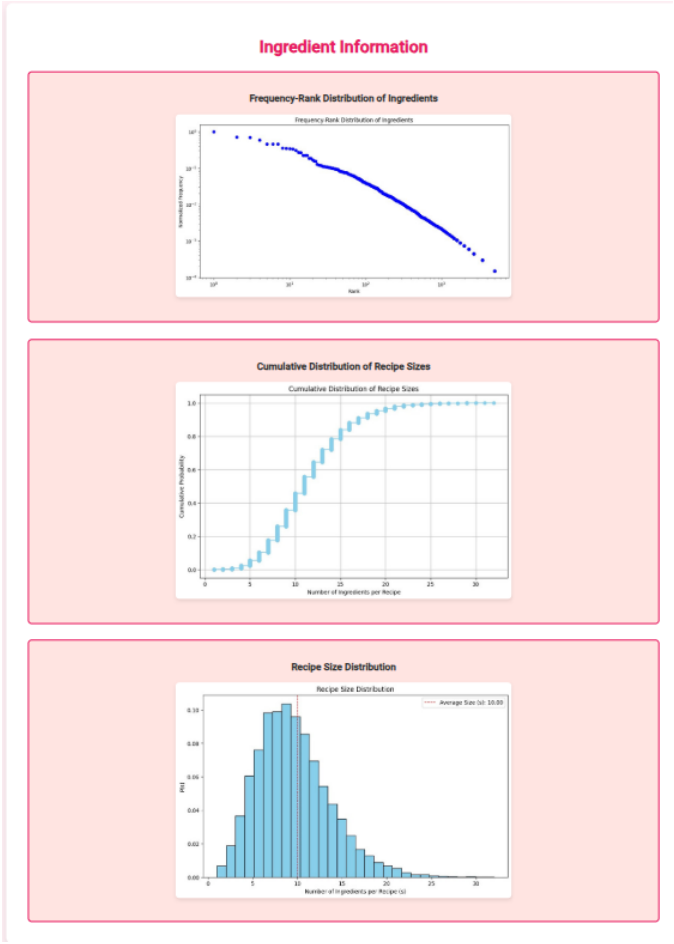

Fig. 6. Information Page: Details about the recipe dataset used in the project.

The website was developed using HTML, CSS, and JavaScript for a seamless user interface, with Flask serving as the backend for data processing and visualization generation. These interactive visualizations make the results accessible, actionable, and engaging for chefs, food enthusiasts, and researchers alike.

To run the website on a local server, clone the repository at https://github.com/bhavyanarnoli/cgas. Open the directory containing `app.py`, download the required dependencies from `requirements.txt`, and run the command `python app.py` to launch the server.

VI. Conclusion and Future Work

This project successfully integrates natural language processing (NLP) and machine learning (ML) to analyze ingredient combinations, uncover optimal food pairings, and visualize complex relationships in recipe datasets. By leveraging a curated dataset of 34,000 recipes and applying advanced analytical techniques such as association rule mining, cosine similarity, K-Nearest Neighbors (KNN), autoencoders, DBSCAN, and t-SNE, the study provides valuable insights into ingredient compatibility and clustering patterns.

The visualization tools, including co-occurrence graphs, heatmaps, word clouds, and t-SNE plots, offer an intuitive understanding of ingredient relationships and pairing frequencies. These tools, coupled with an interactive web-based recommendation system, empower chefs, food enthusiasts, and researchers to explore ingredient combinations dynamically, identify substitutes, and make data-driven culinary decisions.

The project demonstrates the potential of computational approaches in the culinary domain, bridging the gap between data science and gastronomy. It provides a scalable framework for analyzing ingredient relationships that can be extended to various cuisines, dietary restrictions, or personalized preferences. Future work could involve expanding the dataset, incorporating flavor chemistry data, or improving the recommendation system with user feedback to further refine ingredient pairing suggestions.

REFERENCES

[1] Donghyeon Park, Keonwoo Kim, "KitcheNette: Predicting and Ranking Food Ingredient Pairings using Siamese Neural Networks" *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, December 2019.

[2] Ryosuke Yamanishia, Naoki Shinoa, "Alternative-ingredient Recommendation Based on Co-occurrence Relation on Recipe Database", *19th International Conference on Knowledge based and Intelligent Information and Engineering Systems*,Procedia Computer Science 60 (2015) 986 – 993

[3] Devansh Batra, Nirav Diwan, *RecipeDB: a resource for exploring recipes*, Database, 2020, 1–10.