

CIS5500 Final Project Proposal

1. Team Members

- **Member 1:** Jiayang Chen, jiayangc@seas.upenn.edu, aaaaadamchen
 - **Member 2:** Yongyin Yang, yongyin@seas.upenn.edu, Pony-Lamb
 - **Member 3:** Yongqi Zhong, yongqizh@seas.upenn.edu, limzo-yq
 - **Member 4:** Shutong Jiang, ryjiang@seas.upenn.edu, pennshutongjiang
-

2. Project Overview

Objective

YumLog is an AI-driven restaurant review and food discovery platform designed to enhance the dining experience by enabling users to search for restaurants, receive personalized meal recommendations, submit reviews, and access detailed recipes. Through the integration of advanced recommendation algorithms, comprehensive restaurant databases, and interactive social engagement features, the platform offers a data-driven, user-centric, and seamless approach to culinary exploration.

Core Functionalities

- **Restaurant Discovery & Recommendations** – Enables users to search for restaurants based on location, cuisine, dietary preferences, and pricing tiers.
 - **User Reviews & Ratings** – Allows users to write restaurant reviews, rate their dining experiences, and upload food photos to share real experiences.
 - **AI-Powered Meal Suggestions** – Generates personalized restaurant and dish recommendations based on user preferences, dining history, and ratings.
 - **Recipe Retrieval Feature** – Enables users to find detailed recipes and cooking instructions by matching dish names from restaurant menus to a recipe database, making it easy to recreate restaurant meals at home.
 - **Community Engagement & Social Features** – Facilitates social interaction by allowing users to follow food influencers, participate in recipe replication discussions, and share cooking experiences.
-

3. Data Sources

The platform integrates **two large, interrelated open datasets** to facilitate restaurant search functionality and meal replication features.

Dataset 1: Yelp Open Dataset (Restaurant & Review Data)

- **Description:** A publicly available dataset encompassing real-world data related to businesses including reviews, photos, check-ins, and attributes like hours, parking availability, and ambiance.
- **Source:** [Yelp Open Dataset](#)
- **Size & Attributes:**
 - **Table business: Over 6 million reviews from 150,000+ businesses.**
 - Attributes: `business_id`, `name`, `address`, `latitude`, `longitude`, `categories`, `stars` (rating), `review_count`, `menu_items`.
 - Statistics: `stars`(mean 3.63, std 1), `review_count`(mean 30.1, std 98.2)
 - **Table user: Over 1 million users**
 - Attributes: `user_id`, `name`, `review_count`, `yelping_since`, `friends`, `fans`, `average_stars`
 - Statistics: `review_count`(mean 23.1, std 79.1), `fans`(mean 1.46, std 14.3), `average_stars`(mean 3.71, std 1.12)
 - **Table tip: Over 1 million tips**
 - Attributes: `text`, `date`, `likes`, `business_id`, `user_id`
 - Statistics: `likes`(mean 0.02, std 0.15)
 - **Table checkin: Over 3.9 million logs of the number of checkins** for different businesses in different time slots
 - Attributes: `business_id`, `weekday`, `hour`, `checkins`
 - Statistics: `checkins`(mean 4.26, std 13.9)

Dataset 2: RecipeNLG Dataset (Recipes & Ingredients Data)

- **Description:** A dataset comprising over 2 million cooking recipes with structured ingredient breakdowns, commonly available grocery items, and preparation methodologies.
- **Source:** [RecipeNLG Dataset](#)
- **Size & Attributes:**
 - **Table Recipe:** Approximately **2.2 million recipes**, more than **2 million ingredients**.
 - Attributes: `title`, `ingredients`, `instructions`, `cuisine`, `meal_type`.

Data Integration Strategy

- **Entity Resolution:** Utilizing NLP techniques to establish relationships between restaurant menu items and comparable recipes within the dataset.
 - **Normalization & Data Cleaning:** Addressing missing values, standardizing formats, and eliminating redundancies using **Pandas and SQLAlchemy**.
 - **Database Design:** Implementing a **relational database model (PostgreSQL/MySQL)** to ensure optimized storage, retrieval, and indexing.
-

4. SQL Queries

Our project may include the following queries (but not limited to):

1. **Retrieve Nearby Restaurants** – Identify restaurants serving vegetarian-friendly dishes within a 5-mile radius.
 2. **Personalized Recommendations** – Generate restaurant suggestions based on a user's past dining history and rating patterns.
 3. **Recipe Matching** – Retrieve recipes that correspond to a specific dish available on a restaurant menu.
 4. **Find a Recipe to Replicate a Restaurant Dish** – Match a dish name from a restaurant menu to a similar recipe from the recipe database, including ingredients and step-by-step instructions.
 5. **Data Insights & Analytics** – Identify the most popular dishes based on user reviews and provide alternative at-home cooking solutions.
-