Labeling Recipes

Sarah Buss

Interest: 10/10

Project Description

Something I find difficult sometimes is constantly picking recipes to make for breakfast, lunch, and dinner. I think it would be a useful project to build a model that can automatically label recipes based on given criteria to help users pick the best recipe for their meal. For example, some people have allergies and have to carefully look through the ingredients before making a recipe. It would be a lot easier if that recipe could be automatically labeled and returned in a search as “gluten-free” or “dairy-free”. The same concept applies to those who are vegetarian, vegan, or just want a specific type of meal (breakfast, beef-based, rice-based, few ingredients, short cook time, etc.).

Potential Features

In order to make this project doable in a semester, I would select a small subset of potential labels and criteria. The labels “breakfast”, “lunch”, and “dinner” seem to be a reasonable subset that would allow the model to learn categorical patterns rather than just search for specific ingredients. We could potentially add other labels as well such as “dessert”, “drink”, or “snack” to add clarity. The model would primarily focus on the ingredients and the description of the recipe. The data would be stored in the following format:

recipe_id	recipe_title	ingredients	description	category
1234	“Omelete”	[“eggs”, ”salt”, “pepper”, “cheese”, “milk”]	“This simple omelete is the perfect way to start off your morning! You…”	“breakfast”

Data Gathering and Labeling

Most of the data from this project could be pulled from Kaggle using the “Food.com - Recipes and Reviews” dataset: https://www.kaggle.com/datasets/irkaal/foodcom-recipes-and-reviews. This dataset includes the ids, titles, ingredients, descriptions, and more for over 500,000 recipes, although I estimate we would only use about 200 of these recipes.

We could edit the downloaded csv file manually to add the categories or use python and scikit-learn to add this additional attribute to each of the recipes we select. It might also be beneficial to remove the attributes we’re not using to make the dataset load faster.