Anamika Lal

BrainStation Capstone Sprint 1

Recipe Recommender

Overview

- Build a recipe recommendation system such that it can:
  - Help reduce food waste
  - Promote mindful / conscious eating
  - Encourage users to try new recipes

Proposed Vision using ML

- "Reduce food waste" –
  - Provide ideas on how to use the ingredients a user already has.
  - 'Rank' recipes that use whole ingredients instead of partial ones higher. (e.g. 1 bunch of celery instead of 1 stalk)
  - 'Rank' recipes that have less ingredients higher.

- "Promote mindful / conscious eating" –
  - By taking into account calorie / nutrition info when making recommendations.

Proposed Vision using ML (cont.)

- "Encourage users to try new recipes" –
  - By recommending options based on their preferences e.g. Diet-type (Vegan, gluten-free etc.)

Impact

- Reduced Food Waste –
  - By promoting efficient use of ingredients and leftovers.
  - Contributes to sustainability efforts and addresses a societal concern.

- Promotion of Healthy Eating –
  - Encourage users to explore and adapt healthier eating habits.
  - Aligns with the growing social emphasis on wellness and nutrition.

Datasets - Criteria

- Criteria for a good dataset:
  - List and quantity of ingredients for a recipe.
  - Calorie / Nutrition information.
  - Rating for the recipe.
  - Steps to make the recipe.
  - Serving size.
  - Diet – type (Vegan, gluten-free etc.)

Datasets – Search and Compare

B	D	E	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
Link	Size - # rows	# columns	Contains ingredient lists	Contains calorie info / nutrition info	Contains serving size info	Contains diet type (Vegan etc)	Contains cuisine type (italian etc)	Has ratings?	Has # of ingredients	Has recipe steps	Has quantity of each ingredient?	EDA attempted?	Additional Notes	Findings		
Epicurious - Recipes with Rating and Nutrition	~20,000	680 OR 11 (JSON)	Yes	Yes	No	No	No	Yes, in same table	Not explicitly	Yes	Yes, in recipe text	Yes	Less # of recipes Kaggle License says Unknown	54% has rating > 4.0 One-hot encoded table lost ingredient amoun		
	~16,000 after removing nulls															
Food.com - Recipes and Reviews	~430,000	28	Yes	Yes - calorie AND nutrition	Yes	No	No	Yes, in same table	Not explicitly	Yes	Yes, but no units	Yes	This has 2 parquet files, how to work with those?	92% of recipes have rating > 4.0		
The above seems for R ?	~170,000 after removing nulls															
Food.com Recipes and Interactions	~230,000 in raw recip	12 in raw_recipes	Yes	Yes	No	(May be able to impute from ingredients)	No	Yes, in different table	Yes	Yes	No	Yes - data is clea	Has tables that have made data numeric Some tokenization already done			
	~178,000 in pp_recip	8 in pp_recipes														
	~1.1 million in raw_in	5														
	~25,000 in pp_users	6														
~~Food.com Recipes with Search Terms and Tags~~	~~~500,000~~	~~10~~	~~Yes~~	~~No~~	~~Yes~~	~~No~~	~~No~~	~~No~~	~~Not explicitly~~	~~Yes~~	~~Yes~~		~~Calories will have to be calculated?~~ ~~No ratings!~~	~~No null values for relevant cols~~ ~~No duplicate rows~~ ~~Largest number of recipes~~		
Diets, Recipes And Their Nutrients	~8,000	8	No	No	No	Yes	Yes	No	No	No			Can this be used to train for cuisine type? ??			
~~Food Ingredients~~	~~~9,000~~	~~29~~	~~NAH~~													
Nutritional values for common foods	~8000	77														

Datasets - Candidates

- Food.com – Recipes and Reviews (Kaggle)
  - 170K recipes after dropping null values.
  - Contains serving size.
  - Lots of nutritional information.
  - Does not have measurement for ingredients.
  - 92% of the dataset has ratings > 4.0

- Epicurious – Recipes with Rating and Nutrition (Kaggle)
  - 16K recipes after dropping null values.
  - Does not have serving size.
  - 54% of dataset has ratings > 4.0

Next Steps

- Decide on one dataset from the 2 that have been shortlisted.
- Data processing to make the relevant columns numeric.