Final Project Proposal | CIS 550

1. A List of group members, email addresses, and GitHub usernames

Name	Email Address	GitHub username
Zhihang Yuan	zhyuan1@seas.upenn.edu	zhyuan11
Yuxi Dai	yuxidai@seas.upenn.edu	linda-XI
Yuxuan LI	liyux@seas.upenn.edu	Yuxuan-Li295
Yuyan Zhou	zhouyy@seas.upenn.edu	zhouyuyann

https://github.com/zhyuan11/CIS550-Project

2. A Description of application/website idea
   A website that can
   1. Recommend suitable recipes according to ingredients, cooking time, nutrition, and a number of steps, name of cook.
   2. View people's reviews of each recipe.
   3. Recommend suitable pastry recipes with pictures of it according to ingredients, cooking time, nutrition, and a number of steps, name of cook.

3. For each dataset you've chosen:

   a. A 1-2 sentence description of the dataset
      i. The first link, Recipe Ingredients and Reviews, contains two datasets. The dataset is about recipes of pastry food, such as bread, pizza, cake, and cookies. And the second dataset is about their reviews.
      ii. The second link Recipes and Interactions, contains two datasets. The first is about all kinds of recipes. And the second one is about their corresponding users' reviews.

   b. A link to where you found the dataset
      i. Recipe Ingredients and Reviews
      ii. Recipes and Interactions

   c. If you're scraping the data, a description of how you will scrape it
      N/A

d. If you're not scraping the data:
    i. Relevant size statistics (e.g. For a table, mb/gb, number of rows, and number of attributes. For a graph, mb, the number of nodes, and the number of edges.

Dataset	Size of File	# rows	# attributes
Raw interactions	349MB	1125284	5
Raw recipes	350MB	115284	4
Recipes	16MB	18544	10
Reviews	350MB	1616884	4

    ii. Summary statistics of several attributes. (e.g., report mean, standard deviation)

    The dataset mainly contains attributes that are string types.
    There's nothing regarding the statistics.

4. A list of at least 5 queries (in natural language) you could write for your datasets. Some of these should require complex SQL (aggregations, subqueries, joins, etc.

1. Shows all recipes and all ingredients.
2. Search recipes by certain ingredients.
3. Show all the reviews of a recipe.
4. Search recipes by a given range of cook time.
5. Filter recipes according to rating.
6. Group by the recipes by the nutritions / raw ingredients.
7. Join the recipes with their corresponding reviews.