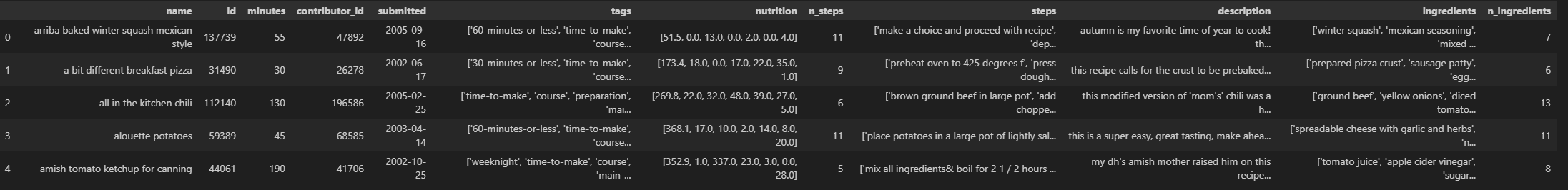
**Exploratory Analysis of Food.com Recipe Scrape**

Joshua Grou, Jason Bauer

**Dataset**

The project of creating a personal chef cognitive assistant will need a large amount of data around what ingredients go with what recipes in order to serve accurate recommendations for recipes to create with the users on hand ingredients. To serve this goal a dataset was found that contains ingredient lists and other metadata for over 230K recipes taken from food.com. This dataset was sourced from Kaggle. The dataset contains twelve columns displayed below. There are a few columns that may be of particular interest to this project. The name column contains a string with the name of the recipe, similarly the description column contains a string describing the recipe. The ingredients column contains a list of ingredients required to make the recipe and the n\_ingredients column counts that list. Similarly the steps and n\_steps column contain the number and description of steps needed to make the dish. The tags and nutrition column may be interesting to explore as some of the tags indicate time expectation and the nutrition column may be utilized to cater to specific dietary needs. The following analysis will mostly be focused on the ingredients column as utilizing this column will be essential to building the personal chef cognitive assistant.



**Analysis of Ingredients**

The ingredients column was first analyzed to understand what types of ingredients are utilized heavily or any trends that may pose some challenges. First the column was cleaned and frequency of ingredients was counted. The top 20 ingredients were plotted as shown below.

A graph of ingredients

Description automatically generated

By far the most common ingredient was salt, appearing in over 80K recipes. This also uncovered some problems with the dataset that will need to be addressed. The obvious problem is that certain ingredients are duplicated such as salt, pepper, and salt and pepper, egg and eggs, garlic cloves and garlic. This presents a challenge as for instance if a user tells the application that they have garlic it is likely that the recipes containing garlic cloves should not be excluded from the search. A slightly less obvious challenge may be that users might not recognize certain ingredients as recipe components, for instance salt is something that is present in every kitchen and thus users may not explicitly tell the application that they have salt but again recipes containing salt likely should not be excluded unless the user has identified some dietary restriction around salt.

Following the analysis of the ingredient frequency investigation on the correlation with other variables was done. This returned a perhaps expected finding that the number of ingredients is weakly positively correlated with the number of steps for the recipe indicating that in general the more ingredients that a recipe contains. This could be useful in recommending recipes if for instance a user indicates that they are looking for something simple to make a lower ingredient count recipe will in general yield a recipe with less steps to complete.

A screenshot of a computer

Description automatically generated