Data Generating Process:

Foodnetwork.com has thousands of recipes available on the website and they sort the recipes by alphabetical orders.

I generated 6411 recipes from the website using web scraping in python. Based on the total pages of recipe lists within each alphabetical letter, I proportionally and randomly selected recipes from each page of recipe list under each alphabetical letter. Therefore, my data is pretty unbiased.

(Note: I could have generated more data but just for preliminary analysis, I think around 6411 recipes are good enough to have some basic understanding)

The Data I generated includes recipe ingredients and their corresponding amount and measurement unit if those apply. I also download around 500+ ingredient words as a food dictionary, as well as some sauce dictionary and spice dictionary from Wikipedia.

Foodnetwork.com also provides some basic recipe labels like main ingredient, what type of dish (breakfast, dessert, lunch or main dish), as well as some basic cooking method. I analyzed the main ingredient label and plot a wordcloud with main ingredient appearance frequency among recipes proportional to the font size in the wordcloud plot.

For each recipe, my data process will return a tuple with (‘amount’,’unit’,’ingredient’). Since in the wordcloud, most recipes use chicken as main ingredient, l narrow down my data to the ingredients along with chicken. By listing top ten or so ingredients used with chicken, I could generate the normalized amount distribution and recommend the optional amount with those ingredients. Therefore, later on for my project I can recommend ingredients based on user’s preference. Moreover It could be generalized by using different way to categorize recipes.