Hello everyone, I'm czk from G01, let me introduce to you the acquisition and processing of our project data.

First, We found a data set containing many recipes on a machine learning website called Kaggle. This website has a data set containing a lot of recipes, and we downloaded one as our original data set. This data set contains a lot of recipes, each recipe has a specific id, cuisine and contained ingredients. This is the data structure of a recipe, and I take this for a example, the id of this recipe is 38898, the cuisine is Korean, the ingredients is a list, which includes sugar, ginger, garlic cloves and so on.

Data processing is divided into three steps, first select data set,the second one is preliminary data processing,the third one is relabel.let me introduce select data set first,because there are a total of 49718 recipes in the original data set,and we select some of them as our data set. As you can see, the final data set has 56 recipes and 341 different ingredients.

Then is preliminary data processing.Because the recipe contains many ingredients, the same ingredient may also have singular and plural,tense and other changes. In order to eliminate such differences,we consider filtering ingredients.

First we use regular expression to remove the punctuation of words, and only keep English alphabet.you can see this example, the comma in the name of the ingredient is removed after processing

Second we use WordNetLemmatizer to remove the singular and plural, tense, and only keep the stem of the word.you can also see, after processing, the plural in the name of the ingredient becomes singular, We have done the same treatment to all the ingredients.

Third is relabeling, in all the ingredients, there are many ingredients that are not very different but have different names, so we have relabeled these ingredients, for example, there are water,boiling water,ice water,cold water,warm water and luke warm water,there are so many water,so we relabeled these ingredients and gave these similar ingredients a common label: water.

These are our data processing parts, and then my teammates will continue to introduce our project.