The two-classification datasets we got were obtained from the Kaggle dataset website. The data set contains 2 csv files, one for positive comments and one for negative comments.
Positive file name = 'pos_english.xls'
Negative file name = 'neg_english.xls'
Each data file contains 25,000 comments. In the processing of the file, our processing steps are as follows.

1. Read all the sentences in these two folders, using the nltk.word_tokenize() function to divide the words into words, save them as a list This is the word set for the machine learning training.

2. Give all words that appear in the 2 files a unique number tag, removed duplicate words, sort by the number of occurrences, these numbers will be used to generate the sentences vectors.

3. Comments on 2 files (positive and negative files) are merged into one file, positive comments and negative comments will be added label 1 and 0, and finally random scramble order.

4. For the merged comment, the entire sentence is represented by the corresponding number (the entire sentence is digitized), and saved in all_['doc2num'] column. At this time, the all_ pandas file has a total of 50,000 rows and 4 columns, each column is 0,label,words and doc2num. This pandas type file is used to import machine learning in the LSTM framework, and finally train the model and save it as 'my_new_english_model_10_epoch.h5'. When you need to use the model for prediction, you only need to load the model.