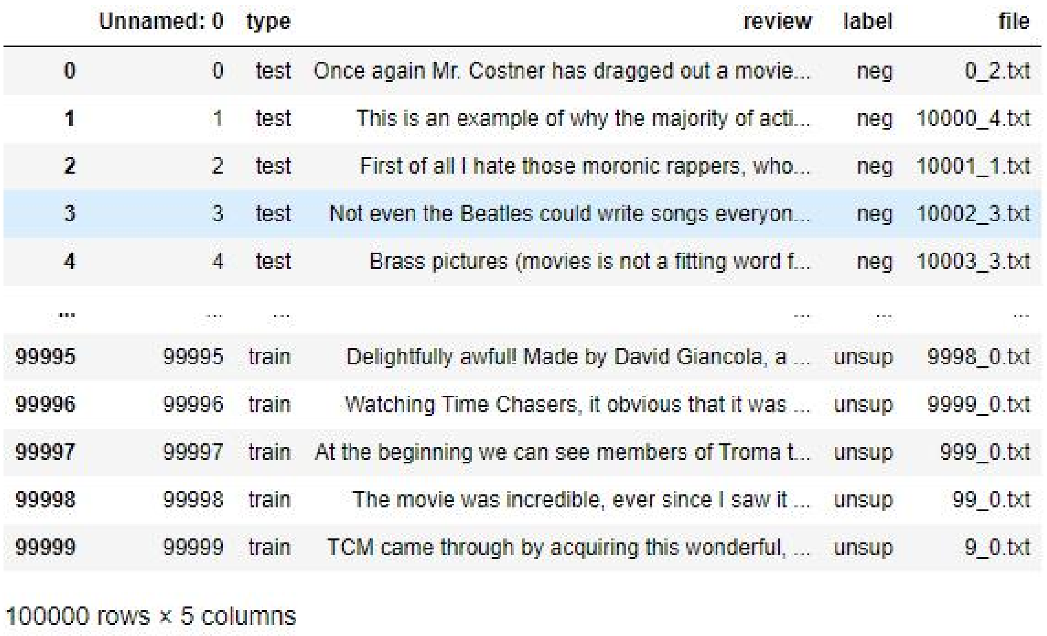
A sentiment analysis system based on the Naïve Bayes classifier

In this assignment, I constructed a sentiment analysis system based on Naive Bayesian classifier. I used two steps in data pre-processing. 1. Stop word removal, that is to remove punctuation marks and some words that have little impact on the prediction results from the original data. 2. Extracting word stems, that is, extracting words with the same meaning but different spelling, so as to improve the accuracy of the system in identifying words. Then, according to the naive Bayesian formula, I wrote the training module and prediction module to predict the data of the test set. Then I independently wrote an F1 function to evaluate the performance of the classifier. Finally, I compare the results of using the text preprocessing technology mentioned above or not. The results show that stop word removal and stem extraction can significantly improve the effect of the classifier.

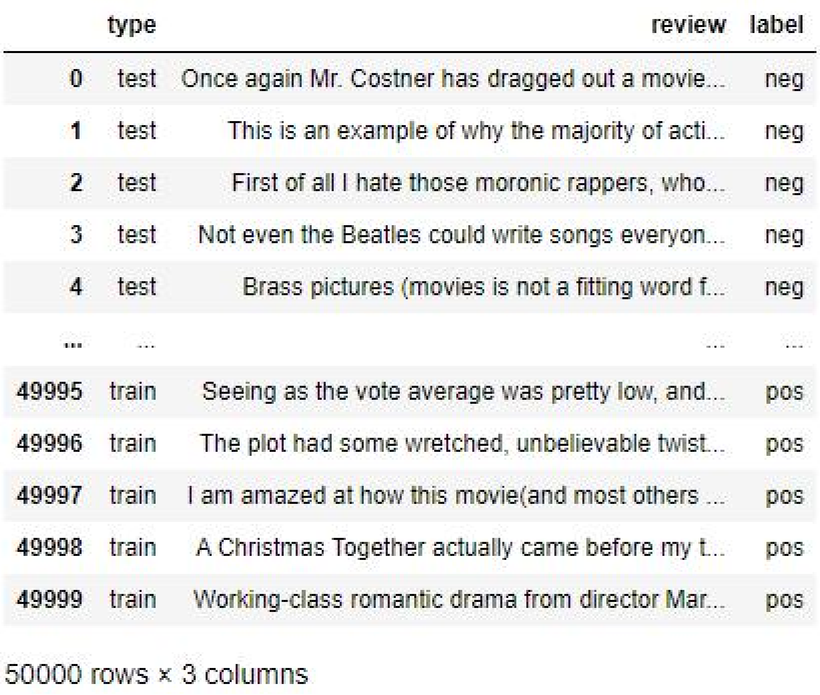
1. data pre-processing



	Unnamed: 0	type	review	label	file
0	0	test	Once again Mr. Costner has dragged out a movie...	neg	0_2.txt
1	1	test	This is an example of why the majority of acti...	neg	10000_4.txt
2	2	test	First of all I hate those moronic rappers, who...	neg	10001_1.txt
3	3	test	Not even the Beatles could write songs everyon...	neg	10002_3.txt
4	4	test	Brass pictures (movies is not a fitting word f...	neg	10003_3.txt
...	...	...	...	...	...
99995	99995	train	Delightfully awful! Made by David Giancola, a ...	unsup	9998_0.txt
99996	99996	train	Watching Time Chasers, it obvious that it was ...	unsup	9999_0.txt
99997	99997	train	At the beginning we can see members of Troma t...	unsup	999_0.txt
99998	99998	train	The movie was incredible, ever since I saw it ...	unsup	99_0.txt
99999	99999	train	TCM came through by acquiring this wonderful, ...	unsup	9_0.txt

100000 rows × 5 columns

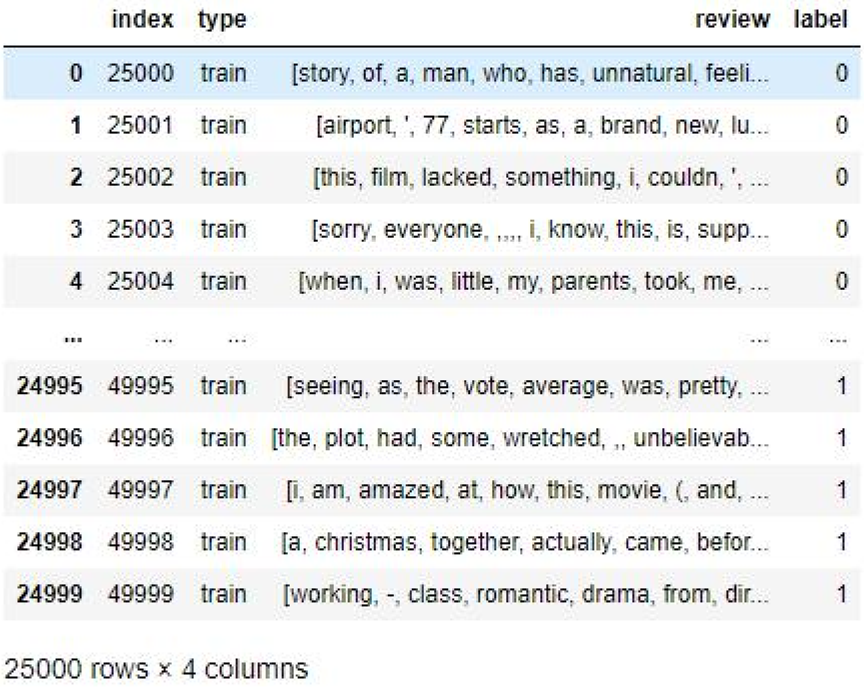
First, we use pandas to read the CSV dataset. After observation, it can be found that there are five columns of data in the data set: index, type, review, label and file. According to the requirements of the job, the only data I need to use is type, review and label, and I don't need to use the data of label which is “unsup”. So I use the filtering function of pandas and only keep the data mentioned above.



	type	review	label
0	test	Once again Mr. Costner has dragged out a movie...	neg
1	test	This is an example of why the majority of acti...	neg
2	test	First of all I hate those moronic rappers, who...	neg
3	test	Not even the Beatles could write songs everyon...	neg
4	test	Brass pictures (movies is not a fitting word f...	neg
...	...	...	...
49995	train	Seeing as the vote average was pretty low, and...	pos
49996	train	The plot had some wretched, unbelievable twist...	pos
49997	train	I am amazed at how this movie(and most others ...	pos
49998	train	A Christmas Together actually came before my t...	pos
49999	train	Working-class romantic drama from director Mar...	pos

50000 rows × 3 columns

Next, for further processing of the data. To compare the results of using and not using the stop word removal step and stem extraction step. I added two Boolean parameters in the data pre-processing module to set whether to use. In the step of removing stop words, I used the wordpunch tokenizer tool in nltk library, and in the step of extracting stem, I used the porterstemmer tool in nltk library. At the same time, before processing, I also convert all the data into lowercase in advance to avoid double calculation. Finally, I divided the data into train_data and test_data according to the value of type.



	index	type	review	label
0	25000	train	[story, of, a, man, who, has, unnatural, feeli...	0
1	25001	train	[airport, ', 77, starts, as, a, brand, new, lu...	0
2	25002	train	[this, film, lacked, something, i, couldn, ', ...	0
3	25003	train	[sorry, everyone, ,,,, i, know, this, is, supp...	0
4	25004	train	[when, i, was, little, my, parents, took, me, ...	0
...	...	...	...	...
24995	49995	train	[seeing, as, the, vote, average, was, pretty, ...	1
24996	49996	train	[the, plot, had, some, wretched, ,, unbelievab...	1
24997	49997	train	[i, am, amazed, at, how, this, movie, (, and, ...	1
24998	49998	train	[a, christmas, together, actually, came, befor...	1
24999	49999	train	[working, -, class, romantic, drama, from, dir...	1

25000 rows × 4 columns

2. Naive-Bayes classifier

In this step, I used a lot of Bayes related formulas from the course. Among them, the values to be calculated in the training module are: 1 The number and probability when review is positive. 2. The number and probability when review is negative. 3. Number and frequency of words when there is positive feedback. 4. Number and frequency of words when negative feedback occurs. 5. Total number of words. Formulas used are as follows:

$$\log \Pi_{i\in positions} P(w_i|c)P(c) = \sum_{i\in positions} \log P(w_i|c) + \log P(c)$$

$$\sum_{i\in positions} \log P(w_i|c) + \log P(c) = \sum_{k\in|V|} n_k \log P(w_k|c) + \log P(c)$$

$$P(w_k|c) = \frac{count(w_k,c)}{\sum_w count(w,c)} \qquad P(c) = \frac{N_c}{N_{doc}}$$

By observing the data, it is found that the direct use of naive Bayes formula will affect the prediction results because the data set is incomplete and some words in the test set fail to appear in the training set. Therefore, we need to use the smooth method to give a basic occurrence rate of each word, using the following formula.

$$P(w_k|c) = \frac{count(w_k,c)+1}{\sum_w count(w,c)+|V|}$$

Finally, by combining the data obtained from the training set, we can get the positive score and negative score of the training set according to the formula. By comparing the two values, we can finally predict whether the test set belongs to POS or NEG.

3. F1 function

After getting a reasonable prediction, we also need a function that can evaluate the quality of the model. F1 score is a measure of classification problems. Some machine learning competitions of multi classification problems often take F1 score as the final evaluation method. It is the harmonic average of accuracy rate and recall rate, with a maximum of 1 and a minimum of 0.

Before writing F1 function, we need to determine several concepts, namely precision and recall. Precision refers to the proportion of positive samples in the positive example determined by the classifier. Recall refers to the proportion of the predicted positive cases in the total positive cases. By comparing the two values, we can get a parameter to evaluate the quality of the model. The formula used is as follows:

$$f1_k = \frac{2 \cdot precision_k \cdot recall_k}{precision_k + recall_k}$$

4. Analysis of various factors

By comparing the data, we get that the F1 score is 0.9077 when the stop word removal and stem extraction steps are not used for the first time. The F1 score was 0.9111 when only stop word removal was used for the second time without stem extraction step. When the stop word removal and stem extraction steps were used at the same time for the third time, the F1 score was 0.9081. It can be found that the use of stop word removal will greatly improve the accuracy of classification, and the help of stem extraction is very limited.

A1808830 Zhengjie He
2022/3/20