**NBC experimental report**

name：ping yu ID：201814848

**1. Key point:**

**1.1 Naïve Bayes**

Assume a target function with f: X🡪V, where each instance x is described by <x1, x2, …, xn>. Most probable value of f(x) is:

Using the Naïve Bayes assumption:

**1.2 Smoothing**

To overcome the following issue:

Smoothing:

**1.3 Simplified Calculation**

↓

**2. Experiment steps**

**Step 1:** Data preprocessing: Tokenization, Normalization, Stemming, Remove stopwords.

**Step 2：**Divide the data into training data and testing data. Testing data account for 20% of the data.

**Step 3:**On the base of the data set structure, store the training set in the following data structure:

List 1 = [{a:1,…, b:2},…,{a:2, …,b:3, …,c=2}]

The list contains 20 objects and each object is a class of training set. Each dictionary in the list shows the number of every word in this class.

Store the testing set in the following data structure:

List 2 = [[{},…,{}]…[{a:1,…, b:2},…,{a:2, …,b:3, …,c=2}]]

The outer list contains 20 objects and each object is a class of testing set. The length of each inner list shows the number of the document in this class. Each dictionary demonstrates the weights of all words in a document.

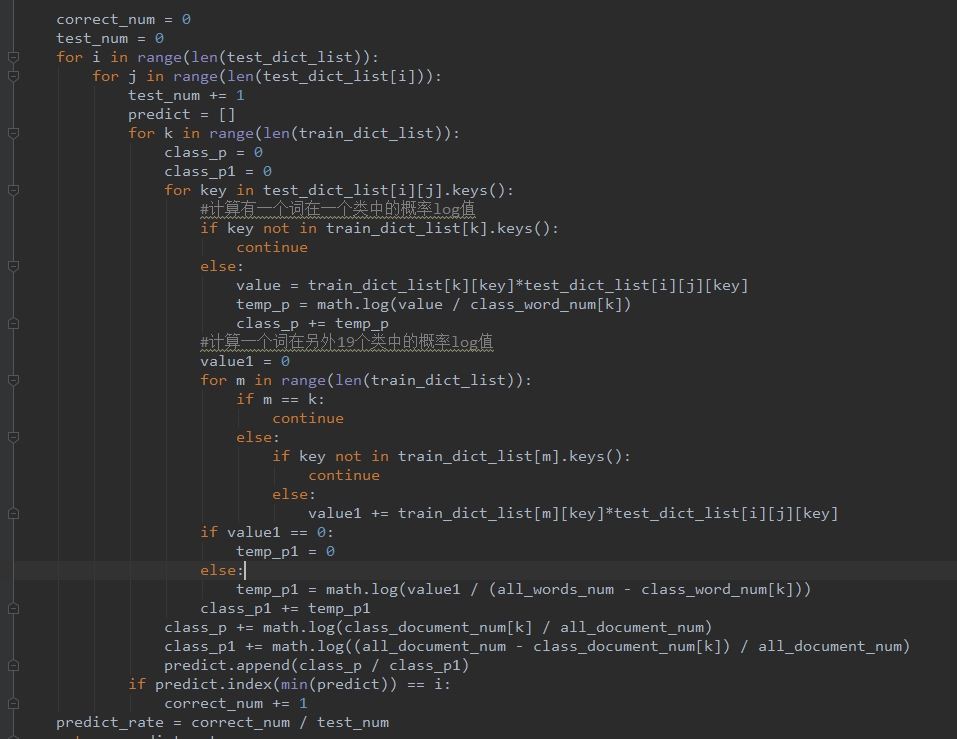
**Step 4:**

For each document in testing set, compute the weight the document belongs to each class.

The weight(W) a document belongs to a class:

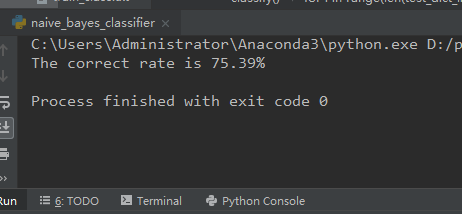
Classify the document to the class corresponding to the min(W).

Related code:



**3. Experiment result：**

**The final correct rate is 75.39% when I use the function above.**



**4. Conclusion**

Based on the results of this experiment alone, the accuracy of Bayes Classifier is lower than K-NN under the same preprocessing of the same specific data set.