**Datamining HW#1**

**201433687 선승연**

**Brief experiment description**

We divide 16000 documents out of 20000 documents by using tokens, and then learn the words that come out from each category. Then, after dividing the remaining 4000 documents using tokens, we predict the category by word only with the data we have learned.

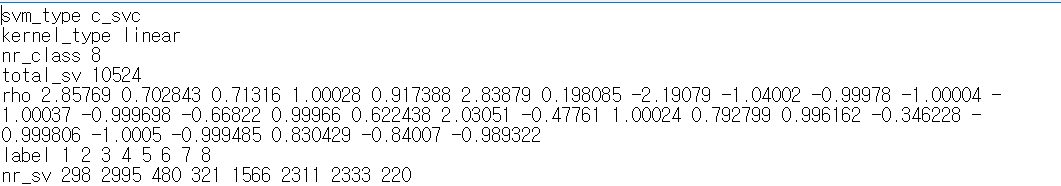
**Pretreatment explanation**

Token: .,“’-\n{}[]()<>

Algorithm: Using ‘dictionary’ to store the number of words in each category when used chi-square

**Performance analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| Language | Train / test time | Line | Accuracy |
| Python | 6 min / 1.25 min | 500 | 63.4438 |



In the model.txt file generated by svm, what total\_sv means is the number of features referenced to categorize. I can see that the categories are categorized by referring to about 10,000 features.

**Error analysis**

False positive & False negative

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PREDICTED VALUE | | | | | | | | |  |
| ACTUAL VALUE |  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | False negative |
| **1** | 41 | 22 | 7 | 3 | 6 | 15 | 7 | 0 | 41.41 |
| **2** | 4 | 939 | 7 | 2 | 27 | 54 | 200 | 2 | 76.03 |
| **3** | 7 | 29 | 38 | 9 | 20 | 15 | 22 | 0 | 27.14 |
| **4** | 4 | 20 | 4 | 56 | 7 | 8 | 3 | 2 | 53.85 |
| **5** | 9 | 116 | 9 | 7 | 342 | 52 | 45 | 16 | 57.38 |
| **6** | 8 | 147 | 3 | 4 | 36 | 659 | 25 | 1 | 74.63 |
| **7** | 15 | 311 | 22 | 6 | 49 | 34 | 439 | 3 | 49.94 |
| **8** | 1 | 17 | 1 | 0 | 10 | 5 | 9 | 32 | 42.67 |
|  | False positive | 46.07 | 58.65 | 41.76 | 64.37 | 68.81 | 78.27 | 58.53 | 57.14 |  |

Conclusion

There was an error counting nine documents when testing. I will improve this part in the future and then think about how to split the features and make the code a bit more concise, so that I can improve performance.