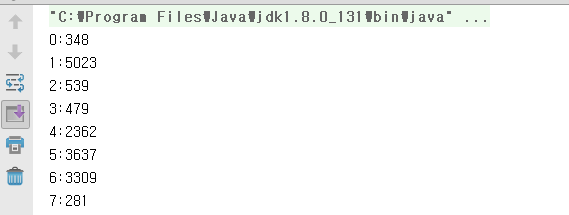
**Data Mining Hw 2**

**201233334 조남진**

**CPU :** 2.3 GHz Intel Core i3

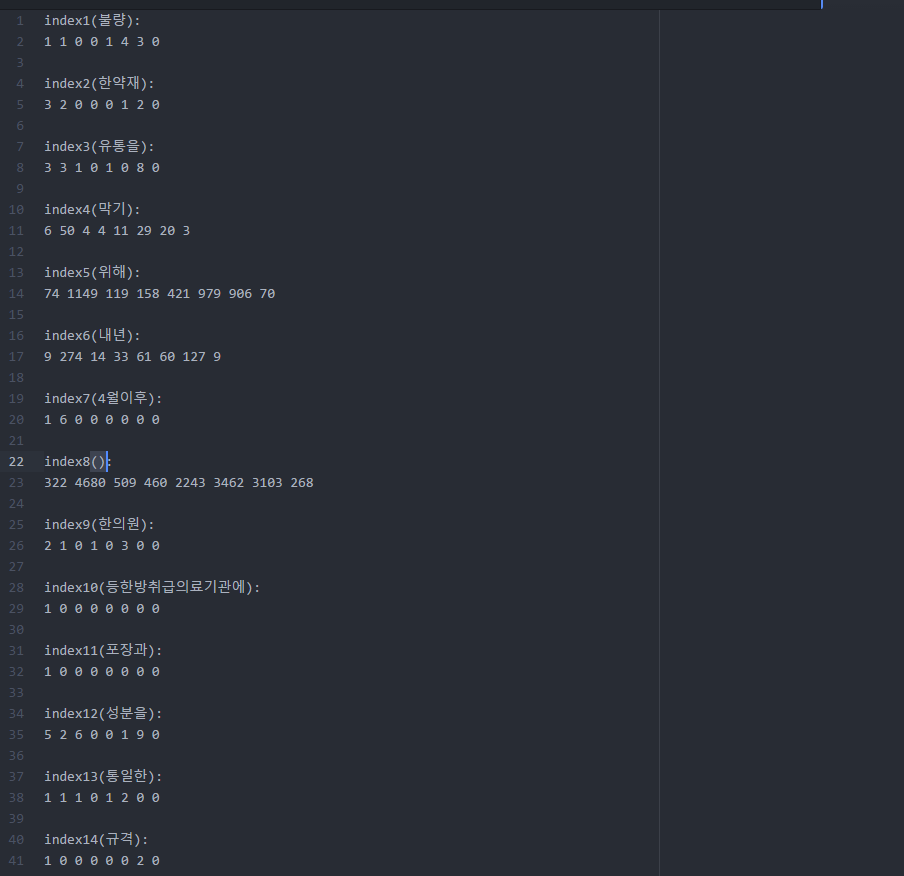
**RAM :** 8GB

**Number of each category**



**Pre - processing**

delete Special characters ('-' , ',' , '/' , ' .' , '<' , '>', ' " ' , ' [' , ']' , ' {', '}') and tags (<a>,<search>) and dividing line to word by token – “ “

****

**Performance**

|  |  |
| --- | --- |
| Language | Java |
| Number of Lines | 540 lines |
| Time to create Training Set | 110 seconds |
| Time to learn Training Set | About 8 minutes |
| Time to classify New Input | About 22 seconds |

**Accuracy**



**Number of False Positive and False Negative**

|  |  |  |
| --- | --- | --- |
|  | **False Negative** | **FalsePositive** |
| Category 1 | 60 | 29 |
| Category 2 | 322 | 428 |
| Category 3 | 98 | 62 |
| Category 4 | 47 | 27 |
| Category 5 | 168 | 188 |
| Category 6 | 169 | 98 |
| Category 7 | 352 | 395 |
| Category 8 | 37 | 26 |

**Conclusion**

My accuracy is about 68 %

I need more elaborate pre-processing like deleting ‘은,는,이가’

**Pseudo code**

Hashmap map(key : String, value : Integer) // Mapping Category string to index

**//creating map Table**

For(the number of documents)

Line += get each line of document

If Line contains "CAT'03"

label= map (string of category)

If Line has ('-' , ',' , '/' , ' .' , '<' , '>', ' " ' , ' [' , ']' , ' {', '}'

Change character to " "

If Line get all String of a document

List features= MapTable.split(token : “ ”)

For(List size)

MapTable.counting(List features[i])

**//get sorted ChiSquare's result**

PriorityQueue<Feature> resultQueue = ChiSquare.getChiResult()

Feature[the number of features] resultFeatures

HashMap resultMap (key : string, value : Integer)

newIndex = 0;

While(resultQueue is not empty)

Feature f = resultQueue.remove()

f.newIndex = newIndex

**resultFeatures[newIndex] = f**

newIndex += 1

ResultMap.put(MapTable.I\_to\_Smap(f.index), newIndex)

Do bubble sort about index

**//change documents to train and test data**

For(the number of documents)

Line += get each line of document

If Line contains "CAT'03"

label= map (string of category)

Continue

If Line has ('-' , ',' , '/' , ' .' , '<' , '>', ' " ' , ' [' , ']' , ' {', '}'

Change character to " "

If Line get all String of a document

String[] split = Line.split(" ")

Change split to Feature object

Sort from feature's index and write on the new file