Problems Found in Submissions for Assignment 1

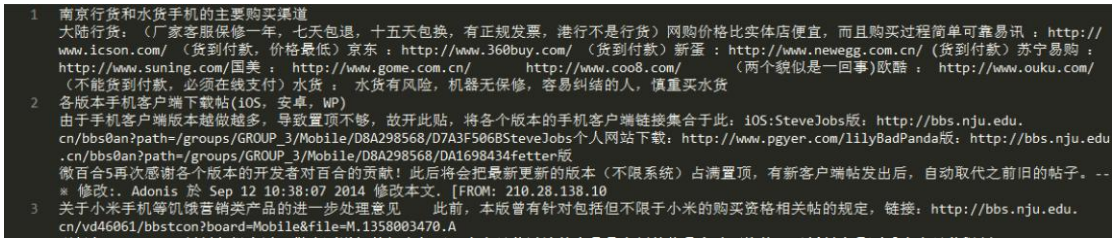
Here are some notes about problems or suggestions on the basis of submissions for Data Mining course Assignment 1. Before reading this note, please read the course and assignment page carefully.

Course web: http://cs.nju.edu.cn/zlj/Courses.html

Assignment web: http://lamda.nju.edu.cn/qianh/dm15.html

Note 1: About opening raw datasets files

Please use Editors like *Sublime* or *Notepad++* to open raw data files. It may be hard to distinguish between lines posts if we open them via *notepad (NOT suggested)*. For example, if we use Sublime or Notepad++ to open them, the raw datasets are shown as below:



Note 2: The TF-IDF output

Instead of computing TF-IDF for each word in each class, we need to compute TF-IDF for each post and combine them as a sparse vector. It must be sparse because most words in the word dictionary may not appear in a post and their TF-IDF values will be zero. Then, we will have a sparse feature vector for each post, which can easily be used in following tasks such as classification.

A Wrong result (computing values for each word separately) is shown below:

In D_Computer.txt

```
1  maven 1.0
2  认识 7.0
3  书写 1.0
4  大处 4.0
5  明确 2.0969100130080567
6  学会 2.0969100130080567
7  寻觅 1.0
8  好了 1.3979400086720377
9  较多 3.0
10 招收 10.0
11 最酷 2.0
12 因病 2.0
13 百思买 1.0
```

A Right result:

	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	.....
Post1	0	0	0.3356	0	0	0	0.3999		
Post2	0.1246	0	0	0	0	0	0.2378		
Post3	0	0.2467	0	0.6835	0.1376	0	0	0	
...									

1. The TF-IDF feature result only need to contain their values, and the *word, post and line number should not be contained*.
2. Please only use *three or four decimals* (like 3.120, 1.1132) to show your results.
3. Directly write sparse vectors into a file may occupy a lot of spaces. We RECOMMEND you to use a sparse representation of the sparse feature vectors:
   <index1>:<value1> <index2>:<value2> ...
   (whether use sparse representation or not will not affect your scores)

Note 3: Dimension for each vector

Dimensionality for each sparse vector must be same. Suppose there are N words in the word dictionary, then the length of TF-IDF feature vector for each post should be N.

Note 4: Please check your TF-IDF results carefully

Please check your results ***carefully*** after preprocessing the data. There are some values like *0.0020900630184019 83YYYYYYYYYYYYYYYY* and *–INFINITY* in the existing uploaded files. All the TF-IDF value *must be* a real number (type of double).

Note 5: About stop words

There may be some messy code in your dictionary or in some posts, you can regard them as stop words and delete them. We have provided a stop word list, but you can use your own one for better results.

Note 6: Please check your file before uploading it

Please check your file carefully before uploading. Some uploaded zip file miss the code or data results. The final score will depend on your code, data results and report.

Note 7: The file format

Please use *zip* file format to compress your file. The same format will help us evaluate your results in a batch style.

Note 8: File and folder name

MG1533001.zip
- code folder(contains all your code or project files)
  -- code file (such as tf_idf.py)
- ReadMe.txt (about how to execute your code)
- report.doc
- result folder (10 txt file for TF-IDF feature vectors)

-- Basketball.txt
-- Mobile.txt
-- ...