ReadMe file: Data Mining Assignment 1

Authors: Dhanvi Athmakuri (athmakuri.1), Harsh Gupta (gupta.749)

Date: 09/20/2016

* The code is compatible with Python 2.7
* **Note : The code reads the files directly from the internet. Please ensure proper internet connection before running the code. If there seems to be a lag in the code execution, it is because of slow internet connection.**
* You can run the code using the command
* **./python preprocessor\_assignment.py** followed by four parameters depending on the kind of files you want to generate and the number of reuters documents you want to run the code on. The script and the relevant output files are present in the following folder :
* ‘**/home/8/athmakuri.1/athmakuri.1\\_DM\\_Lab\\_1**’
* All the relevant output files are in ‘.csv’ format. They can be viewed using the ‘cat’ command or be opened through any other application.
* The files whose names start with the word ‘sample\_’ are generated only for 1000 articles to give a sample of how the output format looks like. The larger files might be difficult to open in any application because of their size. These are the csv files whose names do not start with ‘sample’ . They can be viewed in linux through the following command :
  + **more -n <no\_of\_lines that you want to view> <file\_name>**
* The first three parameters is for
* Writing spare title file
* Writing sparse bag of words file
* Writing TF-IDF matrix

respectively. If you give a value of 1 for each it will write the file and if you plug 0 for any respective file it won’t write that file.

* The fourth parameter is the number of reuters documents (.sgm files) you want the code to run on. If you want only 10 documents (.sgm files) to be preprocessed you can plug in the value 10 in it.

Example:

* **python ./preprocessor\_assignment.py 1 1 1 5**

The above command will execute on the first 5 documents in reuters directory and will output the sparse title file, sparse bag of words file and TF-IDF matrix.

* **python ./preprocessor\_assignment.py 0 0 1 21**

The above code will execute on the first 21 documents in the reuters directory and will output only the TF-IDF matrix. This choice is given to the user, because generating all the files for all the 21 .sgm files it time consuming. The user can collect output for the files that the user is interested in much faster.

**Disclaimer:**

* If you don’t give any parameter with “./**python preprocessor\_assignment.py**” it will execute all the files in the reuters directors and will write all the three files**.**
* You can also have the code generate all the files for all the ‘.sgm’ files, simply by typing **make**.