ReadMe file: Data Mining Assignment 4

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

Date: 11/15/2016

* 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
* **./minhash.py** followed by three parameters depending on the number of shingles you want to generate, the number of reuters documents you want to run the code on and the number of hash functions. The script and the relevant output files are present in the following folder :
* ‘**/home/8/athmakuri.1/athmakuri.1/DMLab3Submission’**
* All the relevant output files are produced in the command line itself.
* The first three parameters is for
* N grams
* Number of files you want the code to run on
* Number of hash functions
* If you don’t give any parameter then the default values will be take
* N gram = 3
* Number of file = 1
* Number of Hash Function = 16

Example:

* **python ./minhash.py**

The above command will execute on the first’.sgm’ file, (1000 articles) in reuters directory with n grams = 3 and number of hash functions = 16 and will output the results on the command line itself.

* **python ./minhash.py 3 1 16**

The above code will execute on ngrams = 3, the first .sgm file (1000 articles) in the reuters directory with hash functions = 16. The code will show, dynamically, the number of files processed, each in a new line, for calculating the jaccard similarity matrix followed by min wise hashing.

**NOTE : The analysis was presented using random sampling of documents, however, on running the code, random sampling is NOT performed over the parsed files.**

**NOTE : To verify the correctness of the code, it is recommended use the above to commands, as they run faster than other arguments. (Expected time : 210 seconds)**

**NOTE :** Please run the code only for **file\_count = 1,2,3,4,10 or 20 (i.e 1000,2000,3000,4000,10000 and 20000) files as the appropriate prime numbers have been computed only for those corresponding shingles.** The results would still be displayed for other files, but with the use of a much larger prime number.