CS598 AITR Assignment 5
MapReduce programming and word association discovery

- Code written and tested on Single Node Hadoop Cluster set up on OSX
- API Version: 1.2.1 (this is different from the altocumulus box which has 0.19)
- Java Files : wordCount.java , pairCount.java , mutualinformation.java
- Data Set : Running on a 4000 line pruned data set of apsrc.txt
- Expected Outputs: wc.dat (wordcount) ,wpc.dat(wordpaircount) and mutualinfo.txt(file containing result)
- **Where can you find the outputs**: (hadoop relative) /home/raguram2/

  ```
  hadoop fs -ls /home/raguram2
  ```

  Where can you find the code: (linux default home path) /home/raguram2/

Steps in running code

- **Step 1**
  Generate Pruned data set using head -4000 apsrc.txt
- **Step 2**
  Generate jar using hadoop libraries

  ```
  javac -classpath /usr/local/Cellar/hadoop/1.2.1/libexec/lib/commons-cli-1.2.jar:
  /usr/local/Cellar/hadoop/1.2.1/libexec/hadoop-core-1.2.1.jar  -d
  miclasses *.java && jar -cvf mi.jar -C miclasses/ .
  ```

- **Step 3**
  Use the wordCount class that uses map reduce and derive counts on occurence in total documents and total number of occurrences.Word Counts go into "wc/" folder .You can find the wc.dat file used in my hadoop relative path /home/raguram2/wc.dat

  ```
  hadoop jar mi.jar wordCount apsrc_4k.txt wc
  ```

- **Step 4**
  Use the pairCount class that uses map reduce and derive pair counts. Pair Counts go into wpc/ folder .You can find the wpc.dat file used in my hadoop relative path /home/raguram2/wpc.dat

```
hadoop jar mi.jar pairCount apsrc_4k.txt wpc
```

- **Step 5**
  Use the mutualinformation class and derive mutual information for a given word pair as per assignment instructions. Mutual Information gets written in mutualinfo.txt

  ```
  hadoop jar mi.jar mutualinformation apsrc_4k.txt wc.dat wpc.dat mutualinfo.txt
  //note that wc.dat and wpc.dat are renamed results inside the wc/ and wpc/ results
  ```

  I have attached the source files as well as sample output files (100 lines) of wc.dat,wpc.dat and mutualinfo.txt