This Program is made to run on Linux (Highly required)

This program runs on Python 2.7 (required)

**Steps:**

1. Run the hadoop
2. Make directory inside haddop dfs  
   Hadoop dfs –mkdir –p /user/(username)/section/
3. Upload the datasets you want to do jobs on.  
   Hadoop dfs –put ~(Dataset directory) /user/(username)/section
4. Run the first job  
   hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-2.8.1.jar

-Dmapred.reduce.tasks=1

-Dstream.num.map.output.key.fields=2

-input /user/trend/section/\*.txt

-output /user/trend/section/output\_run

-mapper ~(first mapper file with directory in your machine)

-reducer ~(first reducer file with directory in your machine)

1. Run the second job  
   hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-2.8.1.jar

-Dmapred.reduce.tasks=1

-Dstream.num.map.output.key.fields=2

-input /user/trend/section/output\_run/part-00000

-output /user/trend/section/output\_final\_answer\_resault

-mapper ~(second mapper file with directory in your machine)

-reducer ~(second reducer file with directory in your machine)

1. The final output file is generated and you can explore it at hadoop dfs at   
   /user/trend/section/ output\_final\_answer\_resault /part-00000
2. To get the output file from hadoop local web application at link  
   http://localhost:50070/webhdfs/v1/user/trend/section/output\_final\_answer\_resault/part-00000?op=OPEN