**ASSIGNMENT 2**

Setup everytime on local machine:

export SPARK\_HOME=<insert\_your\_path\_here>

# when it is on cluster, only use <your\_path>.

Then start spark on local:

${SPARK\_HOME}/bin/pyspark

When on cluster:

pyspark

(before running, remove all outputs in hdfs filesystem):

hdfs dfs -rm -r output\*

do spark-submit out of pyspark console!

When on local:

${SPARK\_HOME}/bin/spark-submit wordcount.py wordcount-1 output-1

When on cluster:

spark-submit wordcount.py file\_path\_on\_cluster/wordcount-1 output-1

(file\_path\_on\_cluster: /courses/732/wordcount-1)

check result:

hdfs dfs -cat output-1/\* |less