**Lab3**

**Worksheet #1: After running the cp command, what file(s) are now in your home directory? There should be at least two: a ".sh" file, and a ".py" file.**

When I type in “ls –l” command, it shows:

total 3

drwxr-xr-x 2 changbai eecs 60 Sep 19 15:13 si618FluxSetup

Files shows as below:

si618FluxSetup ,si618FluxSetup/ngram-job.py and si618FluxSetup/spark-run.sh

Worksheet #2: **What is the name of the last file in the listing for HFS folder /var/si618f17?**

ataset\_review.json

Worksheet #3: **What year was Einstein first mentioned (as a noun) in Google Books data?**

**einstein\_NOUN** 1921 4 4

year:1921

Worksheet #4: **After the Spark job completes, what are three files listed in your Hadoop File System output directory ./ngrams-out?**

Found 3 items

-rw-r-----   3 changbai hadoop          0 2017-09-19 15:38 ngrams-out/\_SUCCESS

-rw-r-----   3 changbai hadoop       5540 2017-09-19 15:38 ngrams-out/part-00000

-rw-r-----   3 changbai hadoop       5492 2017-09-19 15:38 ngrams-out/part-00001

Worksheet #5: **What were the average word lengths observed in books from the years 1520, 1597, and 1598 ?**

(1592, 8.937106918238994)

(1594, 12.142857142857142)

(1520, 10.84)

### **6. Bonus Challenge**

Using the output file that is placed in your HFS output directory, produce a scatterplot of how average word length has changed over the full time period of the dataset (using your favorite spreadsheet or data viz program). Insert it below.