

## Worksheet for lab 5

Haoyou Liu

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.**

spark-run.sh and ngram-job.py

2. **What is the name of the last file in the listing for HFS folder /var/si618w17?**

yelp\_academic\_dataset\_review\_updated.json

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

1921

4. **After the Spark job completes, what are the first three files listed in your Hadoop File System output directory ./output?**

part-00000, part-00001, part-00002

5. **What were the average word lengths observed in books from the years 1563, 1572, and 1575?**

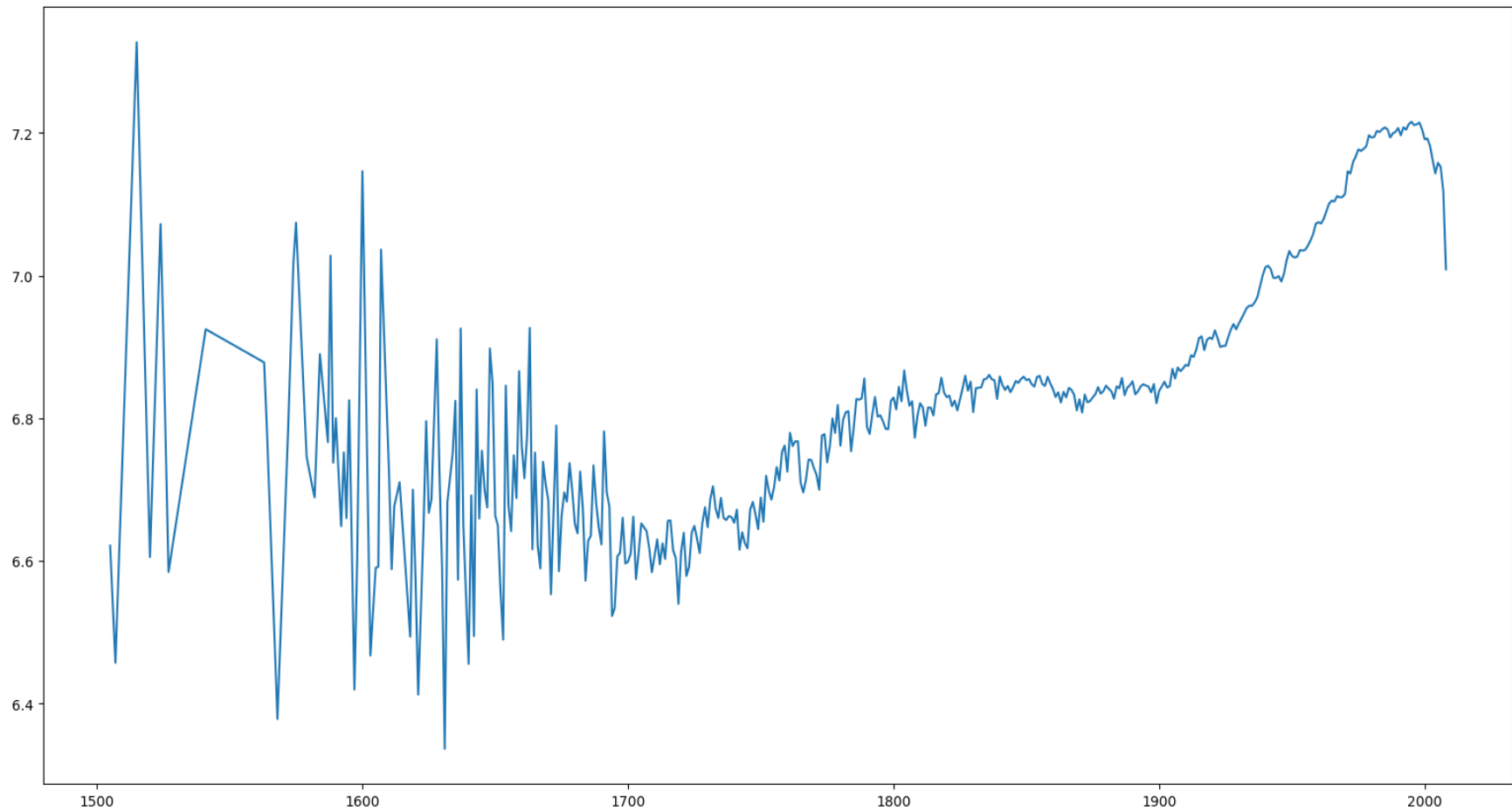
(1563, 6.87824526311463),

(1572, 6.781863482544008),

(1575, 7.074171357961625)

cont'd

6. Bonus Challenge



**Source code:**

```
import matplotlib.pyplot as plt

with open ('ngrams-output.txt', 'r') as f:
    dict = {}
    for line in f:
        line = line.strip().replace('(', '').replace(')',
        '').replace(' ', '').split(',')
        dict[line[0]] = line[1]

    tup = sorted(dict.items(), key= lambda x:x)
    year, word_len = zip(*tup)

    plt.plot(year, word_len)
    plt.show()
```