BIS 420 PROGRAMMING FOR DATA SCIENCE

PRAJAKTA POHARE

CHAPTER 13 EXERCISE 13.9

ILLINOIS STATE UNIVERSITY

The “rank” of a word is its position in a list of words sorted by frequency: the most common word has rank 1, the second most common has rank 2, etc. Zipf’s law describes a relationship between the ranks and frequencies of words in natural languages (http: // en. wikipedia. org/ wiki/ Zipf's_ law). Specifically, it predicts that the frequency, f , of the word with rank r is:

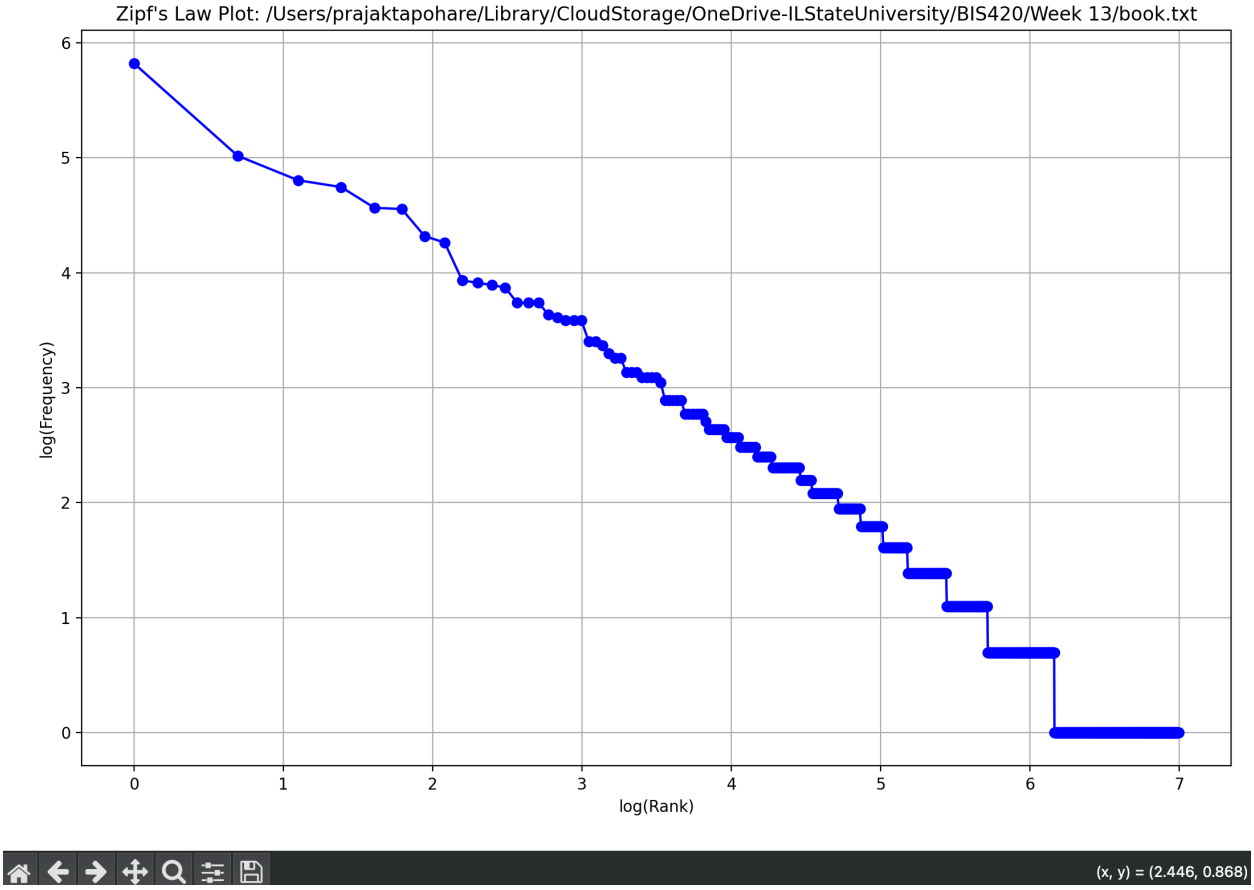
f= cr−s where s and c are parameters that depend on the language and the text. If you take the logarithm of both sides of this equation, you get:

log f= log c−s log r. So if you plot log f versus log r, you should get a straight line with slope−s and intercept log c.

Write a program that reads a text from a file, counts word frequencies, and prints one line for each word, in descending order of frequency, with log f and log r. Use the graphing program of your choice to plot the results and check whether they form a straight line. Can you estimate the value of s?

```
PROBLEMS 1   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

/usr/local/bin/python3 "/Users/prajaktapohare/Library/CloudStorage/OneDrive-ILStateUniversity/BIS420/Week 13/BIS420_PrajaktaPohare_Ch13_13.9.py"
→  ~ /usr/local/bin/python3 "/Users/prajaktapohare/Library/CloudStorage/OneDrive-ILStateUniversity/BIS420/Week 13/BIS420_PrajaktaPohare_Ch13_13.9.py"
Matplotlib is building the font cache; this may take a moment.
Rank  Word            Freq      log(R)    log(F)
--------------------------------------------------
1     the             337       0.0000    5.8201
2     and             151       0.6931    5.0173
3     to              122       1.0986    4.8040
4     of              115       1.3863    4.7449
5     a               96        1.6094    4.5643
6     in              95        1.7918    4.5539
7     exercise        75        1.9459    4.3175
8     is              71        2.0794    4.2627
9     as              51        2.1972    3.9318
10    position        50        2.3026    3.9120
11    with            49        2.3979    3.8918
12    be              48        2.4849    3.8712
13    on              42        2.5649    3.7377
14    it              42        2.6391    3.7377
15    this            42        2.7081    3.7377
16    body            38        2.7726    3.6376
17    action          37        2.8332    3.6109
18    constipation    36        2.8904    3.5835
19    for             36        2.9444    3.5835
20    right           36        2.9957    3.5835
```


Zipf's Law Plot: /Users/prajaktapohare/Library/CloudStorage/OneDrive-ILStateUniversity/BIS420/Week 13/book.txt
log(Frequency)
log(Rank)
0
1
2
3
4
5
6
7
(x, y) = (2.446, 0.868)