Report

Devbrat Anuragi

17078

Before doing any Analysis I have removed all the text before the “CHAPTER I” and after the end of the “CHAPTER XXXV” from the given Corpus.

A) Observed Token-Type ratio = 8.7

B)

a. Class: Words

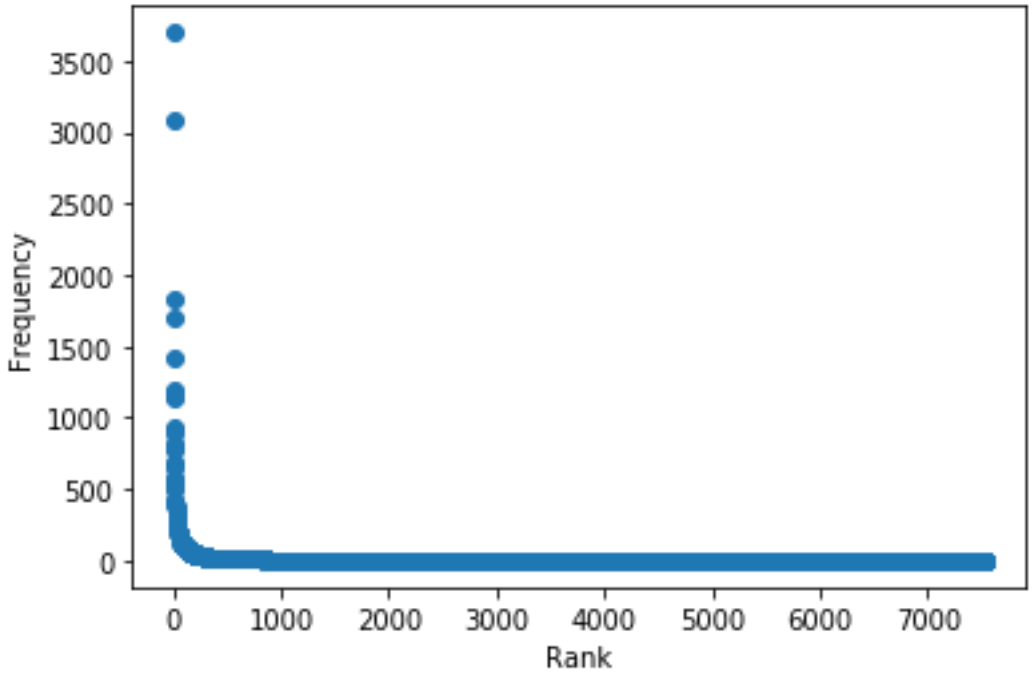
i. calculation of frequency for all word is done in the jupyter notebook file.

ii. Most frequent words are:

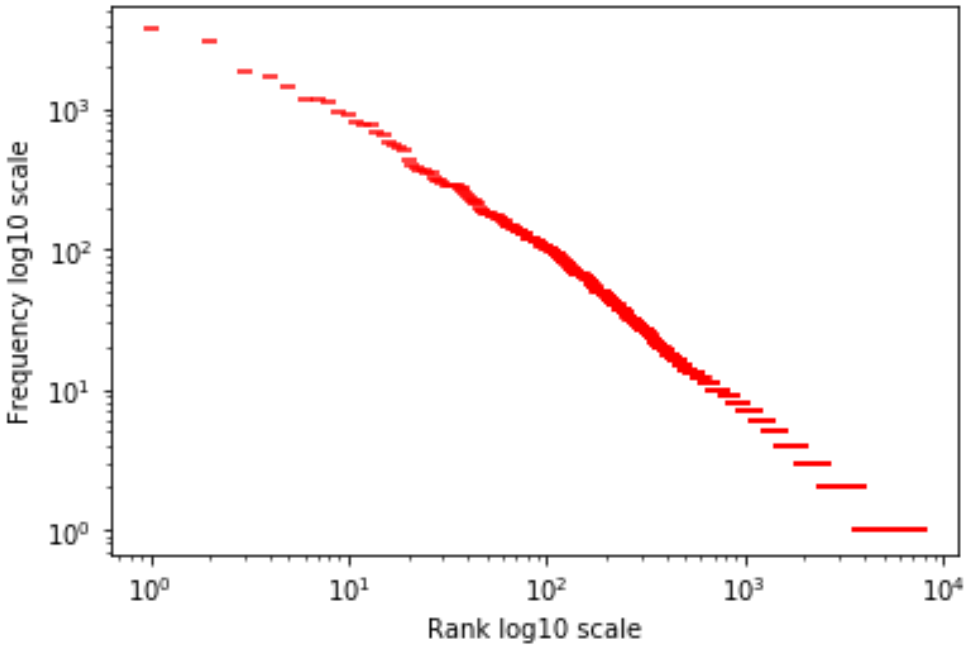
1. the
2. and
3. a
4. to
5. of

words with length ~ 3,2 are most frequent. Most of the frequent word are determinerand preposition.

iii. Plot of Rank vs Frequency



iv. Plot log10(Rank) vs log10(Frequency)



v. Pearsons coefficient of correletion between rank and frequency is -0.17. Negative correlations imply that as **frequency increases**, **rank decreases**, which matches with our observations and also with zipf's law

b. Class: Letters

i. calculation of frequency for all word is done in the jupyter notebook file.

ii. Most Frequent Letters

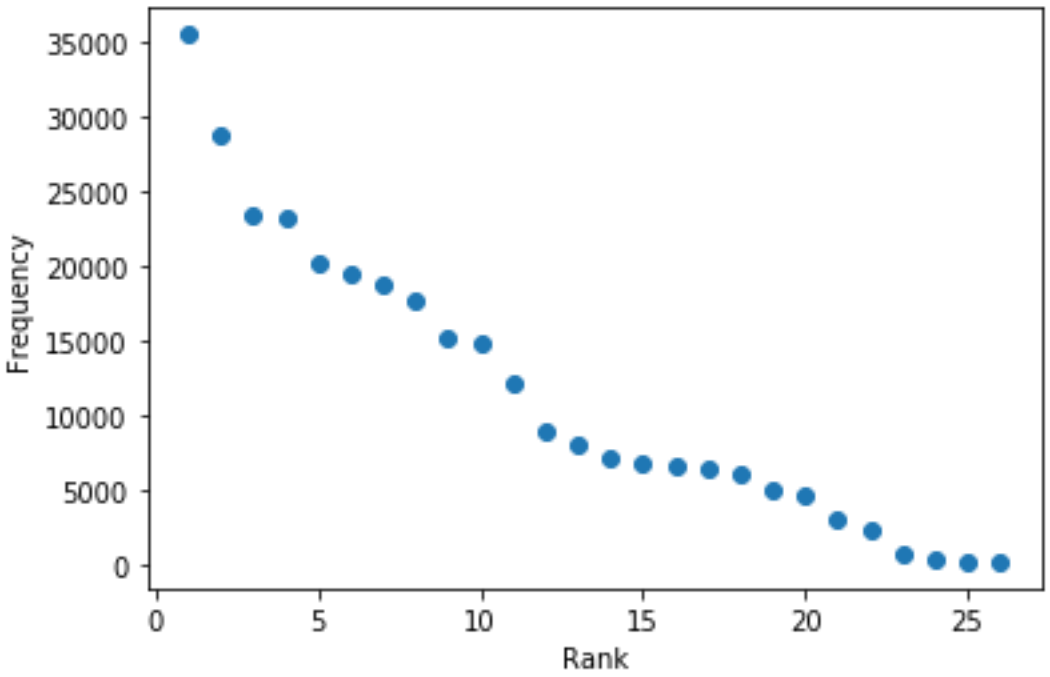
1. e
2. t
3. a
4. o
5. n

    a. **3 out of 5** most frequent letter are vowels.

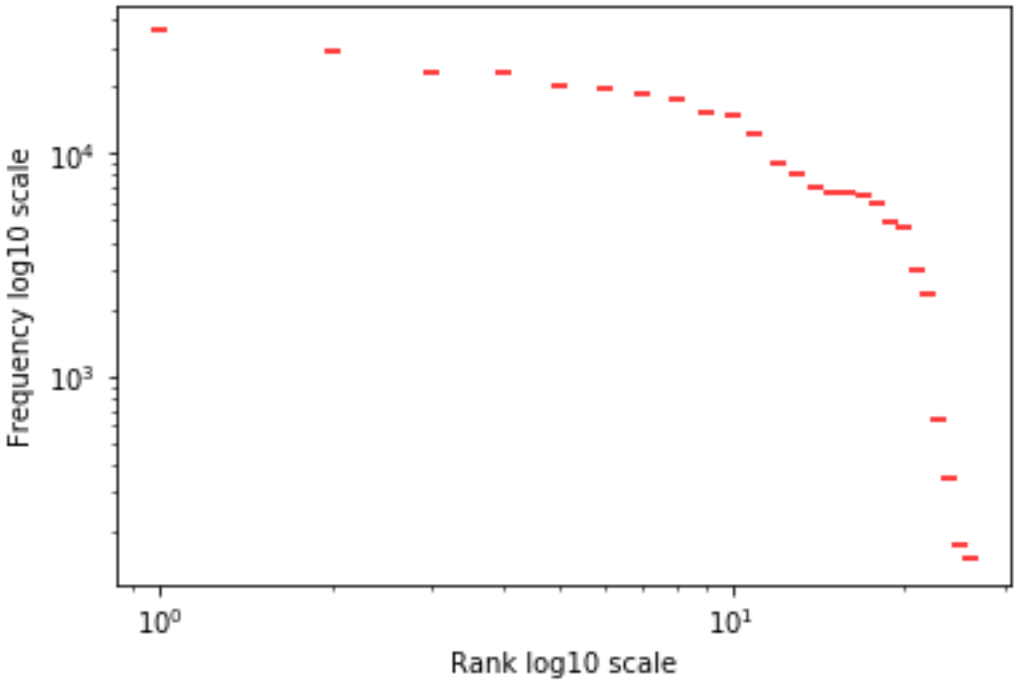
    b. "e" and "t" are the highest because word "the" was also highest and it contains both e and t.

    c. "and" was also among the frequent word that'why "a" and "n" is also among the most frequent letter

iii. Plot of Rank vs Frequency



iv. Plot of log10(Rank) vs log10(Frequency)



v. Pearson's coefficient of correlation between rank and frequency is **-0.95.** This indicates as rank increases the frequency decreases sharply

C) Nearly 37% of the text comprises of Vowels.

Distribution of the letter is different from the distribution of the words. The distributution of letter buldges out more than the distribution of the Lettes. They reasons for this may be there are around ~ 70000 word but there are only 26 alphabets.

"e" and "t" have the highest frequency because word "the" also had highest frequency and it contains both e and t.

"and" was also among the frequent word that's why "a" and "n" is also among the most frequent letter.

zipf's law is not an exact fit for the frequency and rank relation. According to the Zipf's law a straight line is the best fit for log(frequency) and log(rank), as per the observation from the above, a parabola can be a good fit.

Note:

In the code, I have done some extra analysis for the Part A of the assignment. Also, for the Part C I have tried to verify the following example:

For example, this says that the 50^{th} most common word should occur with three times the frequency of the 150^{th} most common word. This