DSE308 Computational Linguistic: Zipf's Law Assignment Report

DEEP POOJA
17074

1 Type-Token Ratio(TTR)

TTR = $\frac{no.of word-type}{no.of tokens} * 100$

TTR = $\frac{8965}{69690} * 100 = 12.86$

type-token ratio is slightly higher than textbook, the textbook has tokens = $71,370$ and word-type = $8,018$, but I got slightly higher value of word-type and lower value of tokens, this is because words like 'couldn't', 'didn't', 'ain't', 'weren't', 'can't', 'don't' are counted as single word which could have been counted as two separate words, which would have result in more higher value of tokens, also these mentioned words have been counted uniquely as 'could not' so we have three counts i.e 'could', 'not' and 'couldn't' in spite of two counts 'could' and 'not', due to this the value of word-type has increased a bit (compared from textbook).

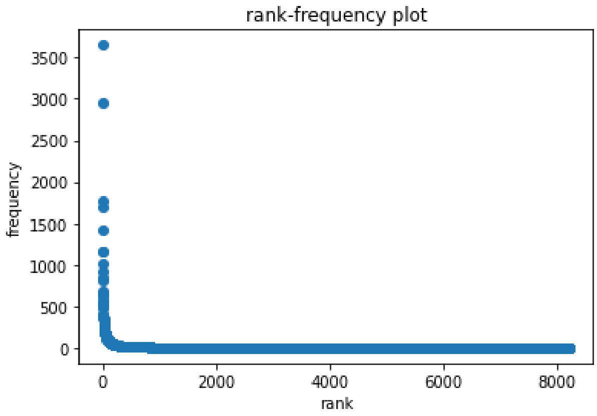
2 Analysis at word level

The most frequent words are listed below with their respective frequency:

Word	frequency
'the'	3651
'and'	2954
'a'	1781
'to'	1693
'of'	1419

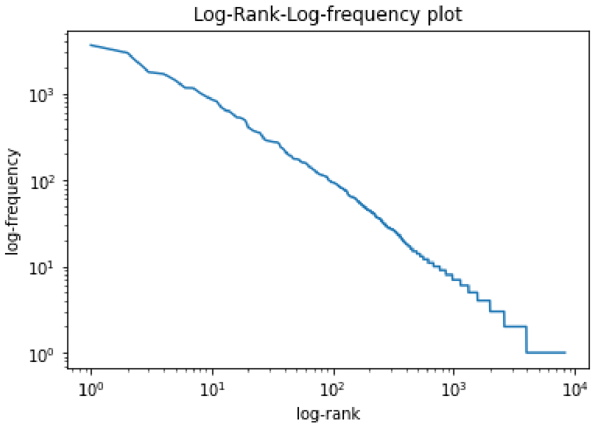
clearly, the most frequent words are function words that is articles, prepositions and conjunctions, Also these words are short in length.

2.1 rank-frequency plot



The rank-frequency plot for words in Tom Sawyer text looks like a hyperbolic in nature that is linear in small proximity close to origin and becomes asymptotic to both axes (rank as well as frequency) for much higher value.

2.2 log-log plot



The log-log plot of rank-frequency for words turns out to be same as Zipf’s law.Besides some variation from linearity in lower right corner as well as in upper left corner, these slight variation from linearity is also seen in Zipf’s study.

2.3 Pearson’s coefficient of correlation

Pearson’s coefficient of correlation for words is -0.164.

This negative correlation (inverse relation) signifies that as the word becomes more frequent(high frequency), the rank of the word decreases(low rank) (and vice versa). This inverse relation has been already pointed out by Zipf as Zipf’s law.
The closer the value is to -1, the stronger the linear correlation

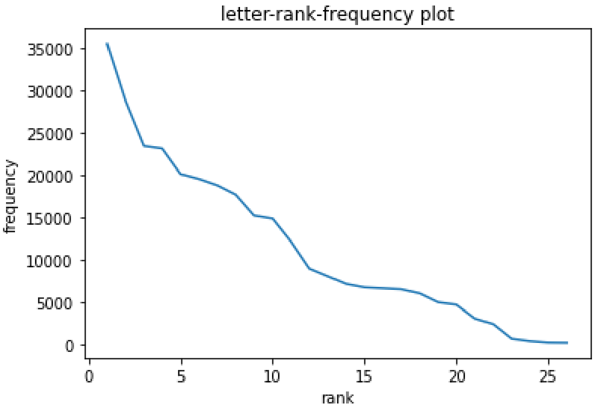
3 Analysis at character level

Most frequent letters are listed below with their respective frequency:

letter:frequency
’e’:35535
’t’:28760
’a’:23442
’o’:23141
’n’:20114

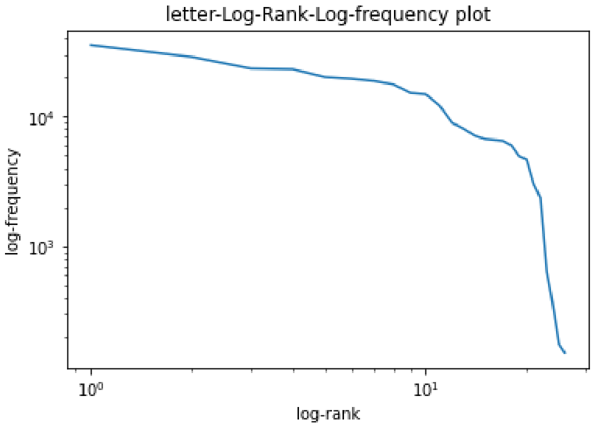
’e’ is the most frequent letter followed by ’t’ and ’a’, since ’the’ is most frequent word followed by ’and’ and ’a’. So high frequency of consonant ’t’ and ’n’ is due to the high frequency of word ’the’ and ’and’ respectively in the text. ’to’ and ’of’ are also among the high frequency words which leads to high frequency of vowel ’o’.

3.1 Rank-Frequency plot



This is interesting since rank-frequency plot for letters is also hyperbolic in nature but with smaller asymptotic tail mainly because of small or constant number of letters in English i.e 26 only.

3.2 log-log plot



This is unexpected as I expected linear behaviour just like for words, It's a non-linear behaviour. It seems somewhat linear when rank is low and it decreases steeply for high rank letters.

3.3 Pearson's coefficient of correlation

Pearson's coefficient of correlation for letters is -0.957.

Note: When coefficient is -1, the relationship is said to be perfectly negatively correlated. In short, if one variable increases, the other variable decreases with the same magnitude (and vice versa).

In our case Pearson's coefficient of correlation for letters pretty close to -1, which can be inferred as very strong inverse relation between frequency of letter and it's rank.

4 Zipf's law

Zifp's law states that the product of frequency of occurrence of words in text and it's rank(most frequent word has least rank) is constant.
Let's denote the frequency of word by 'f' and it's rank by 'r' then from Zipf's law we have:

$$f.r = Constant$$

I have computed the frequencies of each word in Tom Sawyer text by splitting the text by white spaces and it's corresponding rank, it turns out that when I plot frequency and rank on log axes, it follows Zipf's law[see fig 3.2].

I did similar analysis for letters in the text i.e calculating frequency and rank of each letter (26 letters in English alphabets) ad plotting a log-log plot, the results are contrasting as the plot is not linear or it's linear only in small proximity (i.e for low rank and high frequency) and frequency decreases steeply for high rank letters .

I have also plotted frequency and rank on Cartesian axes for both words and letters, it seems plots are quite similar both hyperbolic in nature.

Both words and letters has negative correlation coefficient,but letters has very strong correlation close to -1.
A very interesting observation is letters in English language is not uniformly distributed that is English language has it's own imprint...

5 References

`https://piazza.com/class_profile/get_resource/kdzoqj4hbff3zn/ke838xg2d293p8?`