

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

INFORMATION RETRIEVAL WITH APPLICATIONS OF NLP ENCS4130

**Assignment #2**

Prepared by: Mazen Batrawi - 1190102

Instructor: Dr. Adnan Yahya

Section: 1

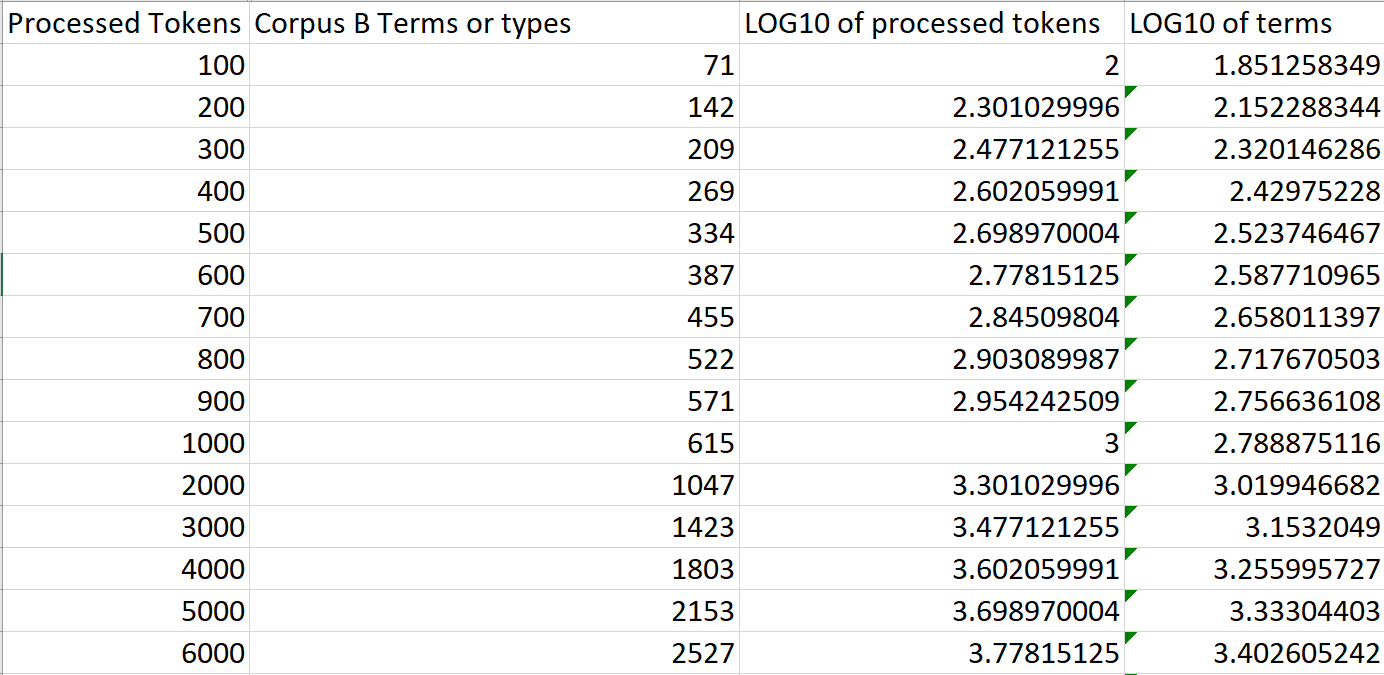
BIRZEIT

May – 2023

1. We have an equation called Heap's Law, which is expressed as M = KTα. In this equation, M represents the number of unique terms or words in a given text, K is a constant that we need to find, T represents the total number of tokens or words in the text, and α is another constant that we also need to determine. To find the values of K and α, we need to follow a few steps. First, we need to gather some data points with sufficiently spaced values. We'll count the number of terms (M) and calculate the number of tokens (T) for each data point.

We take the log for the equation: Log(M) = Log(K) + α log(T).

We can see that the slope of this equation is α and K is 10c, where c comes from the following linear equation: y = ax + c.



The plotted graph resulted as the following:

The equation is: y = 0.6137x + 1.2048  
R² = 0.995

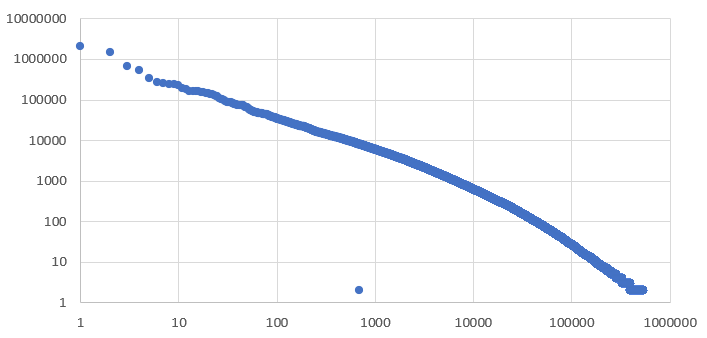
K = 101.2048= 16.025

α = slope = 0.6137

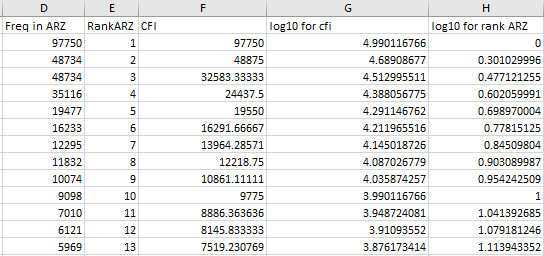
1. Zipf's Law is a commonly used model to describe the distribution of terms in a collection of text. According to Zipf's Law, if we rank the terms in the collection based on their frequency, the frequency of the ith most common term (cfi) is inversely proportional to its rank (i), following the formula cfi = c \* (ik), where c is a constant and k is equal to -1.

To analyze Zipf's Law, we will plot a log-log graph using the data from the selected sheets or corpora. In this graph, we'll plot the log of the term frequency (log cfi) on the y-axis and the log of the term rank (log i) on the x-axis. By taking the logarithm of both variables, we can transform the data and observe any patterns more easily.

Using the formula log cfi = log c + k log i, where k is -1, we can determine the values for log c and plot the points accordingly. Each point on the graph represents a term, with its corresponding log cfi and log i values.

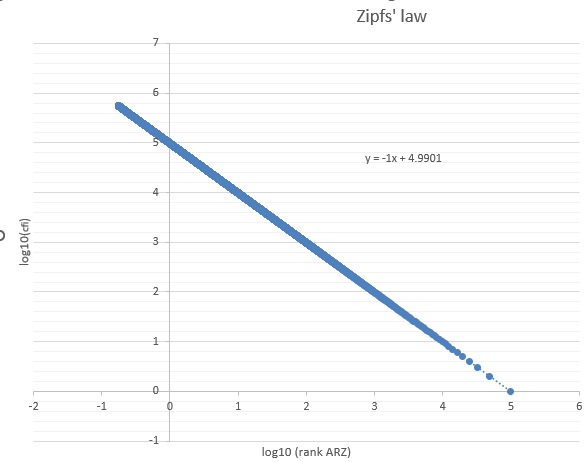


This is plotting the frequency Vs the rank without any simplifications.



We can see that the value of c = 97750.

The plotted graph:



Y = -x + 4.9901

C from this equation is 104.9901 = 97746.2264, which is close to the calculated value (97750).

Cfi = 97750 \* rankARZ-1.

Zipf's law demonstrates a higher level of accuracy for ranks that fall within the middle range, while its accuracy tends to decrease for ranks at the beginning and end. Therefore, although Zipf's law remains reasonably accurate overall, its precision diminishes significantly for ranks that are either very high or very low.