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

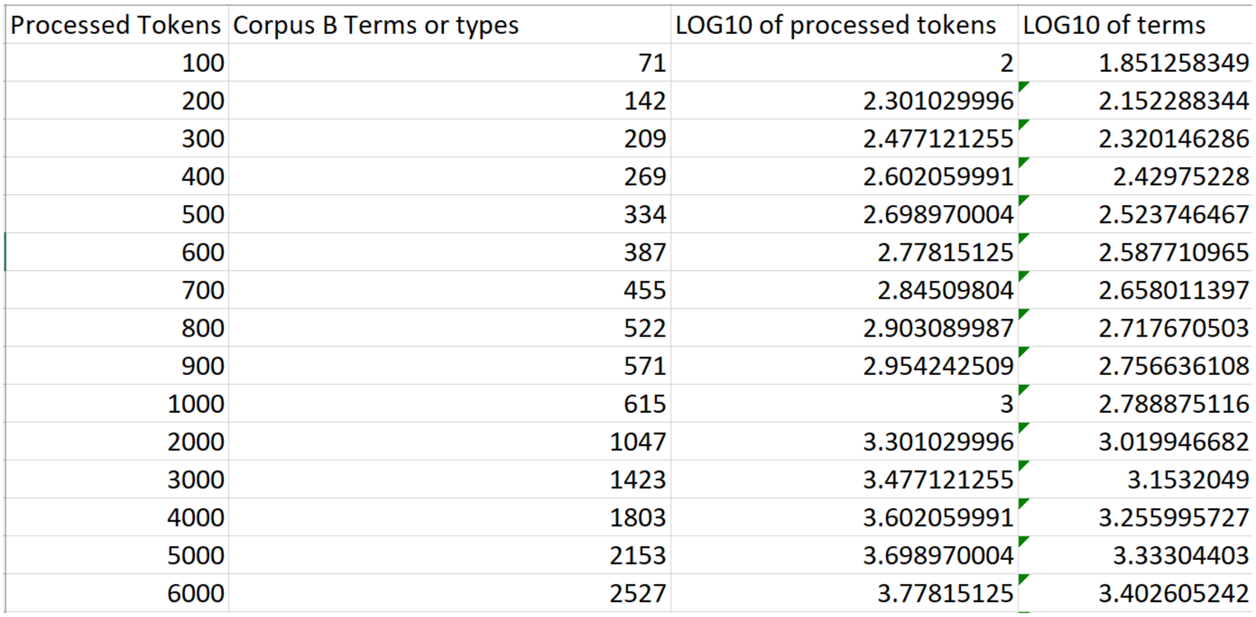
Section: 1

BIRZEIT

May – 2023

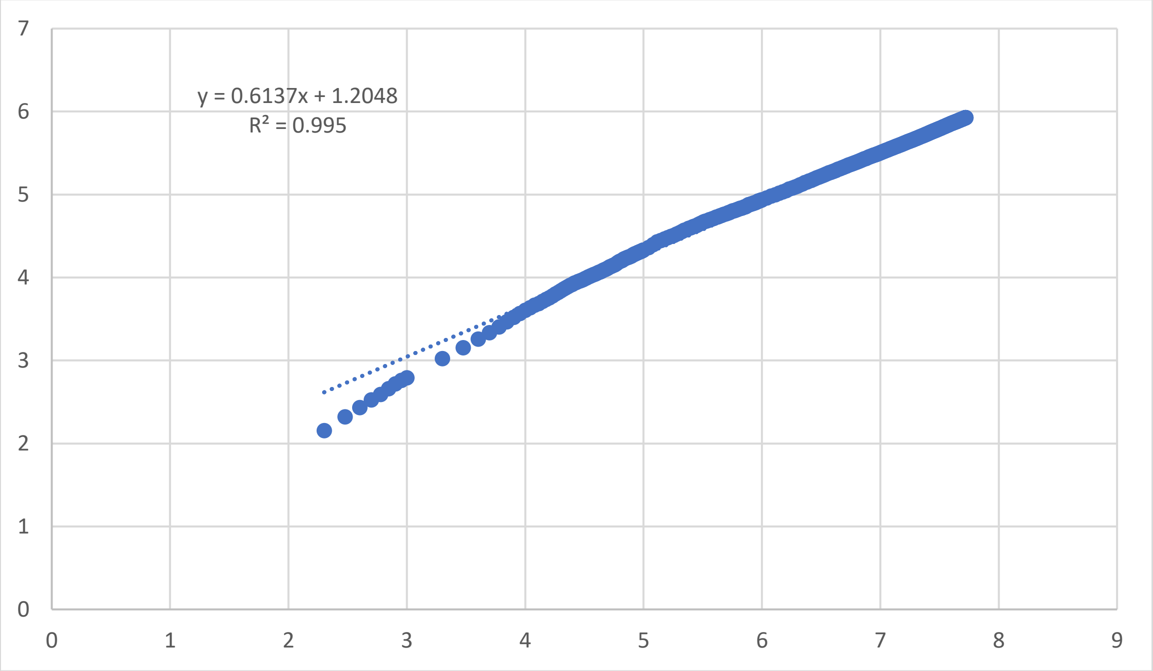
1. We have an equation called Heap's Law, which is expressed as $M = KT^{\alpha}$. In this equation, M represents the number of unique terms or words in a given text, K is a constant that we need to find, T represents the total number of tokens or words in the text, and α is another constant that we 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: $\mathrm{Log}(M) = \mathrm{Log}(K) + \alpha \log(T)$.
   We can see that the slope of this equation is α and K is 10^{c}, where c comes from the following linear equation: $y = ax + c$.



Processed Tokens	Corpus B Terms or types	LOG10 of processed tokens	LOG10 of terms
100	71	2	1.851258349
200	142	2.301029996	2.152288344
300	209	2.477121255	2.320146286
400	269	2.602059991	2.42975228
500	334	2.698970004	2.523746467
600	387	2.77815125	2.587710965
700	455	2.84509804	2.658011397
800	522	2.903089987	2.717670503
900	571	2.954242509	2.756636108
1000	615	3	2.788875116
2000	1047	3.301029996	3.019946682
3000	1423	3.477121255	3.1532049
4000	1803	3.602059991	3.255995727
5000	2153	3.698970004	3.33304403
6000	2527	3.77815125	3.402605242

The plotted graph resulted as the following:



The equation is: $y = 0.6137x + 1.2048$

$R^2 = 0.995$

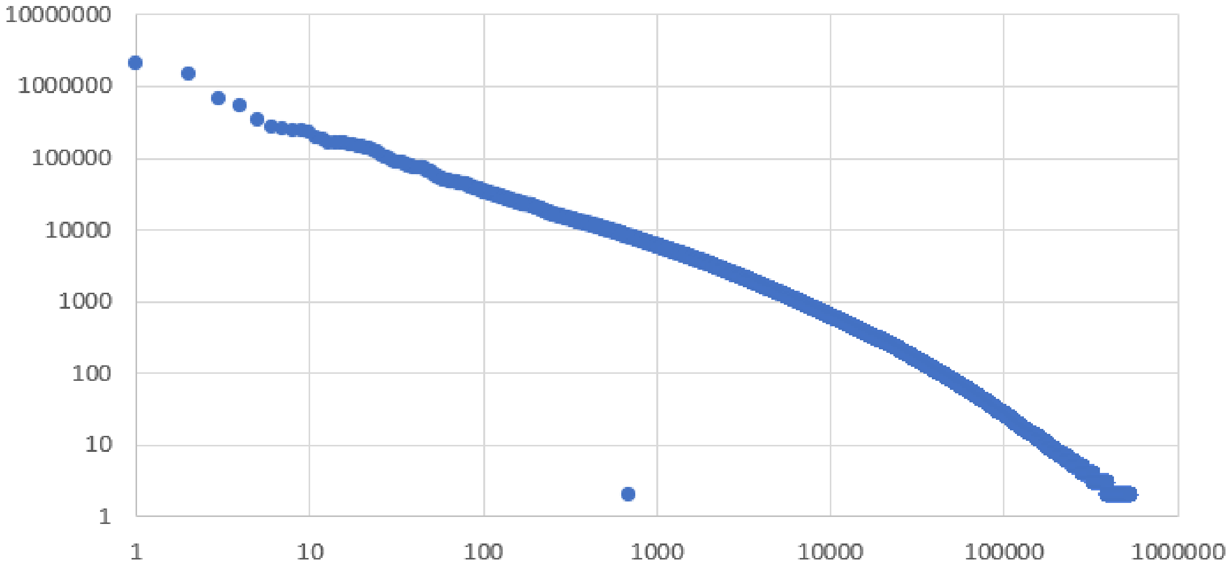
$K = 10^{1.2048} = 16.025$

α = slope = 0.6137

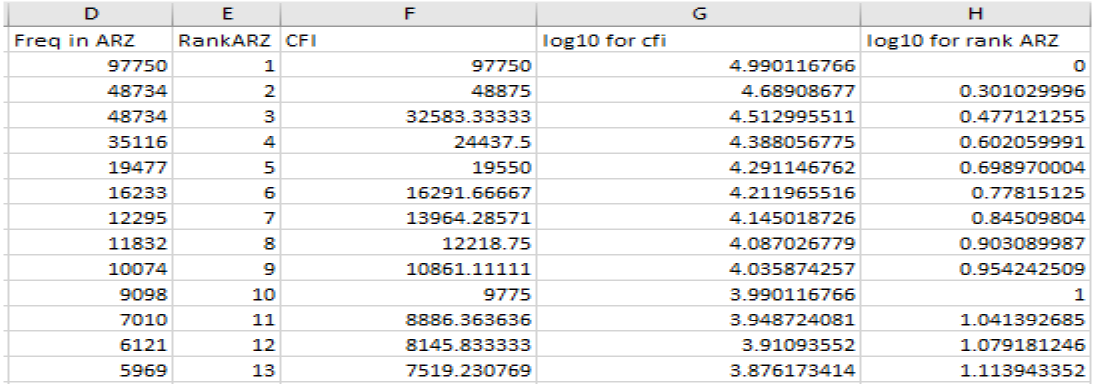
2. Zipf's Law is a commonly used model to describe the distribution of terms in a collection of text. According to Zipf's Law, if we rank the terms in the collection based on their frequency, the frequency of the i_{th} most common term (cfi) is inversely proportional to its rank (i), following the formula cfi = c * (i^k), 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.



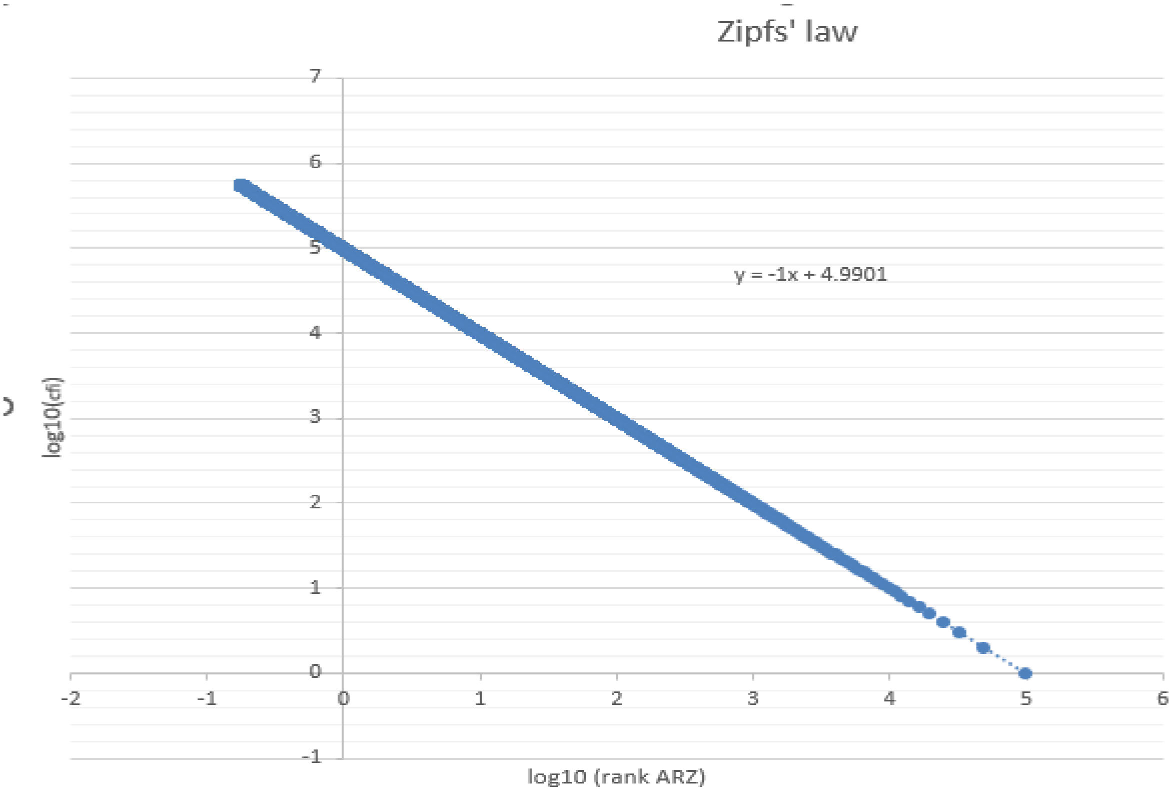
   This is plotting the frequency Vs the rank without any simplifications.



D	E	F	G	H
Freq in ARZ	RankARZ	CFI	log10 for cfi	log10 for rank ARZ
97750	1	97750	4.990116766	0
48734	2	48875	4.68908677	0.301029996
48734	3	32583.33333	4.512995511	0.477121255
35116	4	24437.5	4.388056775	0.602059991
19477	5	19550	4.291146762	0.698970004
16233	6	16291.66667	4.211965516	0.77815125
12295	7	13964.28571	4.145018726	0.84509804
11832	8	12218.75	4.087026779	0.903089987
10074	9	10861.11111	4.035874257	0.954242509
9098	10	9775	3.990116766	1
7010	11	8886.363636	3.948724081	1.041392685
6121	12	8145.833333	3.91093552	1.079181246
5969	13	7519.230769	3.876173414	1.113943352

We can see that the value of c = 97750.

The plotted graph:



Y = -x + 4.9901
C from this equation is $10^{4.9901}$ = 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.