Q2. A)

The text does satisfy zip’s law. As per Zipf’s law Rank \* Probability = Constant.

We see from the data that we have collected in **all\_data.txt** that after 5th rank, the value of probability \* rank is more or less 0.1.

Zipf’s law (rank vs. frequency)

Zipf’s law in log-log space

Q2. B)

As per zip’s law, the formula for number of words that occur n times = 1/n (n+1)

We use this formula to find out words that occur 1,2,3 and 4 times.

1/1(1+1) + 1/2(2+1) + 1/3(3+1) + 1(4(4+1) = 1/2 + 1/6 + 1/12 + 1/20 = **4/5(80%)**

This is the fraction of the total unique words that occur less than 5 times. Total number of unique words = 4/5 \* 2632 = 2106.

The actually number of words that have count less than 5 are 1935

Proportion of words that have count less than 5 are **1935/2632(73.51%)**

Q3. A)

As per Heap’s law,

V = K Nβ

Here,

V = no of unique words

N = total no of unique words

K = constant between 10 and 100

β = constant between 0.4 and 0.6

Now,

V1 = K N1 β … equation (1)

V2 = K N2 β … equation (2)

Dividing equation 1 by equation 2

V1/V2 = (N1/ N2) β … equation (3)

0.9 = (N1/ N2) 0.5 … since V1/V2 = 0.9 and β = 0.5

Squaring both sides,

**0.81 = (N1/ N2)**

Thus, **81%** of the text must be read, before 90% of its vocabulary has been encountered.

Q3. B)

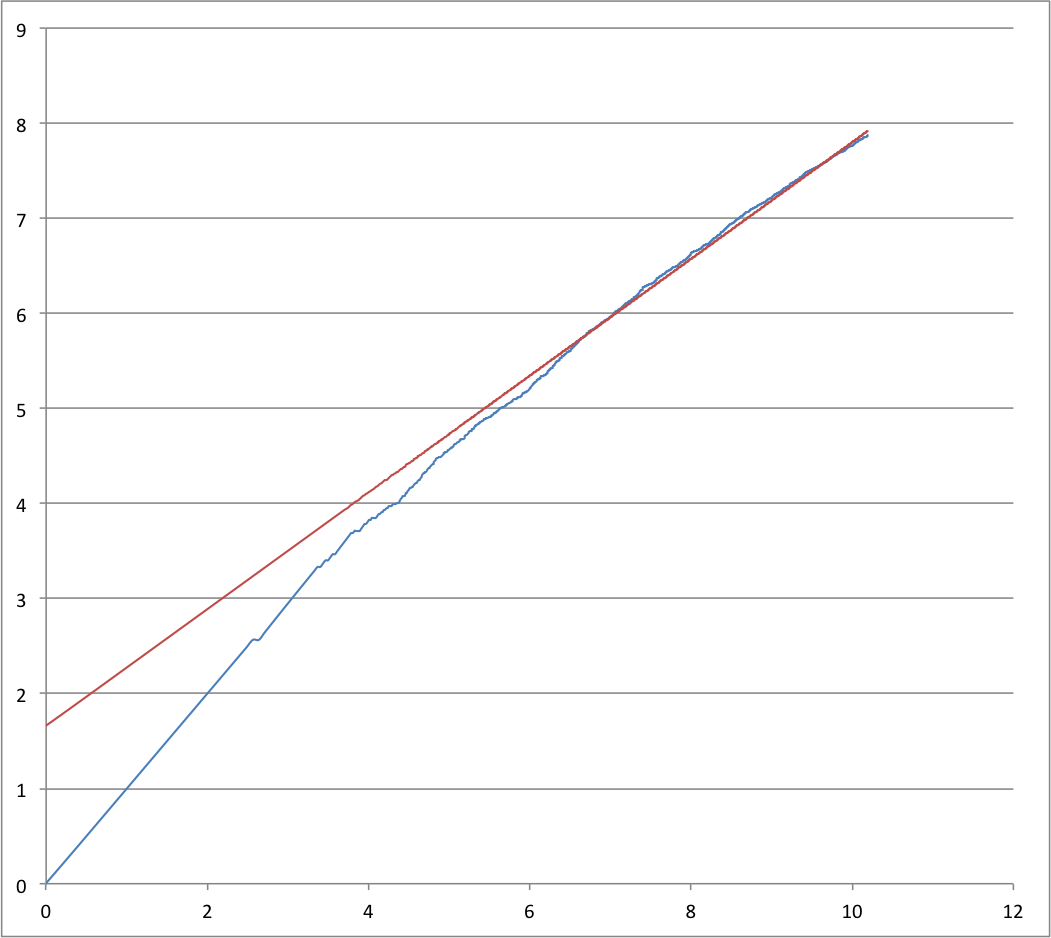
As per Heap’s Law:

V = K N β

Therefore,

log (V) = log (K) + β log (N)

After running heap.py, we get the values of **k as 1.66** and **β as 0.614**



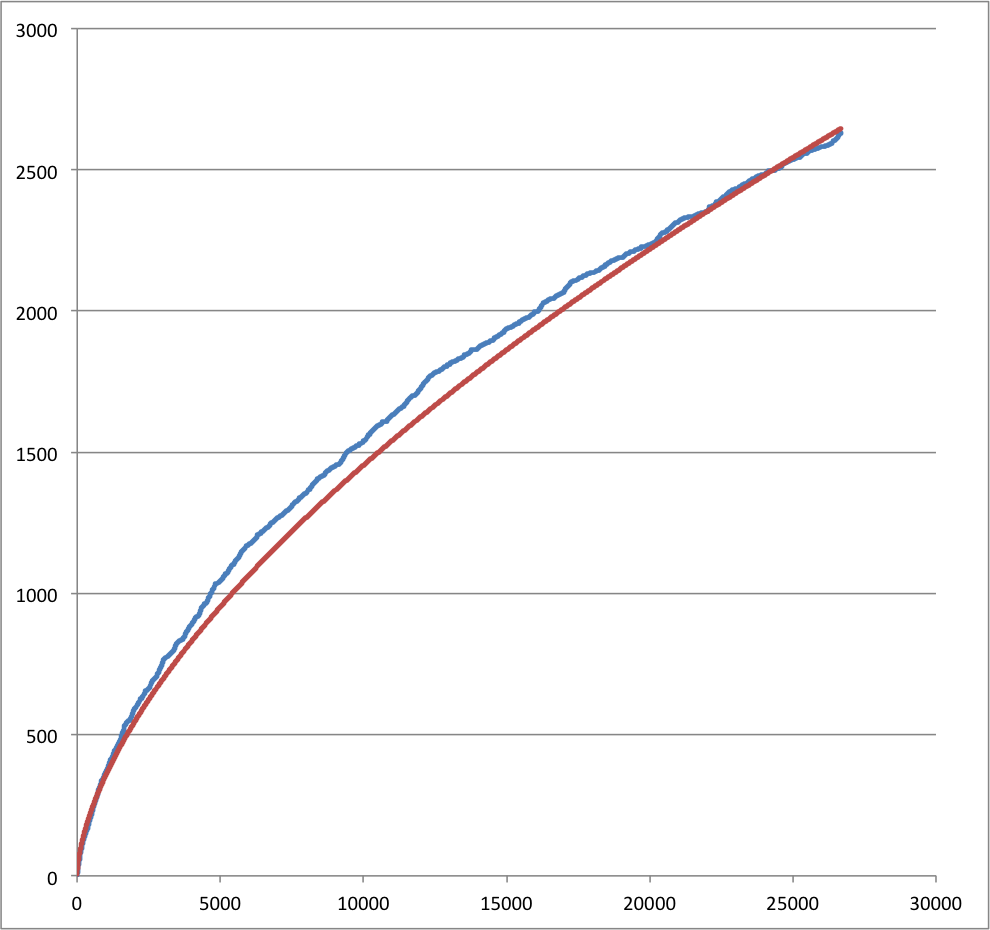
Heap’s law in log-log space

Blue line 🡪 heap’s law (log (words processed) vs log (unique words))

Red line 🡪 Line with best fit for log-log space

Converting back to our normal space,

**K= e1.66 = 5.280**



Heaps law (words processed v/s unique words)

Blue line – heap law (words processed v/s unique words)

Red line – best fit line

**Parameters for best fit, k = 5.280, β = 0.614**

Q4. A)

Search Engine: ixquick

Queries

* Restaurants near me
* Flights to Atlanta

Query 1

Restaurants near me

Search Results found 🡪 40,199,926