IR Assignment -2

GroupMembers:

Rashmi HTI(CS16MTECH11013)

Sherin Thomas (CS16MTECH11016)

Vinod Chhapariya (CS16MTECH11019)

SECTION - I

Inverted Indices created on 1415 files crawled in first assignment. For each index (I1 to I4) following parameters are computed.

Index	Number of Terms	Maximum Length of Posting List	Minimum Length of Posting List	Average Length of Posting List	Size of the file that stores the inverted index
I1	31761	1401	1	11	5.24 MB
I2	31352	1401	1	10	4.18 MB
I3	23731	1401	1	9	3.22 MB
I4	1218	1401	29	144	2.24 MB

SECTION - II

Details about K (K=20) words, their posting list size and average gap between the documents is calculated for each index (I1 to I4). Three text files (Most_Frequent_K_Words.txt, Median_K_Words.txt, Least_Frequent_K_Words.txt) are generated for following K Words.

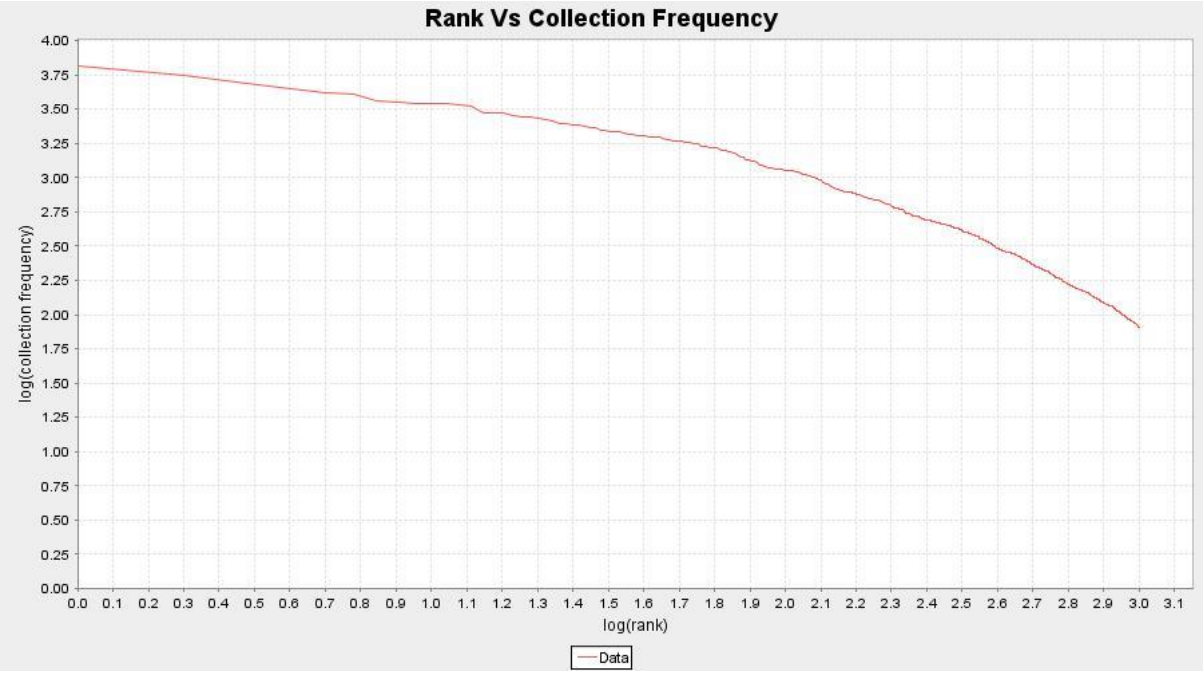
1. **Most Frequent K Words**
2. **Median K Words**
3. **Least Frequent K Words**

SECTION - III

Graph is plotted with top 1000 terms with higher collection frequencies.

X-axis contains log of rank of term in non-increasing collection frequency ordering.

Y-axis contains log of collection frequency of the term.



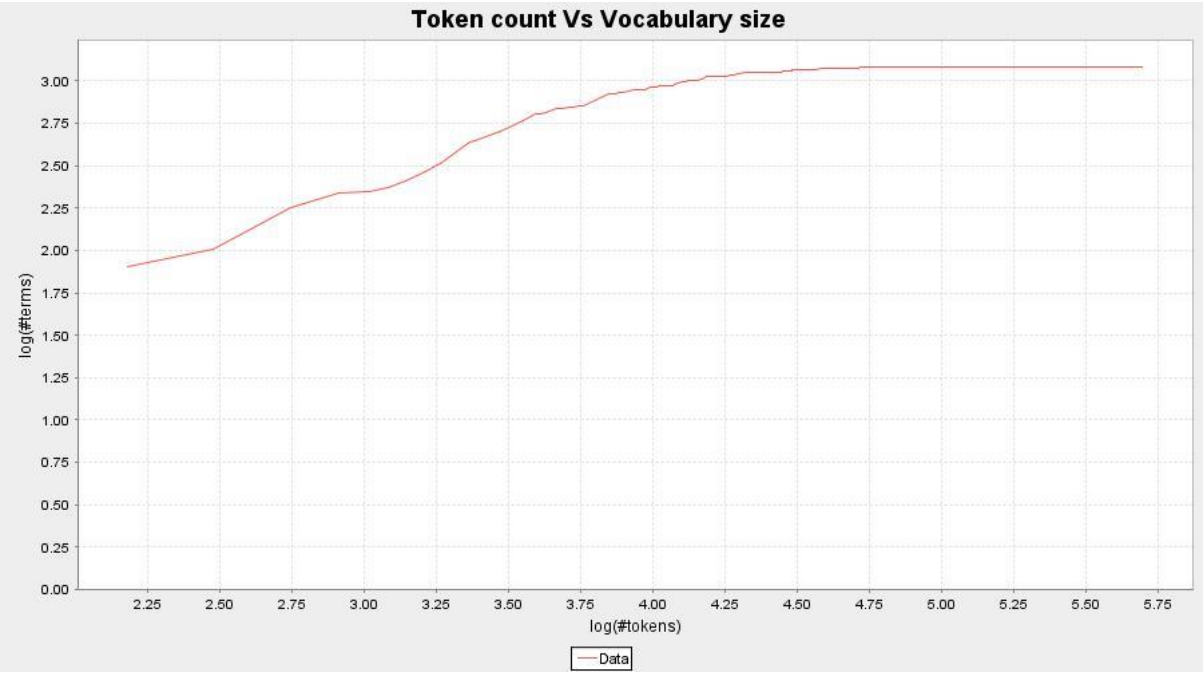
Observation–From the graph we can say that as rank increases collection frequency of the term decreases.

SECTION - IV

Graph is plotted by considering number of terms present and number of tokens already seen in each file

X-axis contains the log of the number of tokens already seen.

Y-axis contains the vocabulary size



Observation- From the above graph we can say that, the number of tokens continues to increases as we process more number of documents in the collection. Whereas the number of terms initially increases and stops at particular point, which shows the terms are already covered in a subset of documents in the collection.