Assignment 2

Mohammed Nauman Sididque

October 15, 2017

Contents

Chapter 1

Problem 4.1

1.1 Problem Statement

Plot rank-frequency curves (using a log-log graph) for words and bigrams in the Wikipedia collection available through the book website (http://www.searchengines-book.com). Plot a curve for the combination of the two. What are the best values for the parameter c for each curve?

1.1.1 Rank-Frequency Curve

Unigrams in the Wikipedia Small Corpus: 225,744

Bigrams in the Wikipedia Small Corpus: 1,429,162

The reason for the graphs to not converge to the Zipf's Law is due to confined corpus size. The value of constant as 1.2 does not allow the curve for unigrams and bigrams to follow the Zipf's Law trajectory. But in case of constant 0.8, the bigrams follow the Zipf's Law tajectory in the middle of their curve.

The graph used to plot the graph is :

$$logcf_i = logc + klogi$$

where k=-1, c is a constant and f is frequency of the word with rank i

The graph with unigram and bigram distribution for constant value of 1.2 is shown in figure. 1.

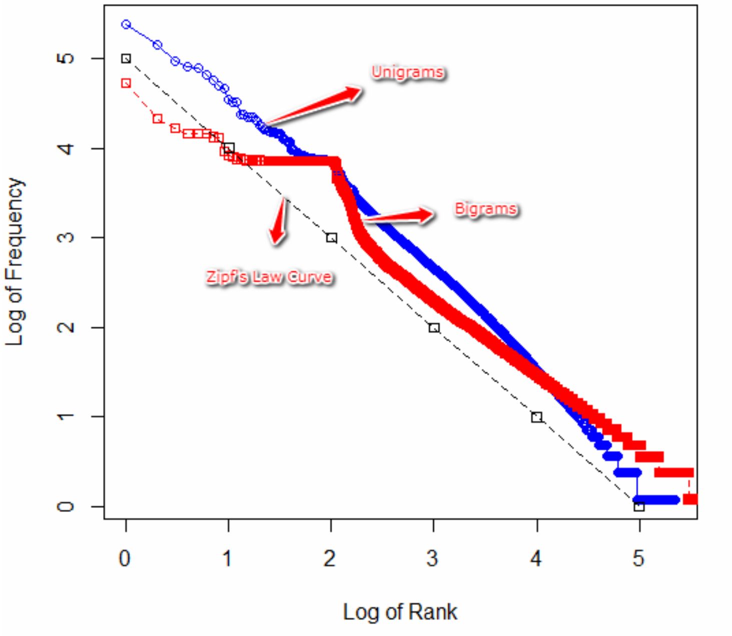


Figure 1.1: rank-frequency curve (c= 1.2)

The graph with unigram and bigram distribution for constant value of 0.8 is shown in figure. 2

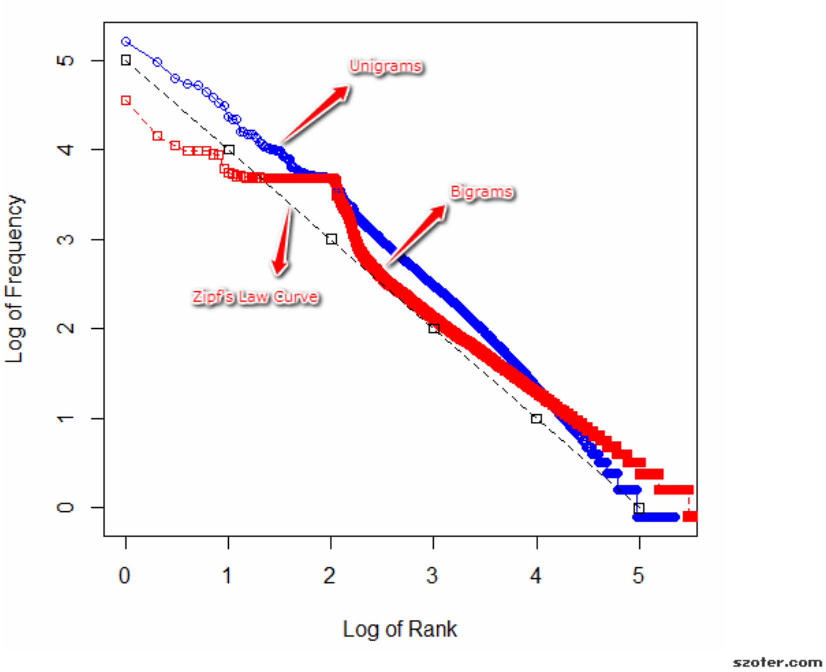


Figure 1.2: rank-frequency curve (c= 0.8)

Chapter 2

Problem 4.2

2.1 Problem Statement

Plot vocabulary growth for the Wikipedia collection and estimate the parameters for Heaps law. Should the order in which the documents are processed make any difference?

2.1.1 Heap's Law Curve

Monograms in the Wikipedia Small Corpus: 225,744 Total Vocabulary in the Wikipedia Small Corpus: 3,902,115

Chapter 3

Problem 4.3

3.1 Problem Statement

Try to estimate the number of web pages indexed by two different search engines using the technique described in this chapter. Compare the size estimates from a range of queries and discuss the consistency (or lack of it) of these estimates.

3.2 Independent Query

3.2.1 Bing Results

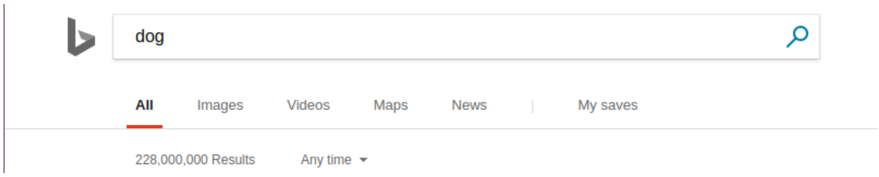
Query: Dog Mumbai

For Query, Dog: 228,000,000

For Query, Mumbai: 89,000,000

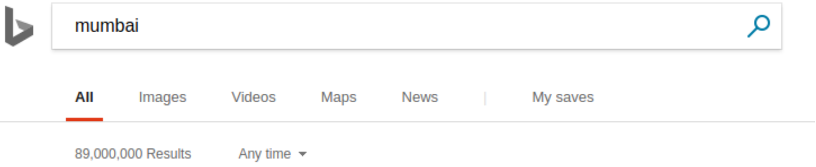
For Query, Dog Mumbai: 19,100,000

Total collection size = 1,062,000,000


dog

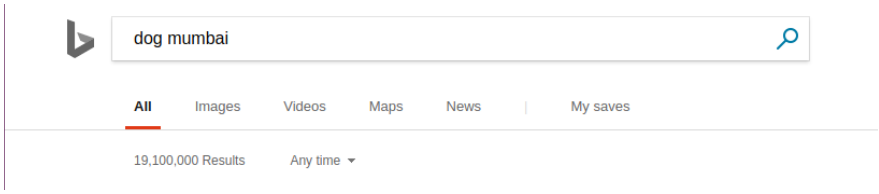
228,000,000 Results

Figure 3.1: Bing Results for Dog


mumbai

89,000,000 Results

Figure 3.2: Bing Results for Mumbai


dog mumbai

19,100,000 Results

Figure 3.3: Bing Results for Dog Mumbai

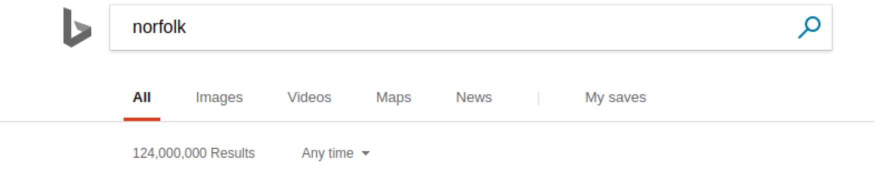
Query:Norfolk Nauman

For Query, Norfolk: 124,000,000

For Query, Nauman: 3,820,000

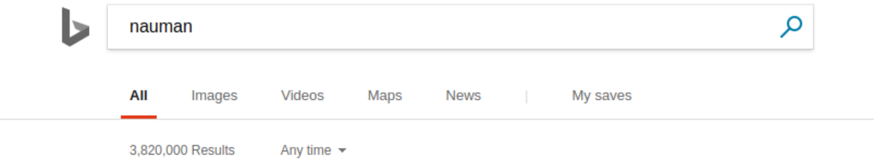
For Query, Norfolk Nauman: 357,000

Total collection size = 3,249,000,000


norfolk

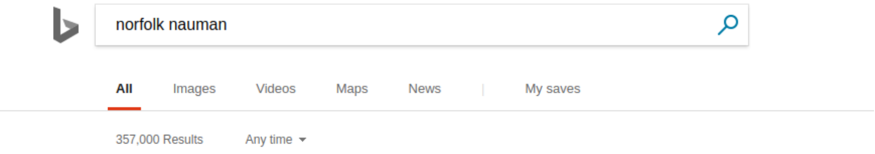
124,000,000 Results

Figure 3.4: Bing Results for Norfolk


nauman

3,820,000 Results

Figure 3.5: Bing Results for Nauman


norfolk nauman

357,000 Results

Figure 3.6: Bing Results for Norfolk Nauman

3.2.2 Google Results

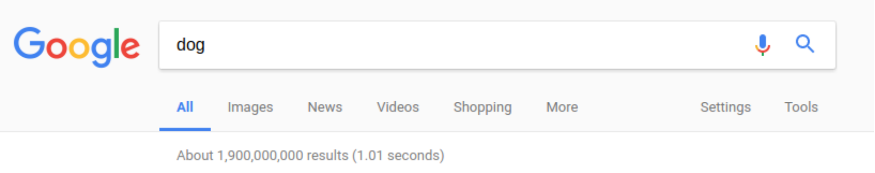
Query: Dog Mumbai

For Query, Dog: 1,900,000,000

For Query, Mumbai: 391,000,000

For Query, Dog Mumbai: 17,000,000

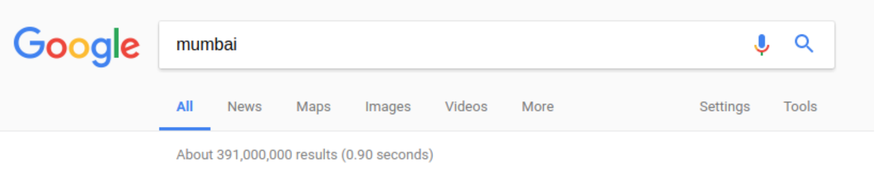
Total collection size = 43,700,000,000



dog

About 1,900,000,000 results (1.01 seconds)

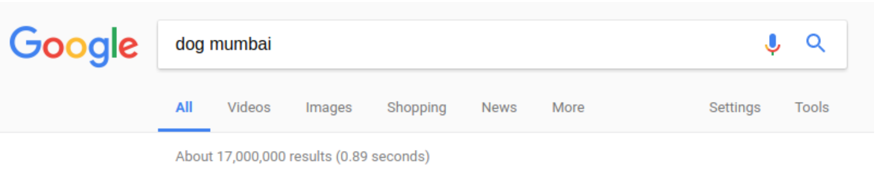
Figure 3.7: Google Results for Dog



mumbai

About 391,000,000 results (0.90 seconds)

Figure 3.8: google Results for Mumbai



dog mumbai

About 17,000,000 results (0.89 seconds)

Figure 3.9: Google Results for Dog Mumbai

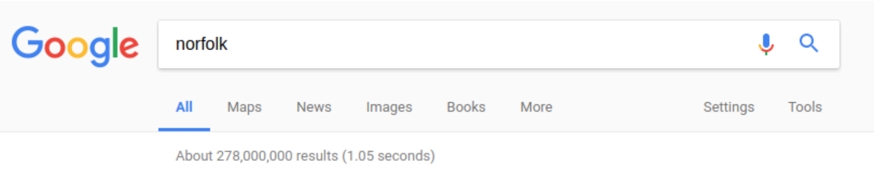
Query:Norfolk Nauman

For Query, Norfolk: 278,000,000

For Query, Nauman: 3,880,000

For Query, Norfolk Nauman: 332,000

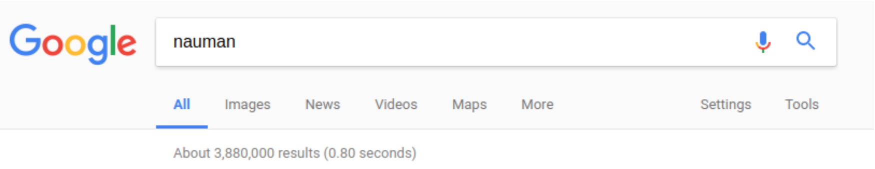
Total collection size = 3,249,000,000



norfolk

About 278,000,000 results (1.05 seconds)

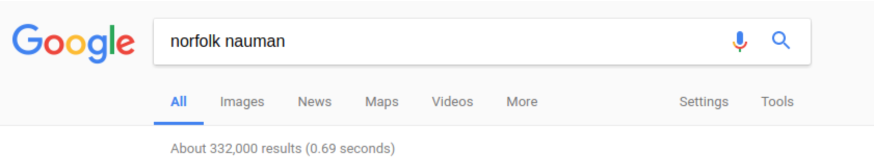
Figure 3.10: Google Results for Norfolk



nauman

About 3,880,000 results (0.80 seconds)

Figure 3.11: Google Results for Nauman



norfolk nauman

About 332,000 results (0.69 seconds)

Figure 3.12: Google Results for Norfolk Nauman

3.3 Correlated Query

3.3.1 Bing Results

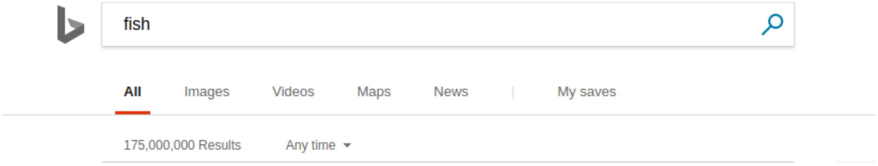
Query: Tropical Fish

For Query, Tropical: 117,000,000

For Query, Fish: 175,000,000

For Query, Tropical Fish: 19,500,000

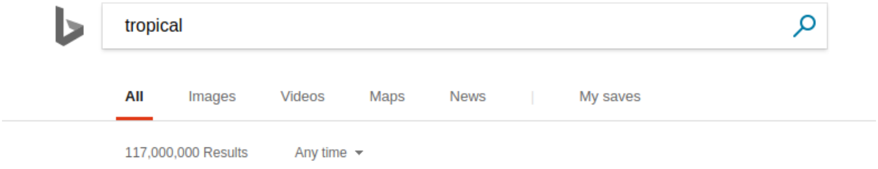
Total collection size = 1,050,000,000



fish

175,000,000 Results

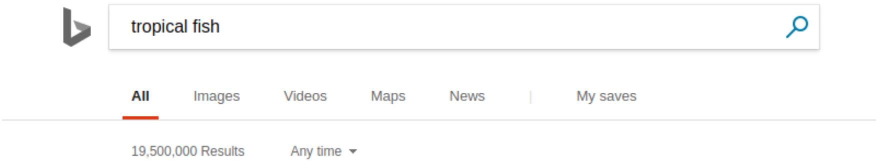
Figure 3.13: Bing Results for Fish



tropical

117,000,000 Results

Figure 3.14: Bing Results for Tropical



tropical fish

19,500,000 Results

Figure 3.15: Bing Results for Tropical Fish

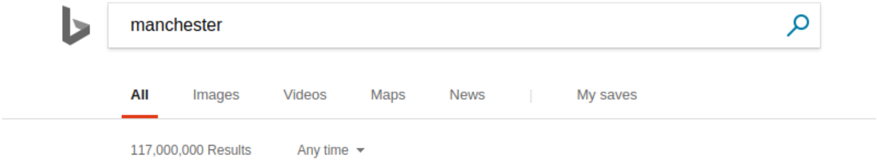
Query:Manchester United

For Query, Manchester: 117,000,000

For Query, United: 409,000,000

For Query, Manchester United: 6,540,000

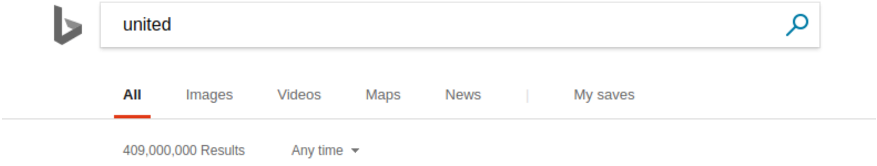
Total collection size = 7,310,000,000



manchester

117,000,000 Results

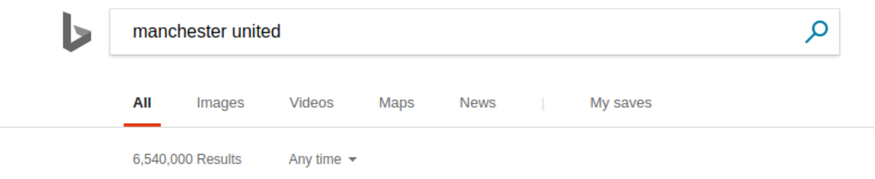
Figure 3.16: Bing Results for Manchester



united

409,000,000 Results

Figure 3.17: Bing Results for United



manchester united

6,540,000 Results

Any time

Figure 3.18: Bing Results for Manchester United

3.3.2 Google Results

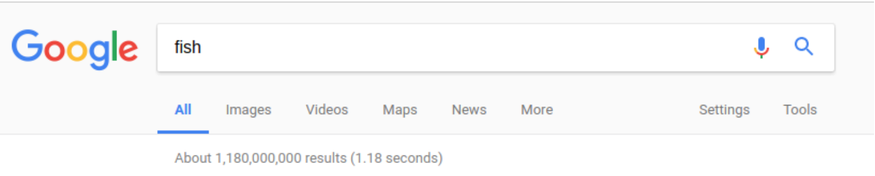
Query: Tropical Fish

For Query, Tropical: 571,000,000

For Query, Fish: 1,180,000,000

For Query, Tropical Fish: 137,000,000

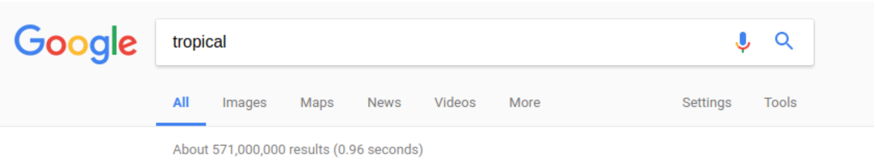
Total collection size = 4,920,000,000



fish

About 1,180,000,000 results (1.18 seconds)

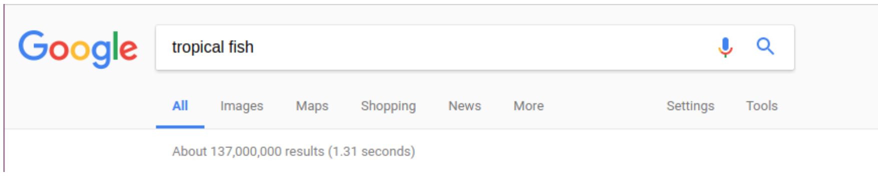
Figure 3.19: Google Results for Fish



tropical

About 571,000,000 results (0.96 seconds)

Figure 3.20: Google Results for Tropical



tropical fish

About 137,000,000 results (1.31 seconds)

Figure 3.21: Google Results for Tropical Fish

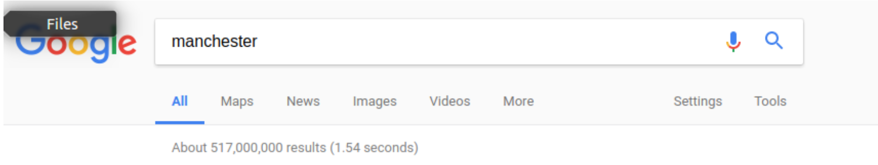
Query:Manchester United

For Query, Manchester: 517,000,000

For Query, United: 4,840,000,000

For Query, Manchester United: 145,000,000

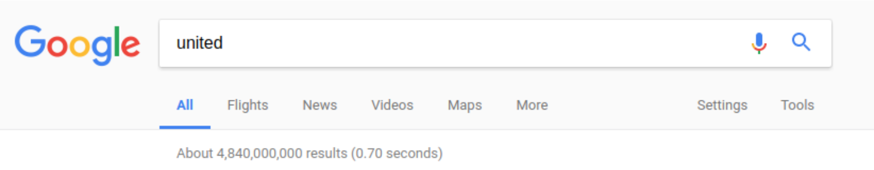
Total collection size = 172,700,000,000



manchester

About 517,000,000 results (1.54 seconds)

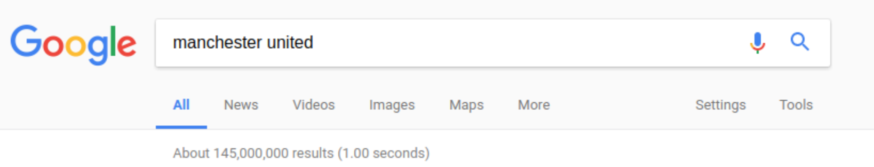
Figure 3.22: Google Results for Manchester



united

About 4,840,000,000 results (0.70 seconds)

Figure 3.23: Google Results for United



manchester united

About 145,000,000 results (1.00 seconds)

Figure 3.24: Google Results for Manchester United

3.4 Conclusion

The formula used to estimate the collection size of a search engine assumes that all the query terms are independent. This assumption is not valid for many query terms which are most likely to be correlated. In the above solution the collection size for Google does not converge to anypoint. It fluctuates by degree of 1, which is not acceptable in estimating the collection size. For Bing the collection size results have a bit convergence but the value of the collection size is very much dependent on the query terms.

Chapter 4

Problem 4.8

4.1 Problem Statement

Find the 10 Wikipedia documents with the most inlinks. Show the collection of anchor text for those pages.

4.2 Top 10 Wikipedia Documents with most Inlinks

The results are based on Wikipedia small dataset. Top 10 Documents with most inlinks are:

Frequency File Name

12088 index.html

12086 Wikipedia%7EAbout_8d82.html

6043 Wikipedia%7EGeneral_disclaimer_3e44.html

6043 Wikipedia%7EContact_us_afd6.html

6043 Wikipedia%7ECommunity_Portal_6a3c.html

6043 Special%7ERecentChanges_e0d0.html

6043 Portal%7EFeatured_content_5442.html

6043 Portal%7ECurrent_events_bb60.html

6043 Portal%7EContents_b878.html

6043 Help%7EContents_22de.html

Anchor texts for all the top 10 Documents are:

1. Anchor text for Index.html: Main Page
2. Anchor text for Wikipedia%7EAbout_8d82.html: About Wikipedia
3. Anchor text for Wikipedia%7EContact_us_afd6.html: Contact Wikipedia
4. Anchor text for Wikipedia%7ECommunity_Portal_6a3c.html: Community portal
5. Anchor text for Wikipedia%7EGeneral_disclaimer_3e44.html: Disclaimers
6. Anchor text for Portal%7ECurrent_events_bb60.html:Current events
7. Anchor text for Portal%7EFeatured_content_5442.html: Featured content
8. Anchor text for Special%7ERecentChanges_e0d0.html: Recent changes
9. Anchor text for Portal%7EContents_b878.html:Contents
10. Anchor text for Help%7EContents_22de.html: Help

These top 10 documents do not lie in the Wikipedia small dataset. The reason for their large number of links is due to their presence on almost every page. There was total 111 inlinks to the pages in the dataset. The top 10 pages with most inlinks are arranged in increasing order:

Frequency	File Name
2	the United States.html
2	township.html
2	townships.html
2	the South Sandwich Islands.html
3	settlements.html
5	Portal.html
8	(at 25 C, 100 kPa).html
9	states.html
18	community.html
54	communities.html

Chapter 5

Problem 5.14

5.1 Problem Statement

In section 5.7.3, we saw that the optimal skip distance c can be determined by minimizing the quantity kn/c + pc/2, where k is the skip pointer length, n is the total inverted list size, c is the skip interval, and p is the number of postings to find. Plot this function using k = 4, n = 1,000,000, and p = 1,000, but varying c. Then, plot the same function, but set p = 10,000. Notice how the optimal value for c changes. Finally, take the derivative of the function kn/c + pc/2 in terms of c to find the optimum value for c for a given set of other parameters (k, n, and p).

5.2 Optimal Skip Distance

For value k = 4, n = 1,000,000, and p = 1,000 and c in [5,50] with step-size of 5. As the value of c increases, the total number of bytes read decreases. With respect to this graph the value of c is too large to be considered for optimal skip distance. The optimal skip distance will lie above 50 in range of approx 60 to 70 for being optimal.

X-axis of graph shows optimal skip distance and Y- axis represents total number of bytes read.

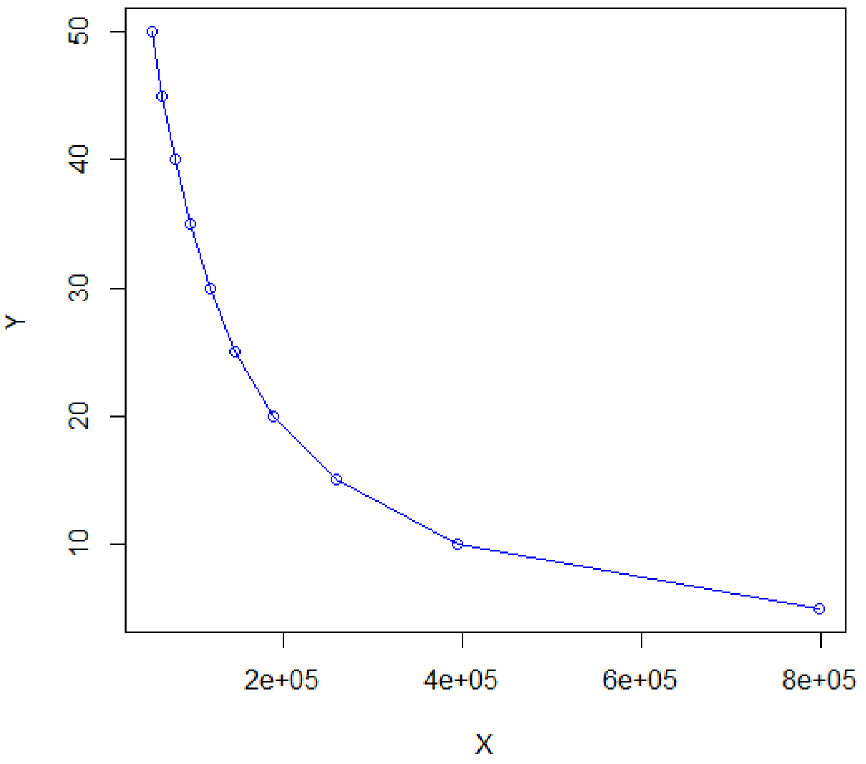


Figure 5.1: Case 1

For value k = 4, n = 1,000,000, and p = 10,000 and c in [5,50] with step-size of 5. The total number of bytes read decreases below 0 when p is increased from 1,000 to 10,000 for the range of C = 5, 10, 15, 20, 25, 30, 35, 40, 45, 50. Value of c above approx. 27 is not acceptable for this function, as further increase in c drops the total number of bytes read to a negative number. With respect to this graph c is optimum in range 20 to 25 for optimal skip distance

X-axis of graph shows optimal skip distance and Y- axis represents total number of bytes read.

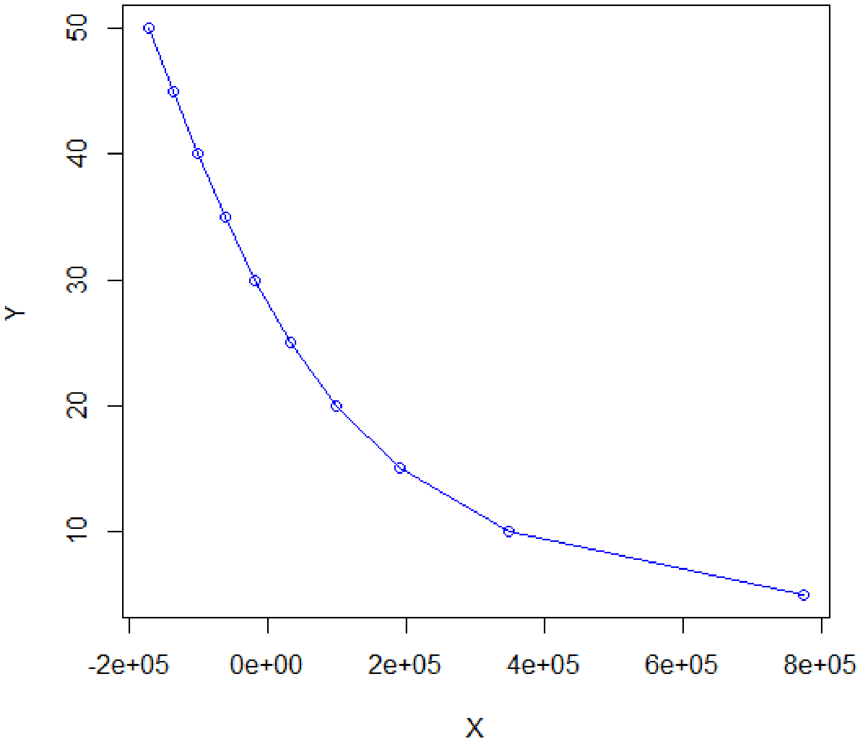


Figure 5.2: Case 2

Differentiating the function:

$$kn/c + pc/2$$

Results to:

$$kn/c^2 + p/2$$

For value = k = 4, n = 1,000,000, and p = 1,000 and c in [0.5,5] with step-size of 0.5. The total value of f bytes read decreases. With respect to this graph c is optimum in around 5 for optimal skip distance

X-axis of graph shows optimal skip distance and Y- axis represents total number of bytes read.

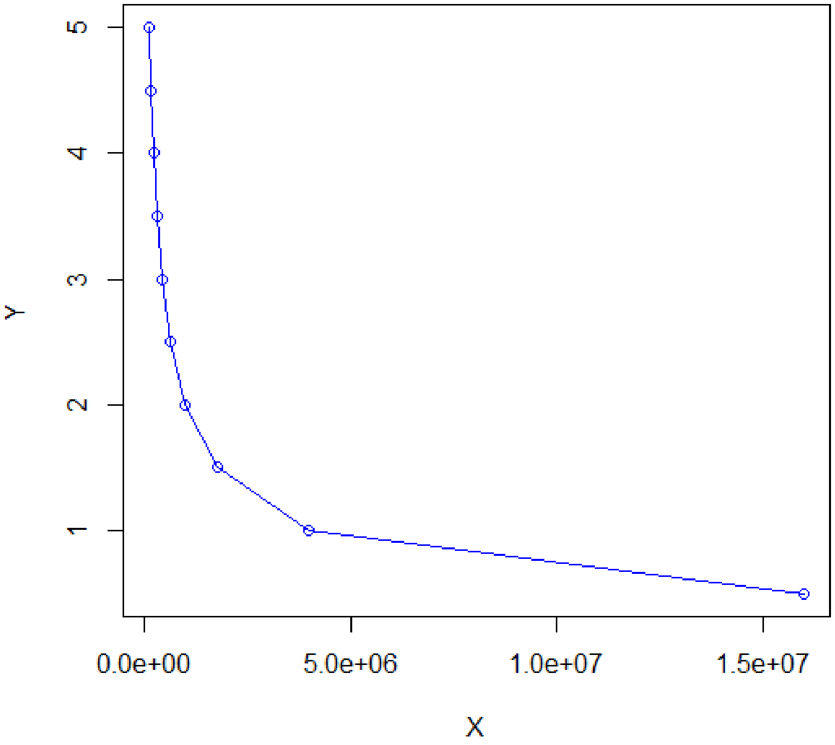


Figure 5.3: Case 3