

# Assignment 2

Mohammed Nauman Sididque

October 14, 2017

# Contents

<b>1</b>	<b>Problem 4.3</b>	<b>3</b>
1.1	Problem Statement . . . . .	3
1.2	Independent Query . . . . .	3
1.2.1	Bing Results . . . . .	3
1.2.2	Google Results . . . . .	6
1.3	Correlated Query . . . . .	8
1.3.1	Bing Results . . . . .	8
1.3.2	Google Results . . . . .	10
1.4	Conclusion . . . . .	12
<b>2</b>	<b>Problem 4.8</b>	<b>13</b>
2.1	Problem Statement . . . . .	13
2.2	Top 10 Wikipedia Documents with most Inlinks . . . . .	13
<b>3</b>	<b>Problem 5.14</b>	<b>16</b>
3.1	Problem Statement . . . . .	16

3.2	Optimal Skip Distance . . . . .	17
-----	---------------------------------	----

# Chapter 1

## Problem 4.3

### 1.1 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.

### 1.2 Independent Query

#### 1.2.1 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

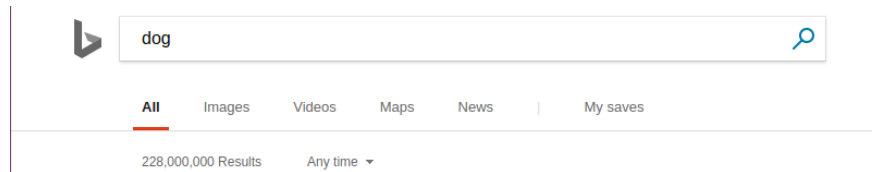


Figure 1.1: Bing Results for Dog

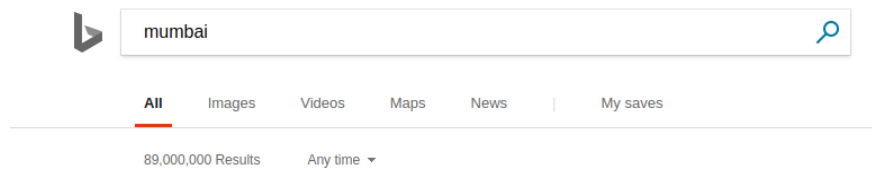


Figure 1.2: Bing Results for Mumbai

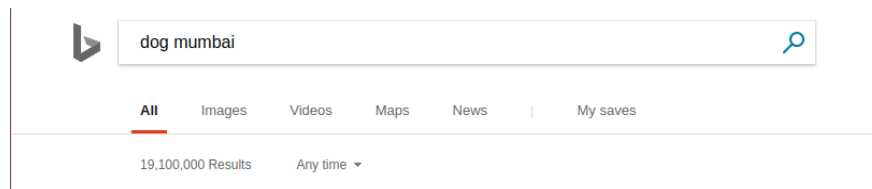


Figure 1.3: Bing Results for Dog Mumbai

**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

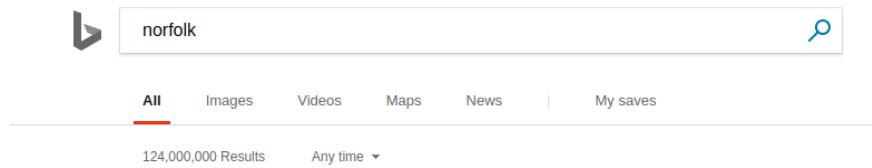


Figure 1.4: Bing Results for Norfolk

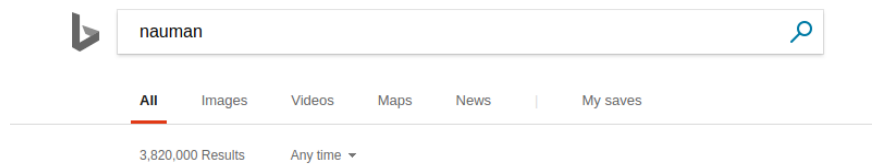


Figure 1.5: Bing Results for Nauman

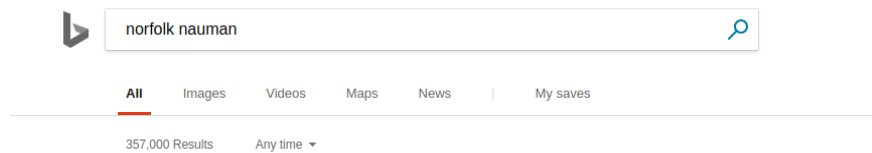


Figure 1.6: Bing Results for Norfolk Nauman

## 1.2.2 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

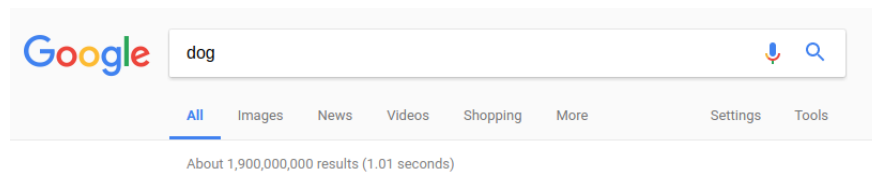


Figure 1.7: Google Results for Dog

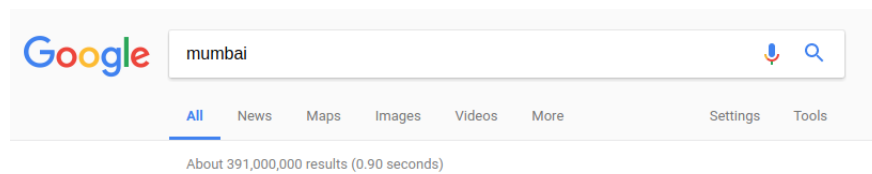


Figure 1.8: google Results for Mumbai

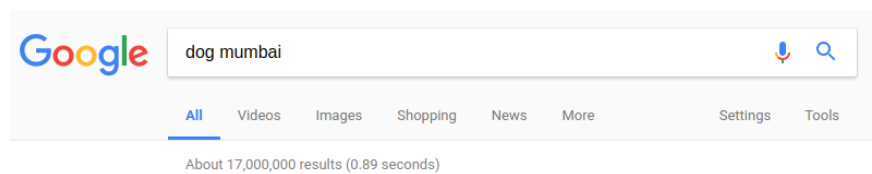


Figure 1.9: Google Results for Dog Mumbai

**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

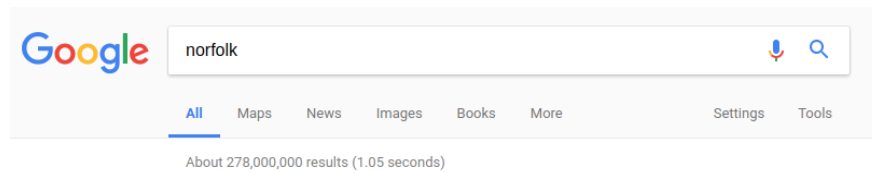


Figure 1.10: Google Results for Norfolk

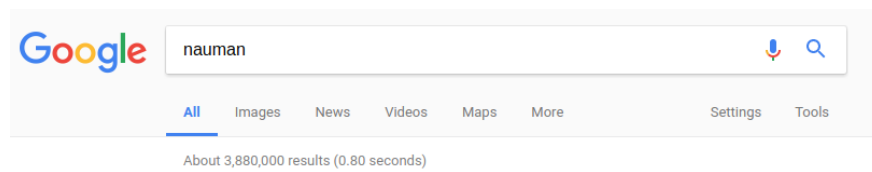


Figure 1.11: Google Results for Nauman

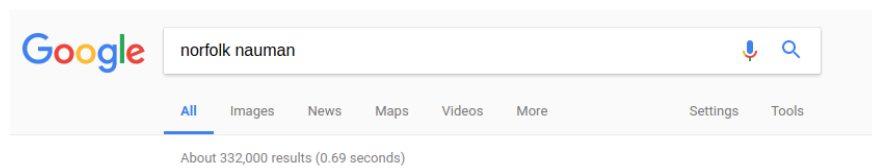


Figure 1.12: Google Results for Norfolk Nauman



## 1.3 Correlated Query

### 1.3.1 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

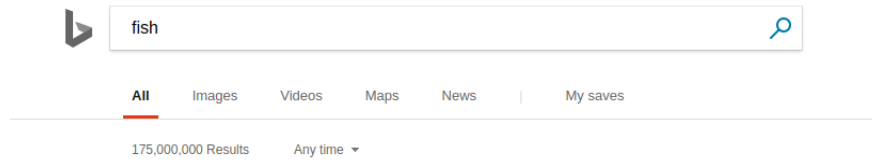


Figure 1.13: Bing Results for Fish

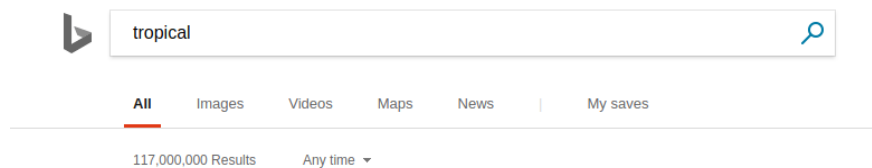


Figure 1.14: Bing Results for Tropical

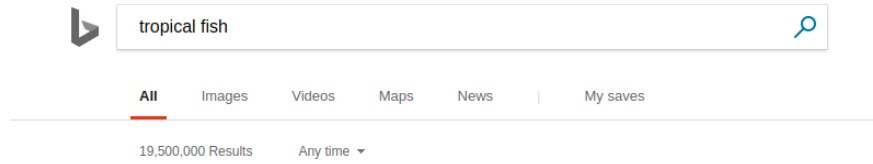


Figure 1.15: Bing Results for Tropical Fish

### 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

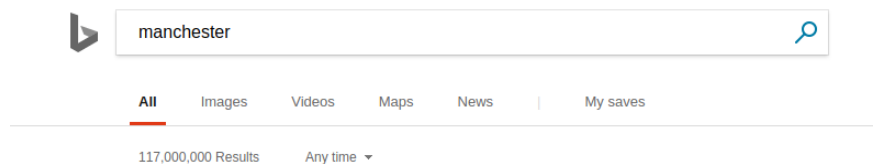


Figure 1.16: Bing Results for Manchester

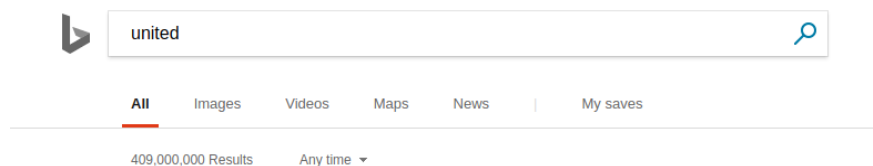


Figure 1.17: Bing Results for United

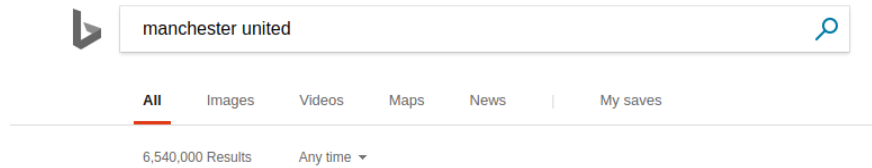


Figure 1.18: Bing Results for Manchester United

## 1.3.2 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

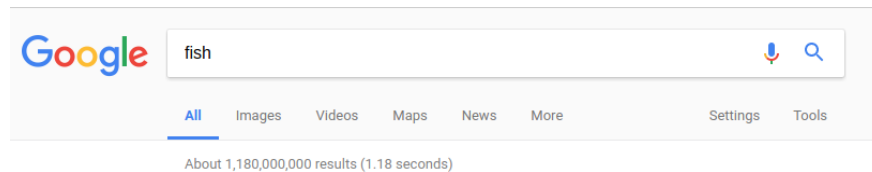


Figure 1.19: Google Results for Fish

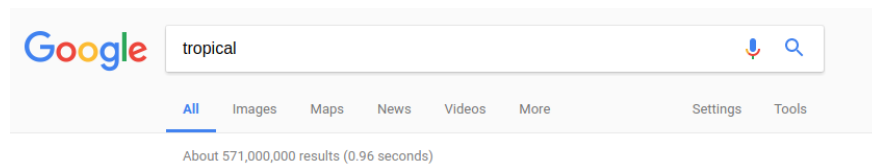


Figure 1.20: Google Results for Tropical

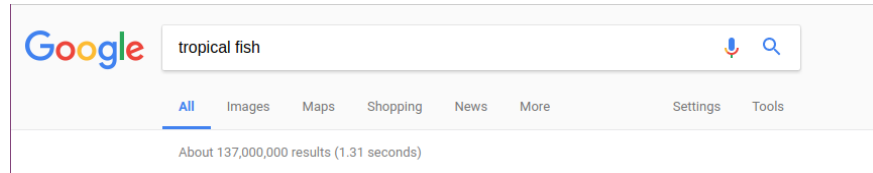


Figure 1.21: Google Results for Tropical Fish

## 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

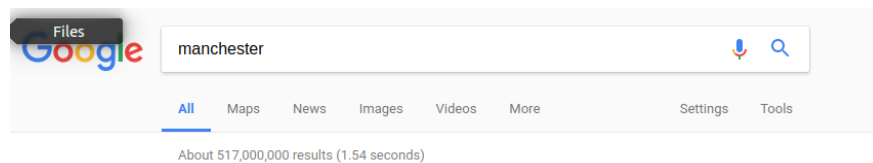


Figure 1.22: Google Results for Manchester

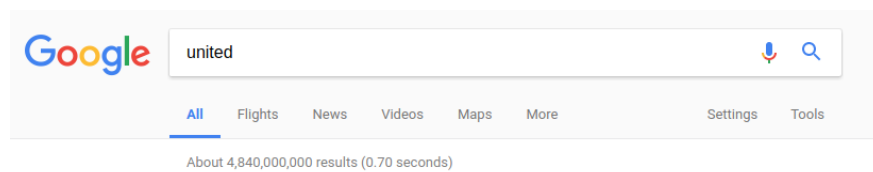


Figure 1.23: Google Results for United

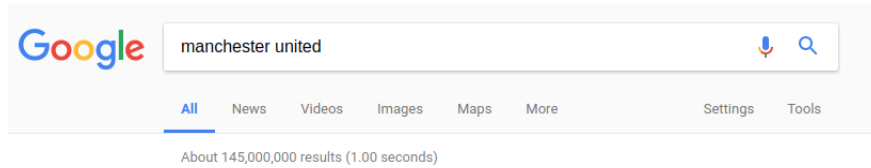


Figure 1.24: Google Results for Manchester United

## 1.4 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 any point. 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 2

## Problem 4.8

### 2.1 Problem Statement

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

### 2.2 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 3

## Problem 5.14

### 3.1 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  $p$ ).

## 3.2 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.

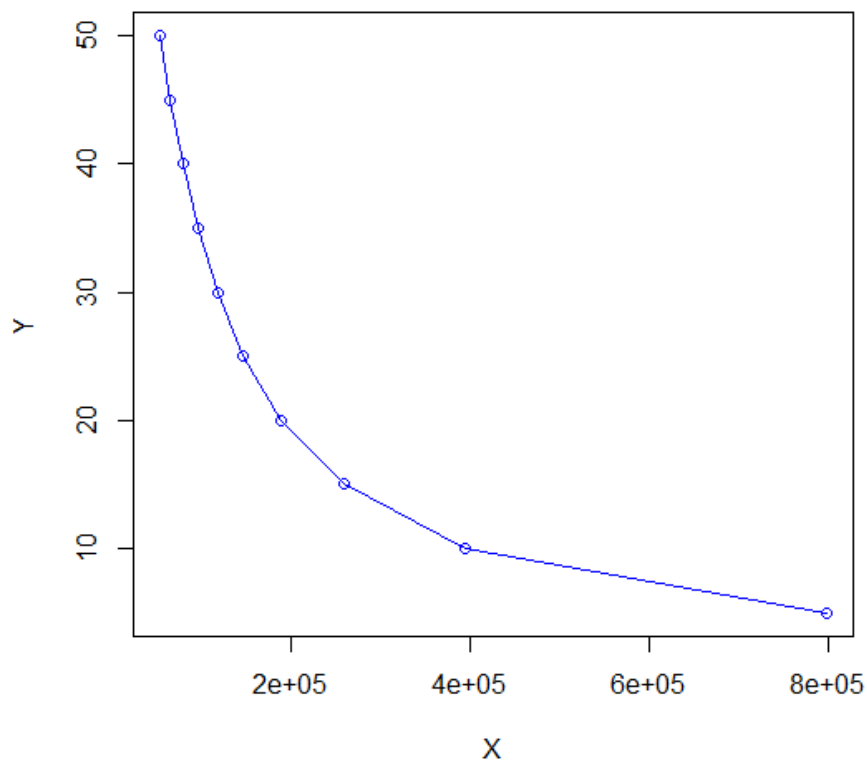


Figure 3.1: Case 1

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.

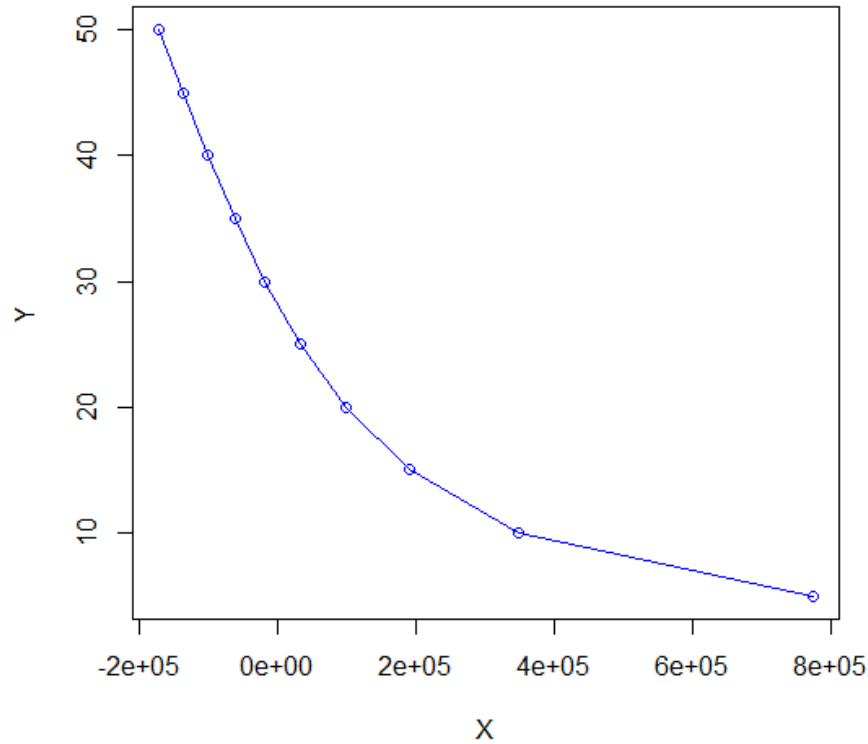


Figure 3.2: Case 2

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.

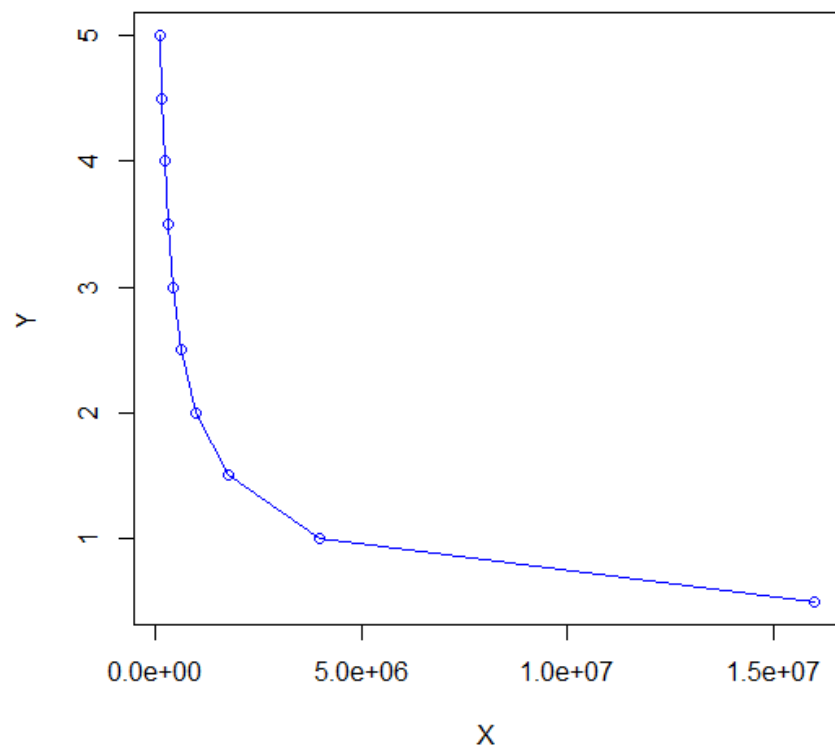


Figure 3.3: Case 3