Report for Assignment 2

Ananya Barat

CS18M524

Task (1) - Comparing Jaccard Similarity with Minhashing

The data from dataword.enron.txt was read in a .csv file called **Input_file.csv** in the same folder. From th is .csv file, a dataframe was constructed having 3 columns for Document ID, Word ID and Word Count. T his implementation is present in CS18M524_PA2_1.ipynb

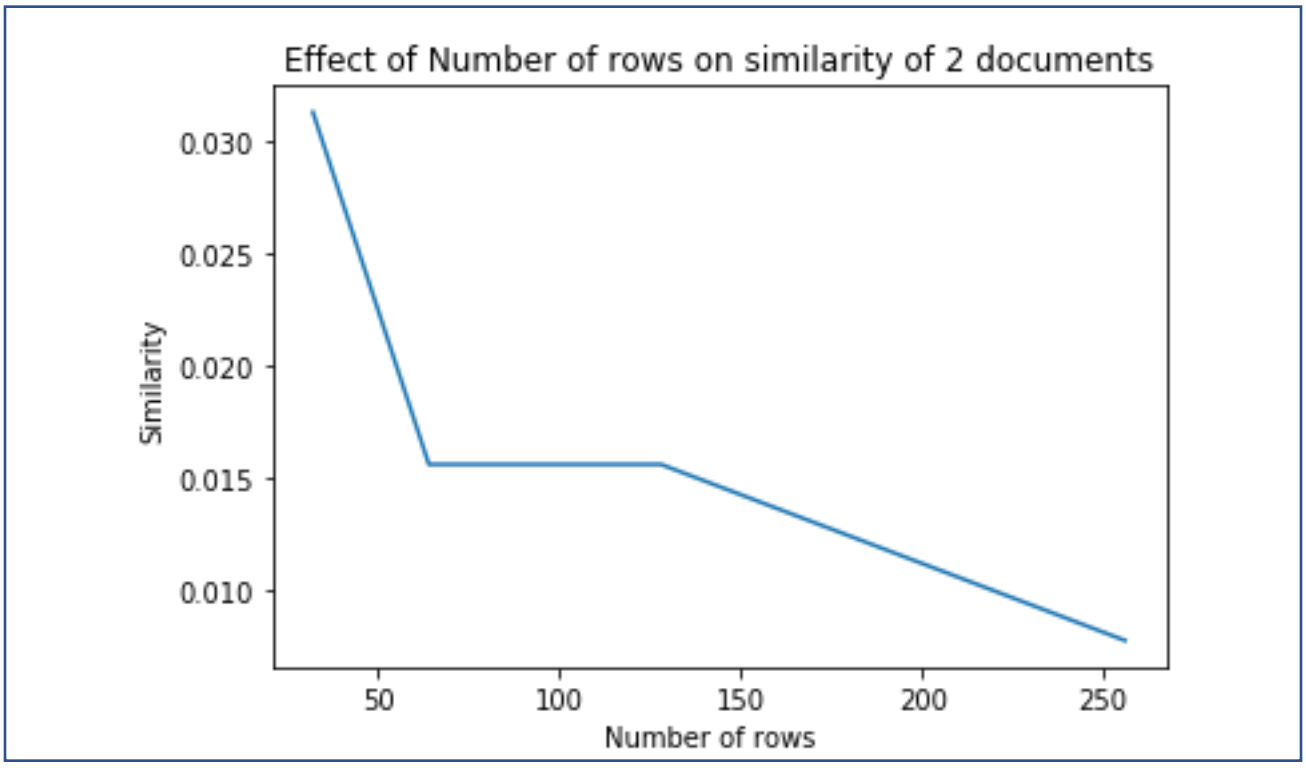
	DocId	Word_Id	Word_Count
0	1	118	1
1	1	285	1
2	1	1229	1
3	1	1688	1
4	1	2068	1

The user has to enter 2 document IDs whose similarity is required to be found.
The selected documents are taken out as chunks from the large dataset by Document ID and processing is done only on these 2 chunks
First the Jaccard similarity between the documents selected is calculated directly. We check for the words common between 2 documents(intersection) and divide it by the number of words in both documents(uni on). This number is reported.

Now the user is asked to enter the number of rows of signature they want and both documents are hashed that many times. For this MinHash function from the datasketch library was used.
The estimated Jaccard Similarity is reported again. The difference between the calculated and estimated J accard similarity is reported as well.

The change in similarity between 2 documents with change in the number of rows of signature matrix , is calculated and displayed.

```
:::::::::: Demonstrating the effect of varying number of rows of Smilarity ma
trix on estimated Jaccard Similarity ::::::::::::
:::::::::: The same 2 documents are used for ease of comparison ::::::::::::
::::::::::::::
The variance in similarity is as follows ::::  [0.03125, 0.015625, 0.015625,
0.0078125]
```



In the above plot we see that the similarity decreases as the number of rows of the signature matrix increases, This is because more number of rows mean more permutations and more the chance that 2 documents can be different unless they are exactly the same.

Now , the use shall select 2 documents by Document ID , the similarity matrix is created for this and the Jaccard Si milarity is calculated directly from the similarity matrix. Every unique word is considered a shingle.

```
Enter the documents for which similarity matrix is
required Documnet Number 1?1
Document Number 2?2
Here we are considering every word as a shingle
:::::::::::::::The Similarity Matrix is ::::::::::::::::::::::::::::
     Word_Id  Doc1
         Doc2 0  118        1           0
1         285     1     0
2        1229     1     0
3        1688     1     0
4        2068     1     1
5        5299     1     0
6        6941     1     0
7        7223     1     0
8        8904     1     0
9        9358     1     0
10       9667     1     0
11       9784     1     0
12      11099     1     0
13      11763     1     0
14      12224     1     0
15      12669     1     0
16      13631     1     1
17      14814     1     0
```

```
18      14816      1      0
19      17208      1      0
20      17872      1      0
21      18139      1      0
22      19190      1      0
23      20240      1      0
24      23028      1      1
25      23481      1      0
26      23893      1      0
27      25611      1      0
28      27283      1      0
29      27359      1      0
..        ...    ...    ...
86      19568      0      1
87      19589      0      1
88      19613      0      1
89      19651      0      1
90      19675      0      1
91      19724      0      1
92      19725      0      1
93      20290      0      1
94      20366      0      1
95      20371      0      1
96      20374      0      1
97      20520      0      1
98      20956      0      1
99      20958      0      1
100     21077      0      1
101     21290      0      1
102     21913      0      1
103     22310      0      1
104     22435      0      1
105     22934      0      1
107     23557      0      1
108     24174      0      1
109     24436      0      1
110     24445      0      1
111     24521      0      1
112     25469      0      1
113     25539      0      1
114     26076      0      1
115     26297      0      1
116     26324      0      1

[114 rows x 3 columns]
Calculated Jaccard Similarity :::::::::::: 0.02631578947368421
```

Then the signature matrix is computed for this similarity matrix with the user requested number of rows. To do this the document is permuted the requisite number of times using the sample function and everyti me the row containing the first 1 is selected as signature for each document. The estimated Jaccard Simila rity is calculated from this signature matrix.

```
Enter the number of rows you want in signature matrix ::120
:::::::::::::::::::: The Signature Matrix for the same 2 documents :::::::
::::::::::::::::::::
     Doc1  Doc2
0      24    24
1      27    52
2      20   112
3      13    43
4       9   106
5      29    87
6      26    75
7      22    53
8      20   100
9       9    60
10     26   108
11      8    52
12     23    94
13      2    72
14     22    96
15      8    79
16     14    34
17     28    68
18      6   100
19     16   106
20     11    37
21     28   113
22     10    46
23     20    71
24      8    63
25     14    59
26     20    72
27     15    60
28     21    62
29     29    43
..    ...   ...
90     22    53
91     24    44
92     15    31
93     12    98
94     20    74
95     29    49
96     15    73
97      6    62
98     20    32
99     14    55
100    29    69
101    12   107
102    19    36
103     1    63
104    17    76
105     7    30
```

```
106     23     86
107     24    105
108     12    109
109      2     61
110     13     48
111      4      4
112      4     92
113      3     96
114      7     97
115      6    104
116     24     24
117     16     51
118      3     38
119      6    107

[120 rows x 2 columns]
Estimated Jaccard Similarity :::::::::::: 0.05
```

The actual Jaccard similarity is 0.02631578947368421 and the estimated Jaccard similarity is 0.05.

Task (2) - Finding nearest neighbors

The same dataset dataword.enron.txt is read again and we convert the entire data to a dataframe having 3 columns Document ID , Word ID and Word Count. This implementation is present in CS18M524_PA2_2 .ipynb

Brute Force Method

The user is asked to select a document and the number of near neighbors of that document that the user wi shes to find. In this section we will simply compare all words of the selected document to the target docu ment in a Brute Force way.
Now, since the number of documents is very high , we will randomly pick 3 times k number of document s. We will then calculate the similarity of each of these documents with the selected document using the B rute force method i.e. compare every word in both documents. We will calculate the Jaccard similarity of the chosen document with each one of the 3 * k randomly picked documents and store the similarity.

Then, out of the 3*k documents, we will choose the top k that have the highest Jaccard Similarity and dis play the same to the user. The running time and the average similarity of the Brute Force method is calcu lated and displayed as well

```
:::::::::::::::::::::::: Brute Force Method ::::::::::::::::::::::::::::
:::::::::
Enter document whose nearest neighbours you want to find :::::::543
Enter the number of neighbours you want to find ::::::::23
K nearest neighbours for the given document are :::::
1  DocId ::: 15085  Similarity ::: 0.04712041884816754
2  DocId ::: 29202  Similarity ::: 0.04262295081967213
3  DocId ::: 31151  Similarity ::: 0.04
4  DocId ::: 31188  Similarity ::: 0.039923954372623575
5  DocId ::: 34751  Similarity ::: 0.03875968992248062
```

```
6  DocId ::: 14283  Similarity ::: 0.03870967741935484
7  DocId ::: 35643  Similarity ::: 0.03759398496240601
8  DocId ::: 30092  Similarity ::: 0.0365296803652968
9  DocId ::: 38106  Similarity ::: 0.035175879396984924
10  DocId ::: 34935  Similarity ::: 0.034722222222222224
11  DocId ::: 4503  Similarity ::: 0.033112582781456956
12  DocId ::: 4037  Similarity ::: 0.03225806451612903
13  DocId ::: 24266  Similarity ::: 0.03201970443349754
14  DocId ::: 24392  Similarity ::: 0.031413612565445025
15  DocId ::: 15659  Similarity ::: 0.029940119760479042
16  DocId ::: 39351  Similarity ::: 0.02631578947368421
17  DocId ::: 16974  Similarity ::: 0.026119402985074626
18  DocId ::: 31757  Similarity ::: 0.026041666666666668
19  DocId ::: 15492  Similarity ::: 0.024691358024691357
20  DocId ::: 32533  Similarity ::: 0.024242424242424242
21  DocId ::: 29929  Similarity ::: 0.022388059701492536
22  DocId ::: 33389  Similarity ::: 0.01948051948051948
23  DocId ::: 15521  Similarity ::: 0.0189873417721519
Running Time in seconds of Brute Force:::::  7.314160108566284
Average similarity of documents using brute force method :::::: 0.03003133
8787957374
```

LSH Method

The user is asked to select a document Id for which the nearest neighbours shall be calculated by LSH. Th e user is also requested to enter the number of bands and the number of rows per band that they want , as well as the number of permutations. All these inputs are then used to calculate the nearest neighbors using the MinHashLSH and MinHash functions from datsketch library. The running time and the average similarity of the LSH method is calculated and displayed as well. The LSH method has a much lower running time with almost same similarity.

```
:::::::::::::::::::::::: LSH Method ::::::::::::::::::::::::::::::::::::
:
Enter the number of bands you want :::::::30
Enter the number of rows per bands you want :::::::1
Enter the number of rows of signature matrix you want :::::::256
Enter the dataset size you want :::::::200
Enter document whose nearest neighbours you want to find :::::::23
Approximate neighbours are [10753, 5633, 12675, 9732, 38152, 37514, 39434,
33661, 33807, 10256, 26640, 39187, 21398, 21016, 21913, 5916, 5926, 14204,
15403, 33197, 19502, 4660, 29108, 14390, 11191, 693, 25274, 7359, 11972, 2
4644, 25414, 33224, 13002, 24267, 25295, 11261, 34645, 25045, 869, 12776,
5609, 15212, 23022, 32367, 17519, 23667, 32116, 18037, 36086, 24956, 7165]


Running Time in seconds of LSH:::::  3.615689754486084
Average similarity of documents using LSH Method :::::: 0.0333333333333333
3
```

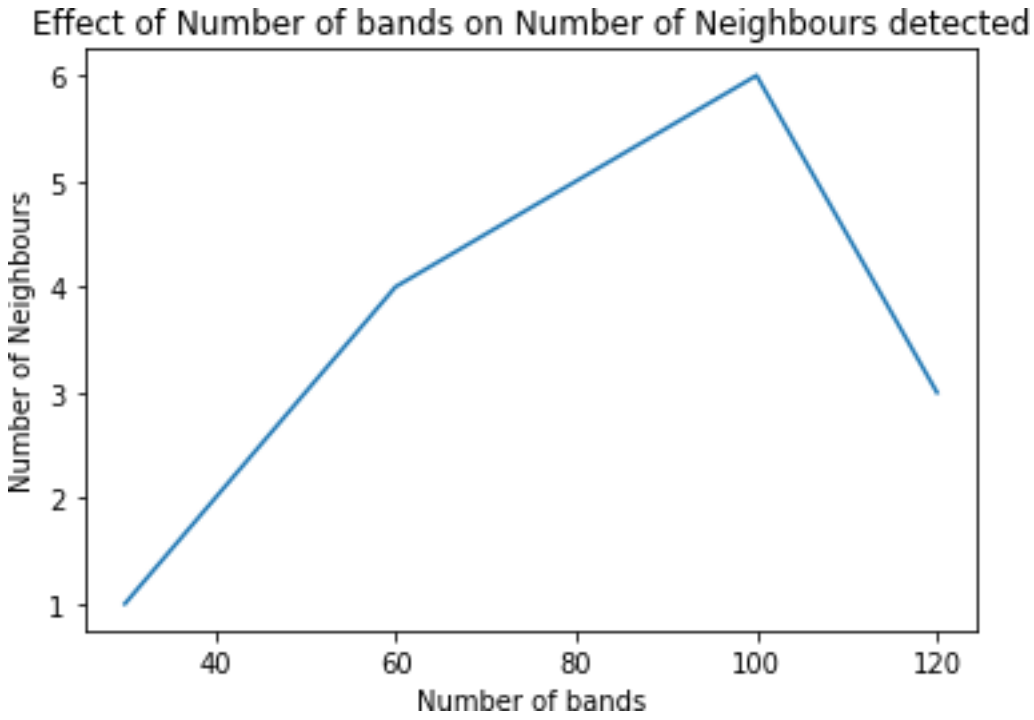
The running time of the LSH method is about half the running time of the Brute force method with the above configurations of bands and rows. And the average similarity is same in both the cases.

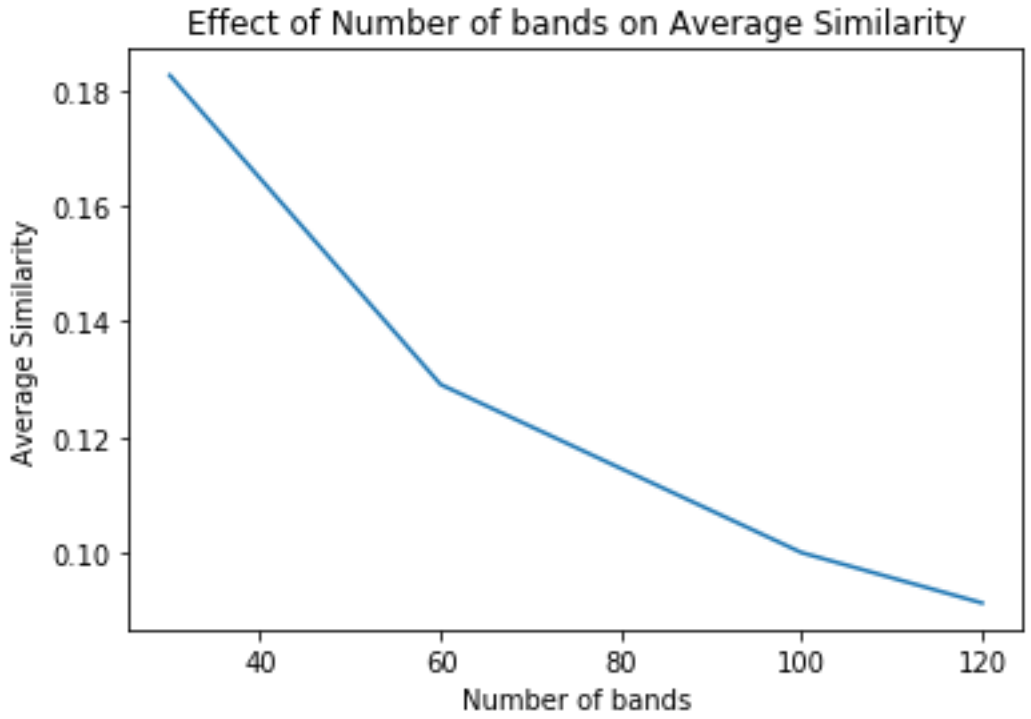
The factors that affect the running time , average similarity and the number of nearest neighbors found are :-

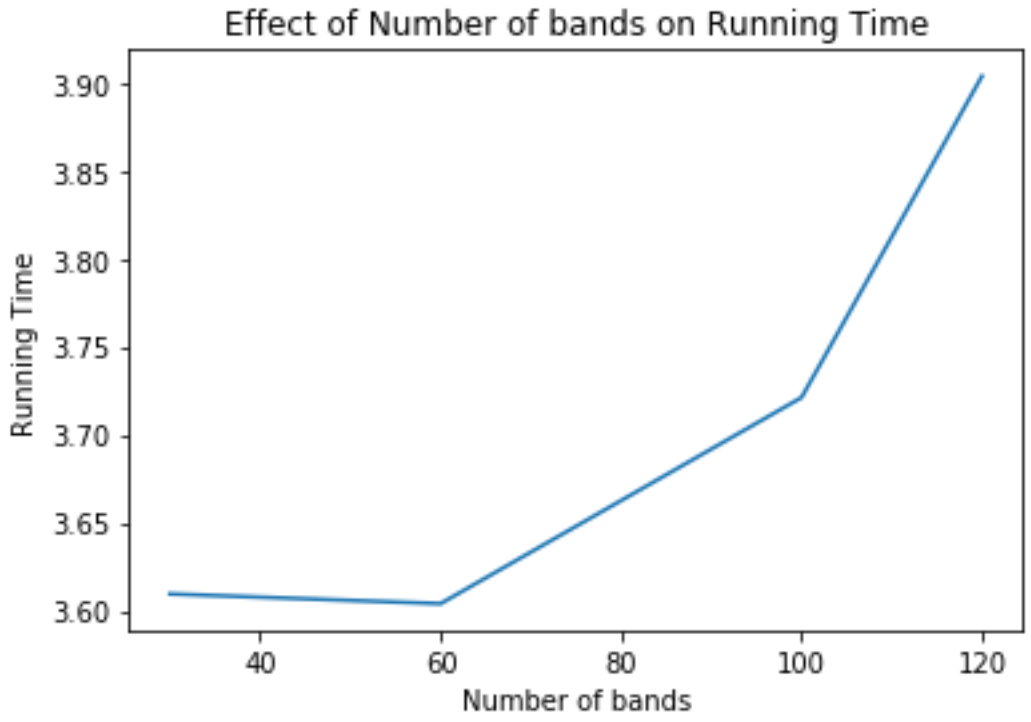
i. The number of bands the signature matrix is divided into
ii. The number of rows in each band of signature matrix
iii. The data size as in the number of total documents under consideration
iv. The number of signatures i.e the number of permutations of the hash functions.

Effect of the number of bands on the average similarity , running time and number of nearest neighbors

```
:::::::::::::::::::::: Demonstrating the effect of varying number of bands :
::::::::::::::::::::::::::::::::
Approximate neighbours are [33682]
Approximate neighbours are [13208, 9769, 33229, 766]
Approximate neighbours are [5318, 9769, 19339, 22862, 22361, 13021]
Approximate neighbours are [2752, 29546, 27508]
Number of neighbours :::: [1, 4, 6, 3]
Average Similarity :::: [0.18257418583505536, 0.12909944487358055, 0.1, 0.091
28709291752768]
Running Time :::: [3.6094839572906494, 3.603961706161499, 3.7211198806762695,
3.9037985801696777]
```



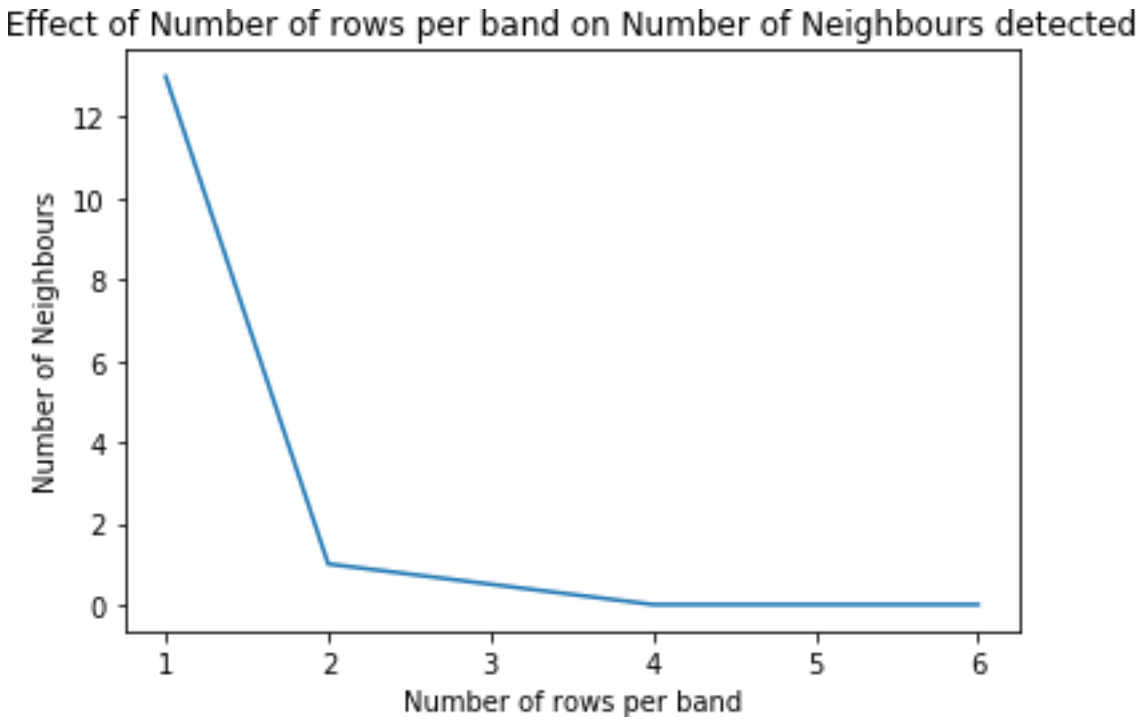


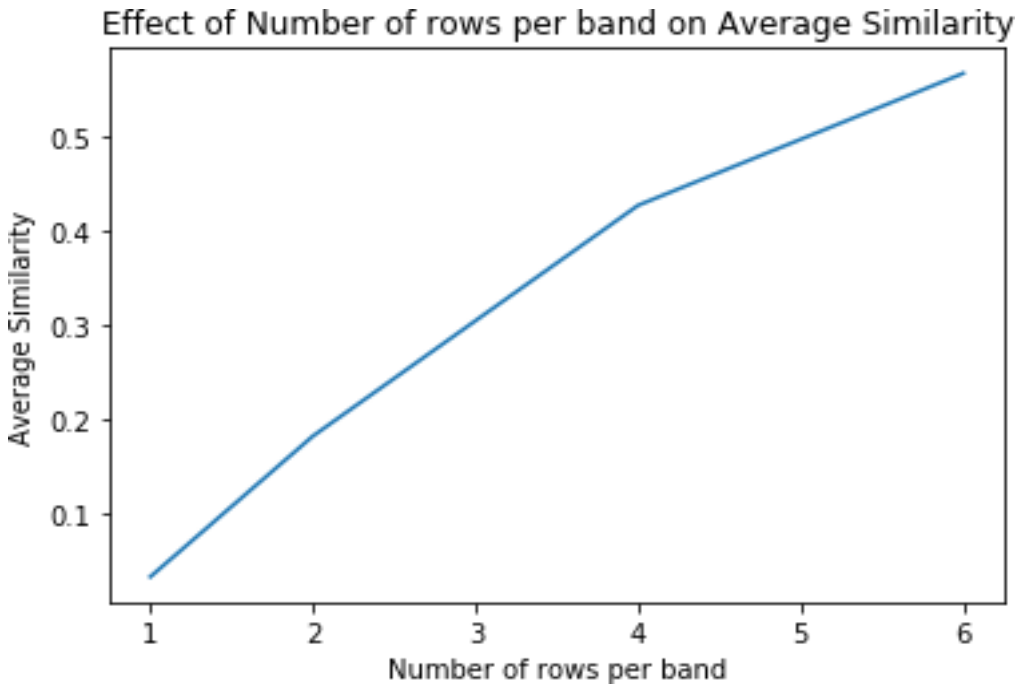


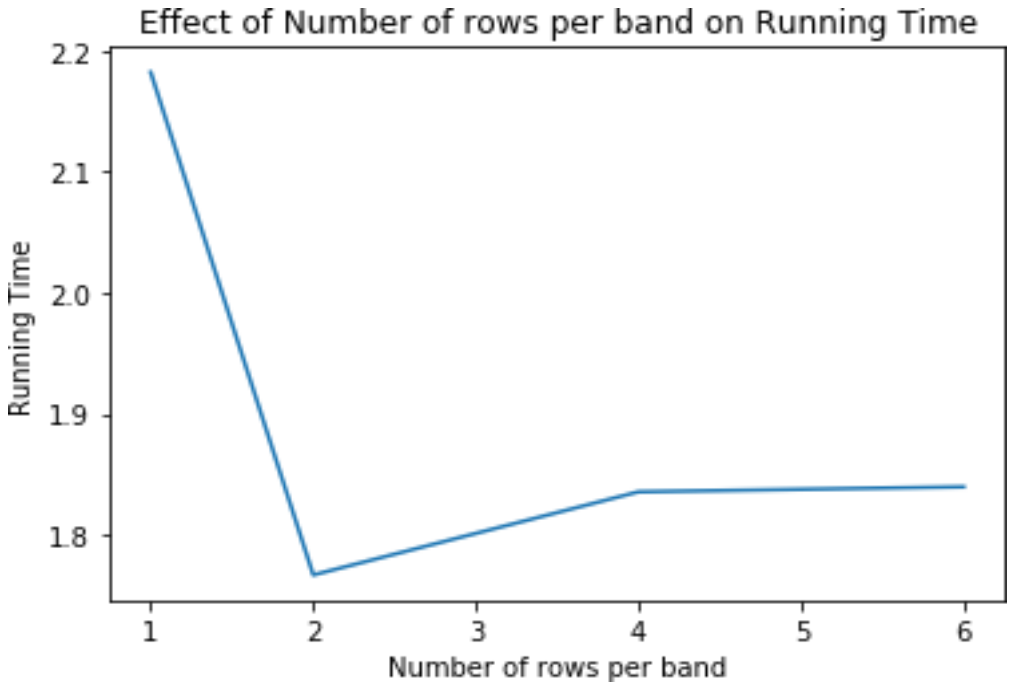
As the number of bands increase, the number of neighbors usually increases. We can look at this like every band is a chance for 2 docs to be similar so more number of bands imply more chances to be similar.
As the number of bands increase, the average similarity decreases and the running time increases. This is justified because similarity is (1/b)^(1/r) and the running time increases because for every band a separate hashing has to be done which increases the calculation.

Effect of the number of rows per band on the average similarity , running time and number of near est neighbors

```
:::::::::::::::::::: Demonstrating the effect of varying number of rows per
band ::::::::::::::::::::::::::::::::
Approximate neighbours are [14881, 16329, 26698, 586, 9743, 39632, 32243, 187
6, 20693, 13237, 11193, 7070, 27263]
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are [13583]
Skipping duplicate doc
Approximate neighbours are []
Skipping duplicate doc
Approximate neighbours are []
Number of neighbours :::: [13, 1, 0, 0]
Average Similarity :::: [0.03333333333333333, 0.18257418583505536, 0.42728700
639623407, 0.5673004449747446]
Running Time :::: [2.1821014881134033, 1.766850233078003, 1.8355274200439453,
1.8396377563476562]
```



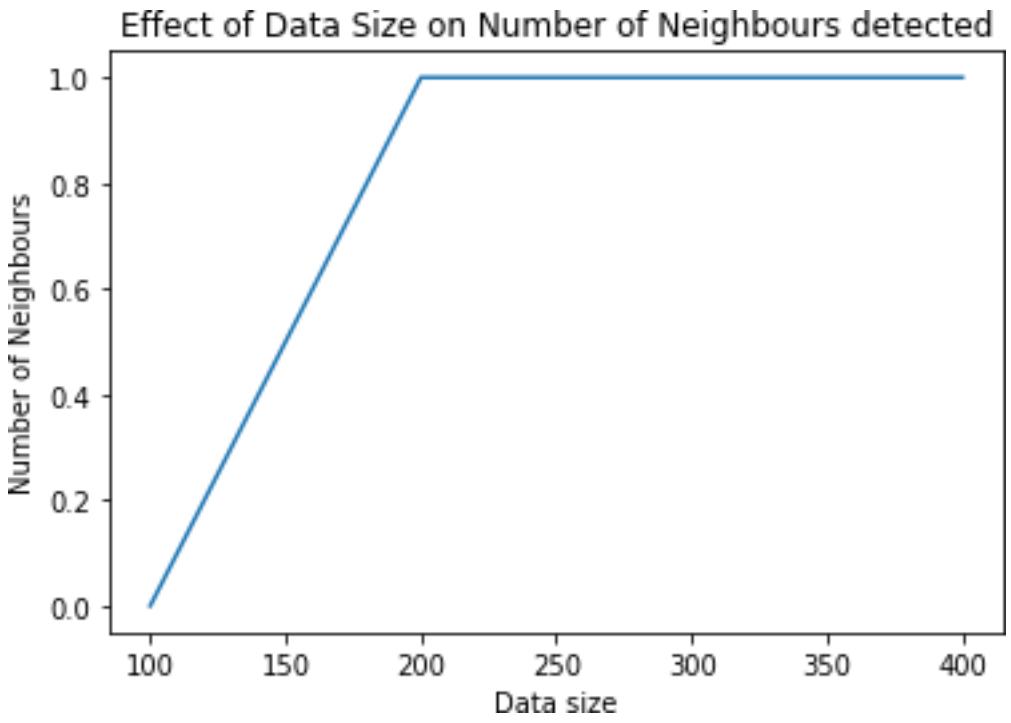


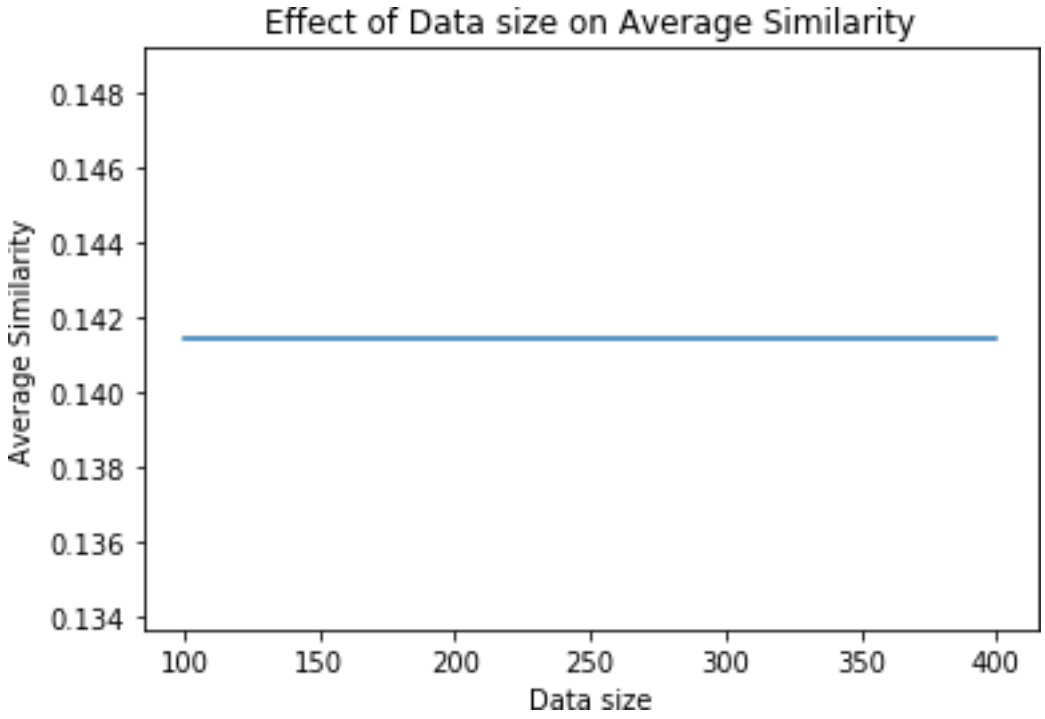


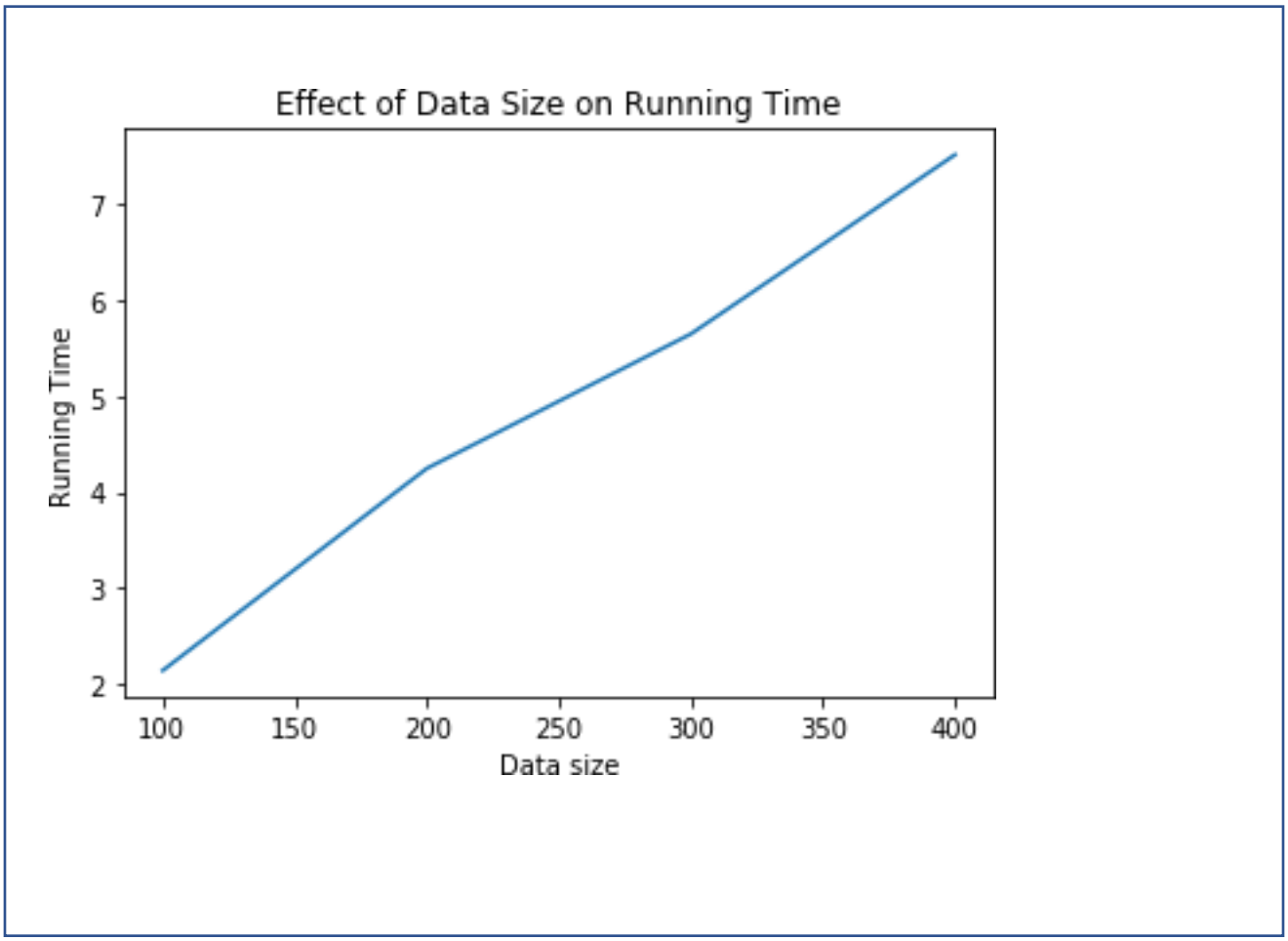
As the number of rows per band increase the number of similar documents i.e the near neighbors decrease. This is because every increase in row implies one more row that 2 dcuments have to be similar in for them to be considered similar. The average similarity increases with increase in r, as it is roughly proportional to (1/b)^(1/r). Running time decreases as more rows per band implies less bands and hence less calculations.

Effect of data size on the average similarity , running time and number of nearest neighbors

```
:::::::::::::::::::: Demonstrating the effect of varying data size :::::::::
:::::::::::::::::::::::
Approximate neighbours are []
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are [13583]
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are [764]
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are [19613]
Number of neighbours :::: [0, 1, 1, 1]
Average Similarity :::: [0.1414213562373095, 0.1414213562373095, 0.1414213562
373095, 0.1414213562373095]
Running Time :::: [2.146419048309326, 4.24815821647644, 5.647330284118652, 7.
511559724807739]
```


Effect of Data Size on Number of Neighbours detected
Number of Neighbours
1.0
0.8
0.6
0.4
0.2
0.0
100
150
200
250
300
350
400
Data size


Effect of Data size on Average Similarity
Average Similarity
0.148
0.146
0.144
0.142
0.140
0.138
0.136
0.134
100
150
200
250
300
350
400
Data size

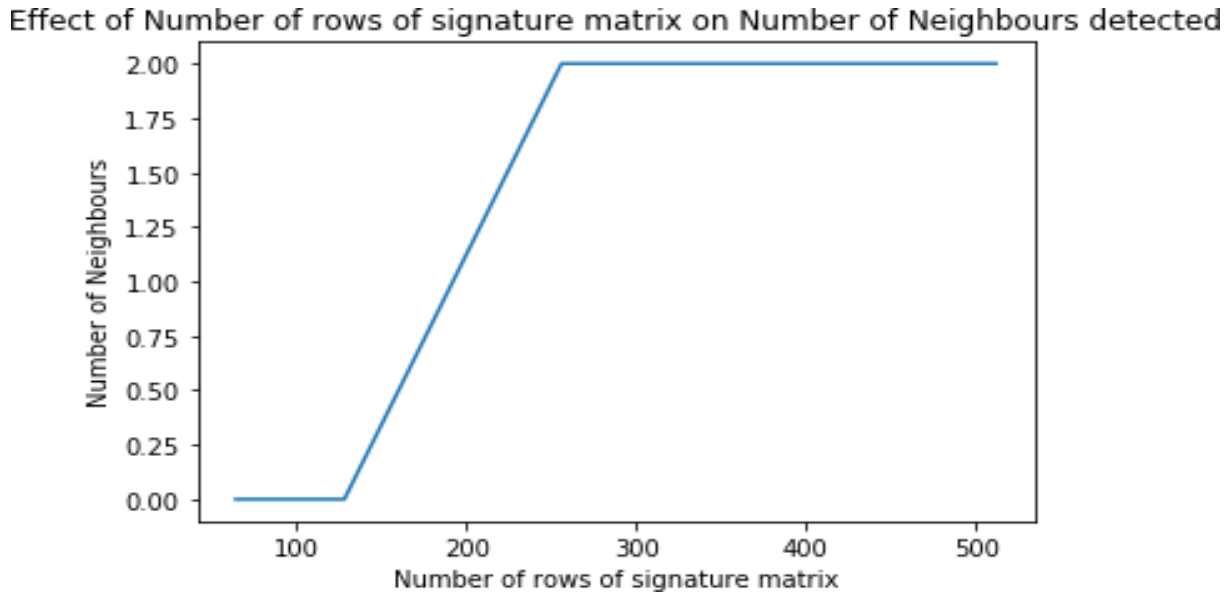


With increase in data size , the number of near neighbors increase because more documents essentially mean more similar documents. The average similarity is constant as average similarity depends only upon b and r and both these parameters are constant. The running time increases because more documents essentially mean mor computation.

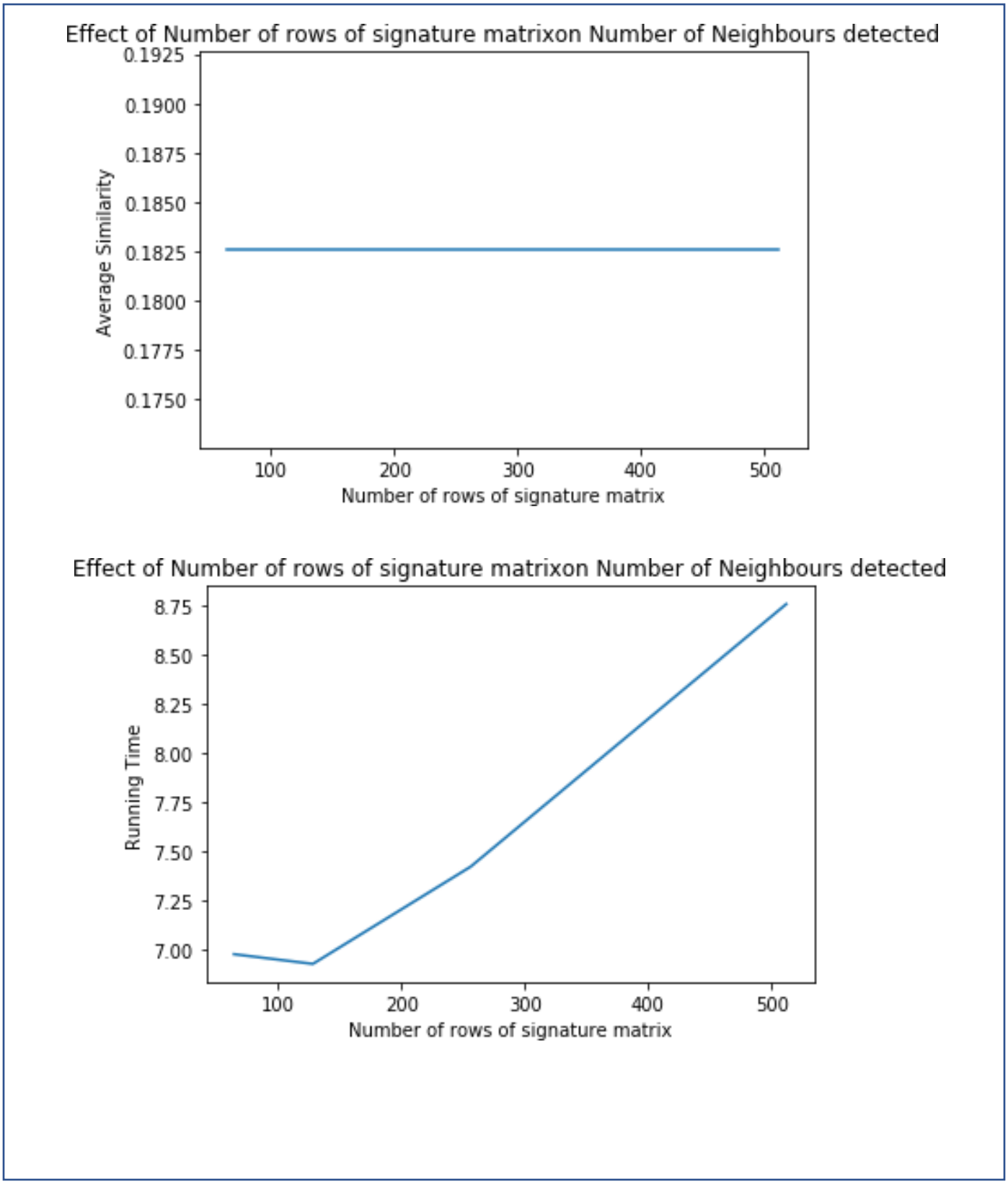
Effect of varying the number of rows of signature matrix (number of permutations) on the average similarity , running time and number of nearest neighbors

```
:::::::::::::::: Demonstrating the effect of varying number of rows of signatu
re matrix ::::::::::::::::::
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are []
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are []
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are [2143, 9503]
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
```

```
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Skipping duplicate doc
Approximate neighbours are [7074, 5507]
Number of neighbours :::: [0, 0, 2, 2]
Average Similarity :::: [0.18257418583505536, 0.18257418583505536, 0.18257418
583505536, 0.18257418583505536]
Running Time :::: [6.976638078689575, 6.927340745925903, 7.421056270599365, 8
.754297971725464]
```



With increase in number of rows of signature matrix, i.e. more permutations the number of near neighbors that get detected increase. The average similarity depends only on b and r and it remains constant as these 2 factors are constant. The running time increases as more permutations will take more time.


Effect of Number of rows of signature matrixon Number of Neighbours detected
0.1925
0.1900
0.1875
0.1850
0.1825
0.1800
0.1775
0.1750
Average Similarity
100
200
300
400
500
Number of rows of signature matrix
Effect of Number of rows of signature matrixon Number of Neighbours detected
8.75
8.50
8.25
8.00
7.75
7.50
7.25
7.00
Running Time
100
200
300
400
500
Number of rows of signature matrix