ASSIGNMENT 4
NAME:-Anand Sharma (MT19059)
Analysis Report

QUESTION 1:-

Dataset used are:-

```
20_newsgroups/sci.med/59199
20_newsgroups/comp.graphics/38703
20_newsgroups/rec.sport.hockey/53700
20_newsgroups/talk.politics.misc/178450
20_newsgroups/sci.space/60196
5000

Total DF dictionary are :- 56729
Total tf_idf Dictionary :-836216
Matrix shape (5000,56729)

Query:-['pretty', 'good', 'opinion', 'biochemistry', 'machine']
5
Pretty good opinions on biochemistry machines
Retrieved Docs are :-  [3305, 973, 2811, 3290, 3981, 2804, 4652,
4968, 2911, 259, 217, 2599, 3426, 978, 75, 3279, 749, 3571, 3352,
2881, 2455, 2871, 3369, 1112, 568, 2817, 1590, 1774, 2367, 2136,
3649, 2874, 2956, 3667, 3791, 2368, 2687, 802, 3360, 3298, 3505,
4848, 611, 3765, 356, 2785, 2373, 3913, 2552, 1271, 852, 101, 2684,
3308, 2052, 1360, 709, 3288, 899, 3728, 12, 2905, 4956, 3376, 944,
4769, 2215, 2835, 2393, 3510, 4053, 4036, 2132, 1642, 59, 2201, 3936,
297, 2851, 3574, 3988, 4377, 2810, 3797, 512, 726, 4193, 13, 4247,
4210, 3962, 912, 3533, 2488, 2679, 1162, 2526, 3655, 1307, 3135]
```

Assumptions:-Assumption of docid are:-

- `(0-999):- 20_newsgroups/sci.med`
  `(1000-1999):-20_newsgroups/comp.graphics`
  `(2000-2999):-20_newsgroups/rec.sport.hockey`
  `(3000-3999):-20_newsgroups/talk.politics.misc`
  `(4000-4999):-20_newsgroups/sci.space`

QUESTION 2:-
(A) for k=100
query="Pretty good opinions on biochemistry machines"
Retrieved DocsID:-[3305, 973, 2811, 3290, 3981, 2804, 4652, 4968, 2911, 259, 217, 2599, 3426, 978, 75, 3279, 749, 3571, 3352, 2881, 2455, 2871, 3369, 1112, 568, 2817, 1590, 1774, 2367, 2136, 3649, 2874, 2956, 3667, 3791, 2368, 2687, 802, 3360, 3298, 3505, 4848, 611, 3765, 356, 2785, 2373, 3913, 2552, 1271, 852, 101, 2684, 3308, 2052, 1360, 709, 3288, 899, 3728, 12, 2905, 4956, 3376, 944, 4769, 2215, 2835, 2393, 3510, 4053, 4036, 1642, 2132, 59, 2201, 3936, 297, 2851, 3574, 3988, 4377, 2810, 3797, 512, 726, 4193, 13, 4247, 4210, 3962, 3533, 912, 2488, 2679, 1162, 2526, 3655, 3135, 1307]

(B)(C) for k=100 and p=10% of k => 10
query="Pretty good opinions on biochemistry machines"
Enter the relevant docs no :-973 259 217 978 75 749 568 802 611 852
Releavnt docs are :-
{802, 259, 611, 75, 973, 749, 978, 852, 568, 217}
Non Releavnt docs are :- {512, 2817, 2052, 12, 13, 2835, 3352, 4377, 1307, 3360, 2851, 2599, 3369, 297, 4652, 3376, 1590, 2871, 2874, 59, 3135, 2367, 3649, 2368, 2881, 2373, 3655, 3913, 1360, 3667, 2132, 1112, 2136, 2905, 2393, 4956, 2911, 3936, 4193, 3426, 356, 101, 4968, 1642, 4210, 2679, 3962, 2684, 2687, 899, 1162, 2956, 3981, 3728, 912, 3988, 4247, 2455, 2201, 4769, 2215, 944, 3505, 3765, 3510, 2488, 4036, 709, 3533, 3279, 3791, 3797, 726, 4053, 3288, 3290, 2526, 2785, 3298, 3305, 3308, 1774, 4848, 3571, 2804, 3574, 1271, 2552, 2810, 2811}
 Iteration 1:-
Retrieved docs are :-
[802, 217, 749, 568, 75, 259, 158, 517, 852, 611, 179, 745, 885, 978, 4591, 4576, 166, 909, 321, 203, 349, 956, 4781, 4806, 310, 768, 1104, 47, 3011, 691, 2923, 973, 1764, 882, 688, 2217, 4660, 709, 944, 746, 3206, 4477, 4568, 886, 44, 151, 4930, 3660, 2227, 4751, 4654, 4215, 4808, 580, 815, 595, 674, 4256, 1371, 3495, 3683, 372, 396, 476, 2279, 4180, 4048, 2688, 2479, 1017, 2656, 1045, 4957, 4115, 2312, 3989, 2272, 2132, 4659, 3170, 2201, 928, 98, 4983, 2011, 804, 204, 441, 333, 4067, 29, 2127, 2390, 57, 3933, 4992, 2140, 2205, 191, 216]

(D) query="Pretty good opinions on biochemistry machines"
      k=100;

OUTPUT:-
Iteration 0
Retrieved docs are :- [3305, 973, 2811, 3290, 3981, 2804, 4652, 4968, 2911, 259, 217, 2599, 3426, 978, 75, 3279, 749, 3571, 3352, 2881, 2455, 2871, 3369, 1112, 568, 2817, 1590, 1774, 2367, 2136, 3649, 2874, 2956, 3667, 3791, 2368, 2687, 802, 3360, 3298, 3505, 4848, 611, 3765, 356, 2785, 2373, 3913, 2552, 1271, 852, 101, 2684, 3308, 2052, 1360, 709, 3288, 899, 3728, 12, 2905, 4956, 3376, 944, 4769, 2215, 2835, 2393, 3510, 4053, 4036, 1642, 2132, 59, 2201, 3936, 297, 2851, 3574, 3988, 4377, 2810, 3797, 512, 726, 4193, 13, 4247, 4210, 3962, 3533, 912, 2488, 2679, 1162, 2526, 3655, 3135, 1307]
Enter the relevant docs no 973 259 217 978 75 749 568 802 611 852
Releavnt docs are :- {802, 259, 611, 75, 973, 749, 978, 852, 568, 217}
Non Relevant docs are :- {512, 2817, 2052, 12, 13, 2835, 3352, 4377, 1307, 3360, 2851, 2599, 3369, 297, 4652, 3376, 1590, 2871, 2874, 59, 3135, 2367, 3649, 2368, 2881, 2373, 3655, 3913, 1360, 3667, 2132, 1112, 2136, 2905, 2393, 4956, 2911, 3936, 4193, 3426, 356, 101, 4968, 1642, 4210, 2679, 3962, 2684, 2687, 899, 1162, 2956, 3981, 3728, 912, 3988, 4247, 2455, 2201, 4769, 2215, 944, 3505, 3765, 3510, 2488, 4036, 709, 3533, 3279, 3791, 3797, 726, 4053, 3288, 3290, 2526, 2785, 3298, 3305, 3308, 1774, 4848, 3571, 2804, 3574, 1271, 2552, 2810, 2811}

 Iteration 1
Retrieved docs are :- [802, 217, 749, 568, 75, 259, 158, 517, 852, 611, 179, 745, 885, 978, 4591, 4576, 166, 909, 321, 203, 349, 956, 4781, 4806, 310, 768, 1104, 47, 3011, 691, 2923, 973, 1764, 882, 688, 2217, 4660, 709, 944, 746, 3206, 4477, 4568, 886, 44, 151, 4930, 3660, 2227, 4751, 4654, 4215, 4808, 580, 815, 595, 674, 4256, 1371, 3495, 3683, 372, 396, 476, 2279, 4180, 4048, 2688, 2479, 1017, 2656, 1045, 4957, 4115, 2312, 3989, 2272, 2132, 4659, 3170, 2201, 928, 98, 4983, 2011, 804, 204, 441, 333, 4067, 29, 2127, 2390, 57, 3933, 4992, 2140, 2205, 191, 216]
Enter the relevant docs no 191 216 57 29 804 204 441 333 98 44
Releavnt docs are :- {259, 75, 204, 973, 333, 978, 852, 216, 217, 29, 57, 802, 611, 98, 804, 44, 749, 568, 441, 191}
Non Relevant docs are :- {512, 2052, 517, 12, 13, 4115, 1045, 2599, 4652, 4654, 47, 4659, 4660, 1590, 568, 59, 3135, 3649, 580, 3655, 75, 3660, 2127, 1104, 3667, 2132, 595, 4180, 1112, 2136, 2140, 2656, 4193, 3170, 611, 3683, 101, 1642, 4210, 2679, 4215, 2684, 2687, 2688, 3206, 1162, 4751, 3728, 4247, 151, 2201, 2205, 158, 4256, 4769, 674, 166, 2215, 2217, 4781, 688, 179, 691, 3765, 2227, 709, 4806, 4808, 203, 3279, 3791, 3797, 726, 3288, 217, 3290, 2272, 2785, 3298, 1764, 2279, 3305, 745, 746, 3308, 749, 1774, 4848, 2804, 1271, 2810, 2811, 768, 2817, 259, 2312, 2835, 3352, 4377, 1307, 3360, 802, 2851, 3369, 297, 815, 3376, 310, 2871, 2874, 2367, 2368, 2881, 321, 4930, 2373, 3913, 1360, 852, 2390, 2905, 2393, 1371, 4956, 349, 4957, 2911, 3936, 3933, 3426, 356, 4968, 2923, 882, 372, 885, 886, 4983, 3962, 4477, 4992, 899, 2956, 3981, 396, 909, 912, 3988, 3989, 2455, 928, 3495, 2479, 944, 3505, 3510, 2488, 956, 3011, 4036, 3533, 973, 4048, 978, 4053, 4568, 2011, 476, 2526, 4576, 4067, 4591, 3571, 3574, 2552, 1017}

 Iteration 2
Retrieved docs are :- [802, 217, 749, 568, 158, 517, 852, 4591, 259, 4576, 166, 75, 179, 611, 333, 745, 885, 216, 151, 29, 4806, 1764, 57, 204, 2217, 768, 4660, 203, 4781, 441, 321, 1104, 3011, 44, 2227, 349, 2923, 4930, 98,

3206, 674, 4568, 310, 956, 944, 4808, 804, 4215, 4654, 3660, 815, 4477, 691,
2479, 1371, 4256, 882, 4751, 3683, 2656, 47, 3495, 595, 3170, 4983, 476, 688,
2312, 4115, 2272, 1017, 709, 2201, 978, 245, 746, 4957, 2011, 2701, 3989,
2132, 2279, 1612, 2390, 4941, 4048, 1045, 2688, 909, 4180, 4491, 1542, 3933,
2127, 4659, 19, 4067, 886, 3221, 4646]
Enter the relevant docs no 19 909 886 746 245 978 709 476 595 47
Releavnt docs are :- {259, 709, 75, 204, 973, 333, 909, 978, 19, 852, 595,
216, 217, 476, 29, 57, 802, 611, 98, 804, 746, 44, 749, 47, 245, 886, 568,
441, 191}
Non Relevant docs are :- {512, 2052, 517, 1542, 12, 13, 4115, 1045, 29, 4646,
2599, 4652, 44, 4654, 47, 4659, 4660, 1590, 568, 57, 59, 3135, 3649, 580,
3655, 75, 3660, 1612, 2127, 1104, 3667, 2132, 595, 4180, 1112, 2136, 2140,
2656, 4193, 3170, 611, 3683, 101, 98, 1642, 4210, 2679, 4215, 2684, 2687,
2688, 3206, 1162, 2701, 4751, 3728, 3221, 4247, 151, 2201, 2205, 158, 4256,
4769, 674, 166, 2215, 2217, 4781, 688, 179, 691, 3765, 2227, 709, 4806, 4808,
203, 204, 3279, 3791, 3797, 726, 3288, 217, 3290, 216, 2272, 2785, 3298, 1764,
2279, 3305, 745, 746, 3308, 749, 1774, 4848, 2804, 1271, 2810, 2811, 768,
2817, 259, 2312, 2835, 3352, 4377, 1307, 3360, 802, 2851, 804, 3369, 297, 815,
3376, 310, 2871, 2874, 2367, 2368, 2881, 321, 4930, 2373, 3913, 333, 4941,
1360, 852, 2390, 2905, 2393, 1371, 4956, 349, 4957, 2911, 3936, 3933, 3426,
356, 4968, 2923, 882, 372, 885, 886, 4983, 3962, 4477, 4992, 899, 4491, 2956,
3981, 396, 909, 912, 3988, 3989, 2455, 928, 3495, 2479, 944, 3505, 3510, 2488,
441, 956, 3011, 4036, 3533, 973, 4048, 978, 4053, 4568, 2011, 476, 2526, 4576,
4067, 4591, 3571, 3574, 2552, 1017}

 Iteration 3
Retrieved docs are :- [802, 217, 749, 568, 158, 4591, 4576, 166, 517, 333,
216, 852, 204, 179, 151, 29, 2217, 1764, 4806, 57, 768, 745, 259, 441, 203,
1104, 75, 2227, 4660, 4930, 674, 245, 595, 885, 4781, 2923, 611, 3011, 476,
19, 4568, 3206, 44, 98, 4215, 4808, 2479, 321, 691, 815, 4654, 944, 746, 3660,
709, 47, 1371, 2656, 310, 349, 4477, 4256, 3170, 804, 2201, 2701, 4751, 886,
882, 2272, 2312, 4115, 4983, 1017, 3683, 3495, 956, 2011, 4957, 3989, 2132,
2390, 2127, 618, 4941, 1612, 4491, 2279, 1045, 688, 4383, 2093, 4048, 2688,
304, 1542, 4659, 3933, 4180, 4067]
Enter the relevant docs no 746 304 709 19 688 47 882 595 476 12
Releavnt docs are :- {259, 12, 909, 19, 29, 802, 804, 44, 47, 304, 688, 568,
57, 441, 191, 709, 75, 204, 973, 333, 978, 595, 852, 216, 217, 476, 98, 611,
746, 749, 882, 245, 886}
Non Relevant docs are :- {512, 2052, 517, 1542, 12, 13, 4115, 1045, 29, 4646,
2599, 4652, 44, 4654, 47, 2093, 4659, 4660, 1590, 568, 57, 59, 3135, 3649,
580, 3655, 75, 3660, 1612, 2127, 1104, 3667, 2132, 595, 4180, 1112, 2136,
2140, 2656, 4193, 3170, 611, 3683, 101, 98, 1642, 618, 4210, 2679, 4215, 2684,
2687, 2688, 3206, 1162, 2701, 4751, 3728, 3221, 4247, 151, 2201, 2205, 158,
4256, 4769, 674, 166, 2215, 2217, 4781, 688, 179, 691, 3765, 2227, 709, 4806,
4808, 203, 204, 3279, 3791, 3797, 726, 3288, 217, 3290, 216, 2272, 2785, 3298,
1764, 2279, 3305, 745, 746, 3308, 749, 1774, 4848, 2804, 245, 1271, 2810,
2811, 768, 2817, 259, 2312, 2835, 3352, 4377, 1307, 4383, 3360, 802, 2851,
804, 3369, 297, 815, 3376, 310, 2871, 2874, 2367, 2368, 2881, 321, 4930, 2373,
3913, 333, 4941, 1360, 852, 2390, 2905, 2393, 1371, 4956, 349, 4957, 2911,
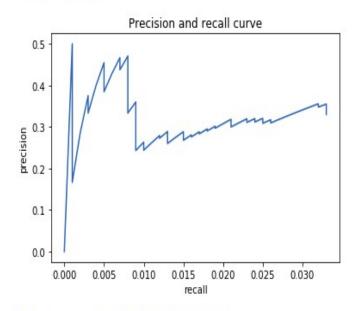3936, 3933, 3426, 356, 4968, 2923, 882, 372, 885, 886, 4983, 3962, 4477, 4992,
899, 4491, 2956, 3981, 396, 909, 912, 3988, 3989, 2455, 928, 3495, 2479, 944,
3505, 3510, 2488, 441, 956, 3011, 4036, 3533, 973, 4048, 978, 4053, 4568,
2011, 476, 2526, 4576, 4067, 4591, 3571, 3574, 2552, 1017

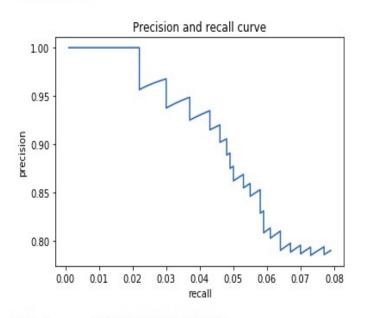**Getting relevant docs process is automated by selecting base folders docs as relevant**

(E) **(1)** **query="Pretty good opinions on biochemistry machines"**
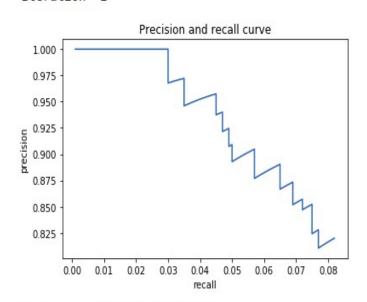**k=100; p=10**

Iteration 0

MAP is :- 0.3379251154374591
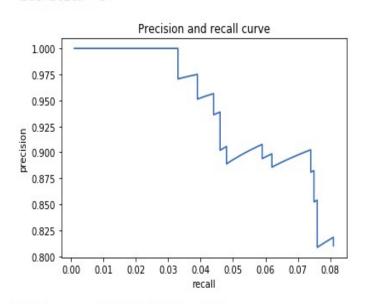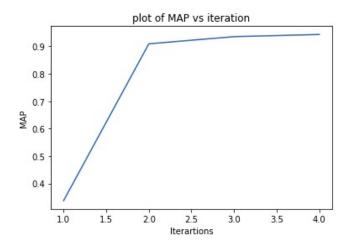
Iteration 1

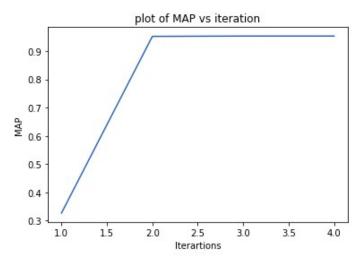MAP is :- 0.9096240804725759

Iteration 2

MAP is :- 0.9357306527639472

Iteration 3
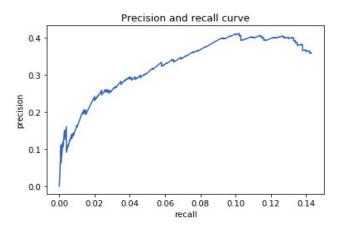

MAP is :- 0.9436640599052544

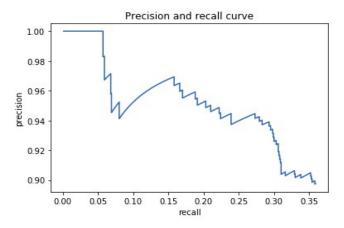**(2)   query=" Scientific tools for preserving rights and body"**
K=400 and p=40

Iteration 0

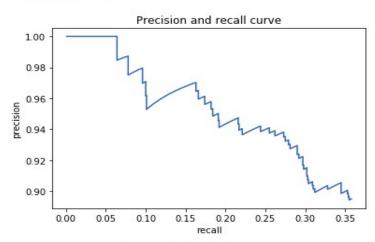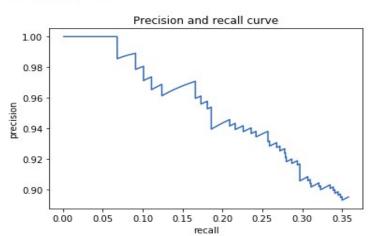MAP is :-   0.3268425272430053

Iteration 1

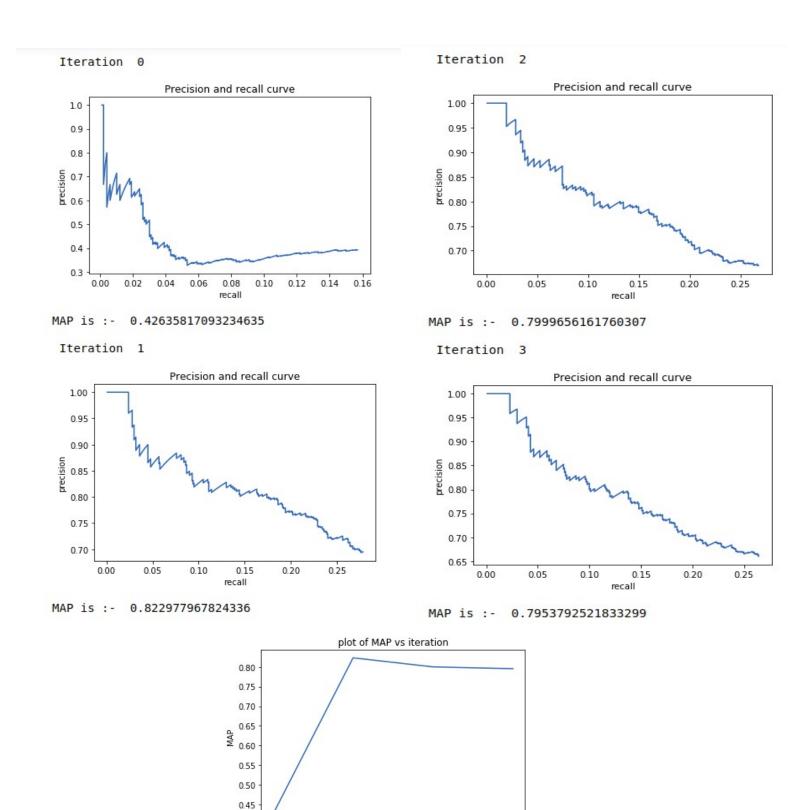MAP is :-   0.9520379584103629

Iteration 2

MAP is :-   0.9532016050898755

Iteration 3

(3)   Query=Frequently asked questions on State-of-the-art visualisation tools
             K=400 and p=40

Iteration 0

MAP is :-   0.42635817093234635

Iteration 1

MAP is :-   0.822977967824336

Iteration 2

MAP is :-   0.7999656161760307
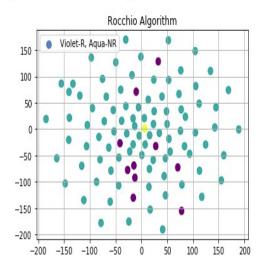
Iteration 3

MAP is :-   0.7953792521833299

(F)
Query vector is getting modified at each iteration by applying Rocchio algorithm
Basically it is shifting towards centroids of Relevant set . Using values of previous query vector and centroid of relevant and non relvant we are shifting our new modified query vector towards relevant docs

Iteration 0
[0. 0. 0. ... 0. 0. 0.]

Rocchio Algorithm

Iteration 2
[0.0033088 0.00375283 0.00315175 ... 0.   0.   0.
]

Rocchio Algorithm

Iteration 1
[0.00176785 0.00202101 0.00180665 ... 0.   0.   0.
]

Rocchio Algorithm

Iteration 3
[0.00481318 0.00542708 0.00463024 ... 0.   0.   0.
]

Rocchio Algorithm