# ANALYSIS FILE

## Pre-Processing Steps:

1) Have used porter stemmer to perform stemming of the documents and the query.2) Have removed all the punctuations across all the docs and replaced it with '' in order to handle cases for numbers like 50,000 .
3) Have performed num2words() to convert the digits to numbers.
4) All the stop words were removed.

## Assumptions:

1) Number of iterations are fixed (I.e. 4) in order to maintain consistency in code while calculating Mean Average Precision (MAP).

2) User knows at the beginning about the number of queries he will be executing.

## Technique Used :

### part 1 : Creating Inverted List

Inverted list is created with each posting of the vocab term contain (doc id ,tf-idf) pair.Inverted List is sorted on tf-idf values.

**Formula used for calculating tf-idf :** tf_idf= (1+math.log(item[1],10)) * df

### Part 2 : Implementing Rochhio Algorithm

**Steps :**

- User is asked to input a set of queries , number of search results required.
- List of ranked documents are returned on the basis of cosine similarity score for the given query.
- User is asked to give input of relevant documents.
- On the basis of partially known relevant and non-relevant docs, updated query vector is calculated.

- And the process repeats.

**Output Sequence**

For each iteration of the user query following results are shown:

- List of searched documents
- Precision Recall values
- Precision Recall Curve
- Tsne plot of relevant , non relevant and query vector
- Average precision of the query for the specific iteration

Once all the iterations for each query finished following results are shown:

- MAP for iteration 1
- Map for iteration 2
- Map for iteration 3

**Note** : Please note that I have hard coded number of iterations(feedback) for each query is set to 3 in order to maintain consistency.

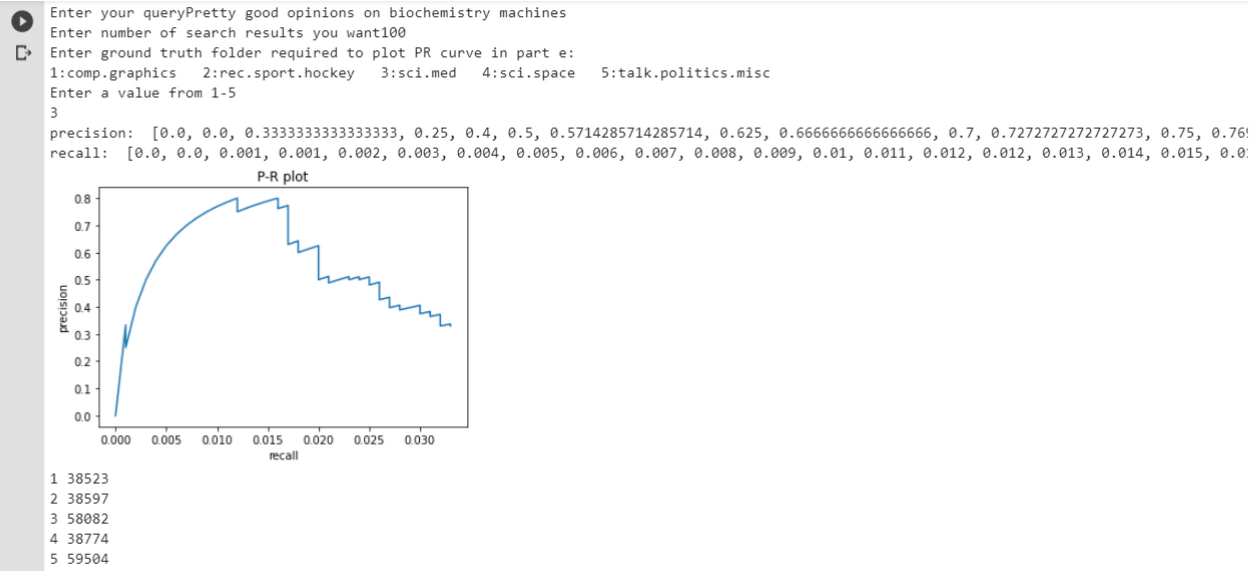
**Inferences Drawn :**

- With each iteration query tends to move towards relevant document  and away from the non relevant documents.
- Cosine similarity works better with tf-idf values.

**Time Taken for each iteration (feedback) for the query** : 5-6 min (max)

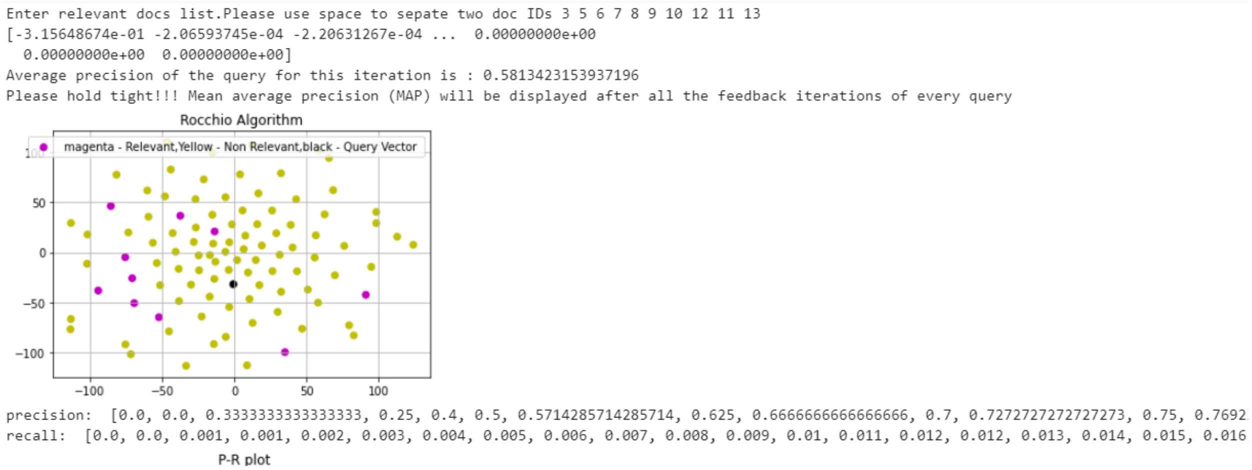
## Sample Output for the query set given in the problem description

## Query 1 iteration 1



```
Enter your queryPretty good opinions on biochemistry machines
Enter number of search results you want100
Enter ground truth folder required to plot PR curve in part e:
1:comp.graphics   2:rec.sport.hockey   3:sci.med   4:sci.space   5:talk.politics.misc
Enter a value from 1-5
3
precision:  [0.0, 0.0, 0.3333333333333333, 0.25, 0.4, 0.5, 0.5714285714285714, 0.625, 0.6666666666666666, 0.7, 0.7272727272727273, 0.75, 0.76
recall:  [0.0, 0.0, 0.001, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.011, 0.012, 0.012, 0.013, 0.014, 0.015, 0.0
```

```
1 38523
2 38597
3 58082
4 38774
5 59504
```

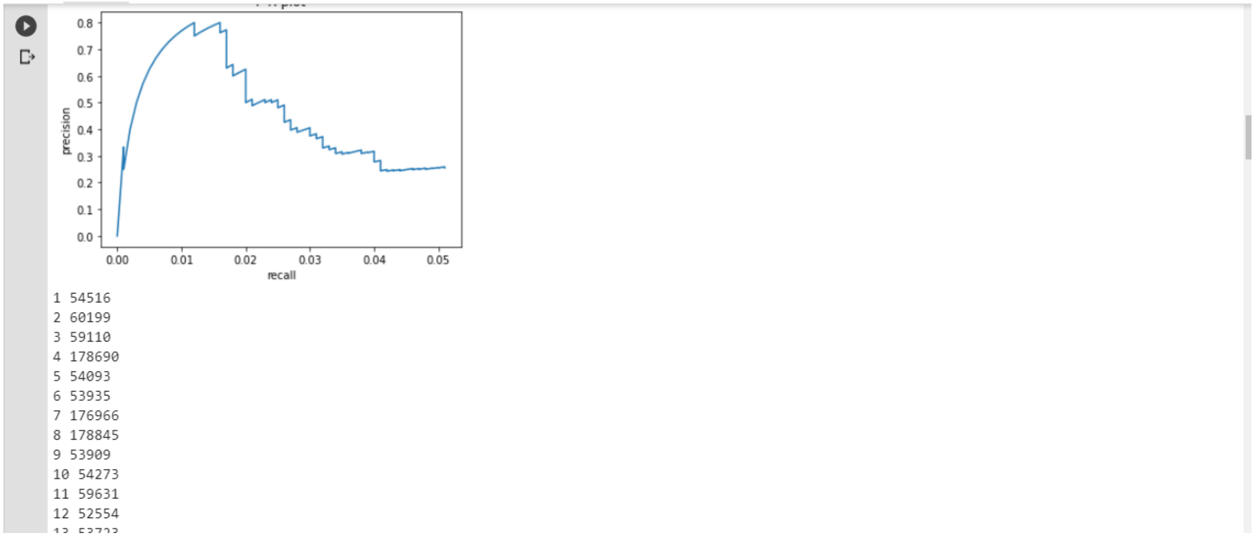


```
Enter relevant docs list.Please use space to sepate two doc IDs 3 5 6 7 8 9 10 12 11 13
[-3.15648674e-01 -2.06593745e-04 -2.20631267e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.5813423153937196
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

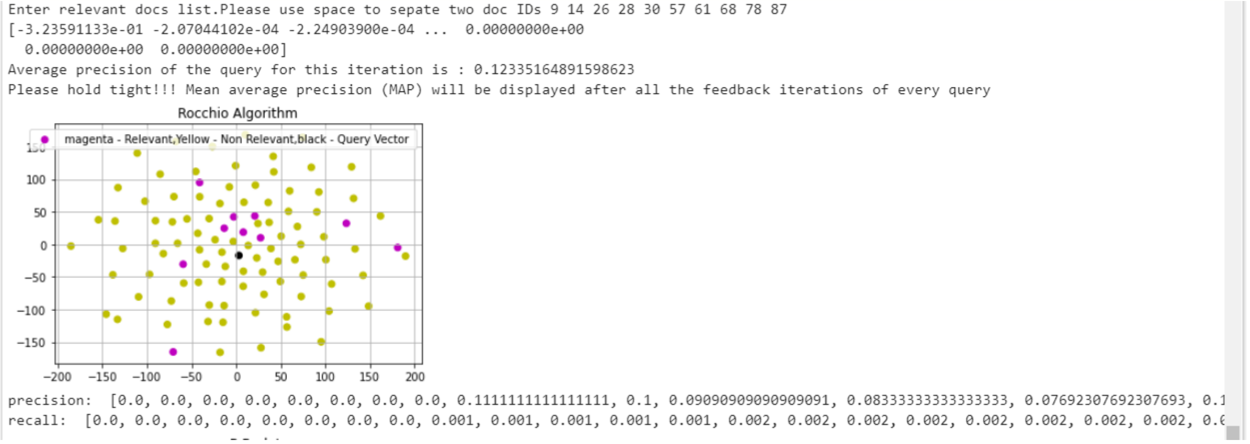
Rocchio Algorithm
magenta - Relevant,Yellow - Non Relevant,black - Query Vector
100
50
0
−50
−100
−100
−50
0
50
100

```
precision:  [0.0, 0.0, 0.3333333333333333, 0.25, 0.4, 0.5, 0.5714285714285714, 0.625, 0.6666666666666666, 0.7, 0.7272727272727273, 0.75, 0.7692
recall:  [0.0, 0.0, 0.001, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.011, 0.012, 0.012, 0.013, 0.014, 0.015, 0.016
```

## Query 1 iteration 2



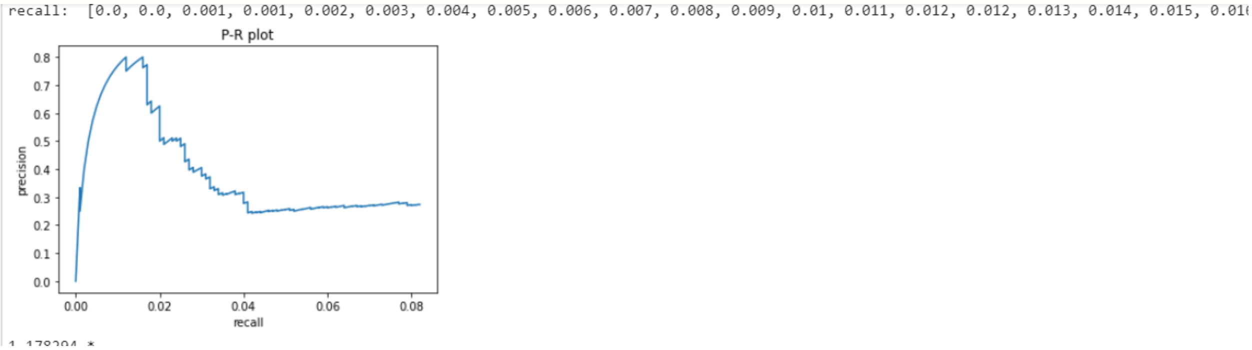
```
1 54516
2 60199
3 59110
4 178690
5 54093
6 53935
7 176966
8 178845
9 53909
10 54273
11 59631
12 52554
13 53723
```



```
Enter relevant docs list.Please use space to sepate two doc IDs 9 14 26 28 30 57 61 68 78 87
[-3.23591133e-01 -2.07044102e-04 -2.24903900e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.12335164891598623
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

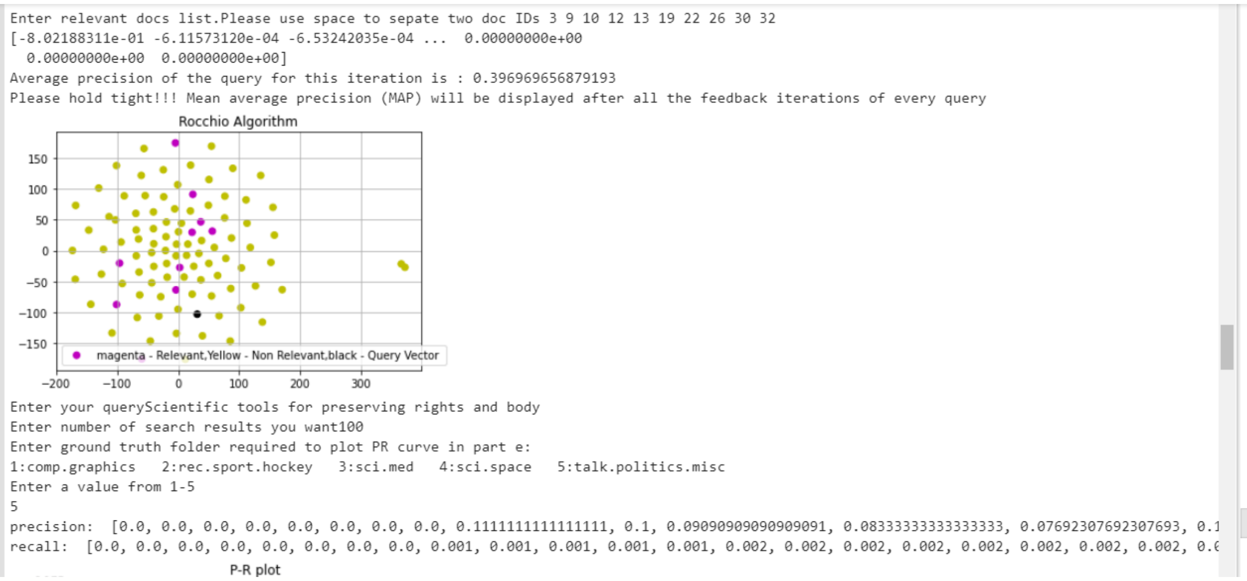
```
precision:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1111111111111111, 0.1, 0.09090909090909091, 0.08333333333333333, 0.07692307692307693, 0.1
recall:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.0
```

## Query 1 iteration 3



```
recall:  [0.0, 0.0, 0.001, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01, 0.011, 0.012, 0.012, 0.013, 0.014, 0.015, 0.01
```

```
1 178204 *
```

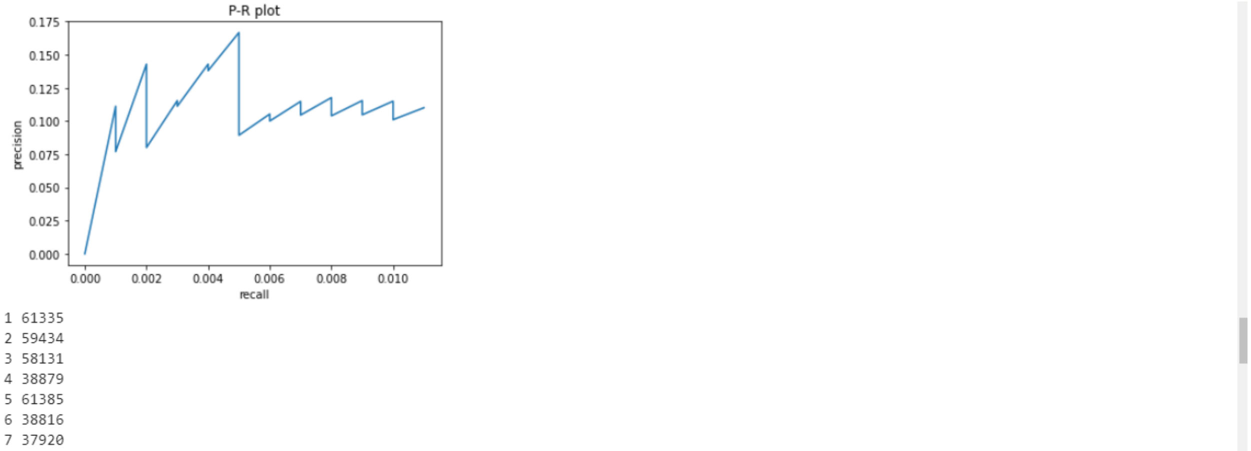


```
Enter relevant docs list.Please use space to sepate two doc IDs 3 9 10 12 13 19 22 26 30 32
[-8.02188311e-01 -6.11573120e-04 -6.53242035e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.396969656879193
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

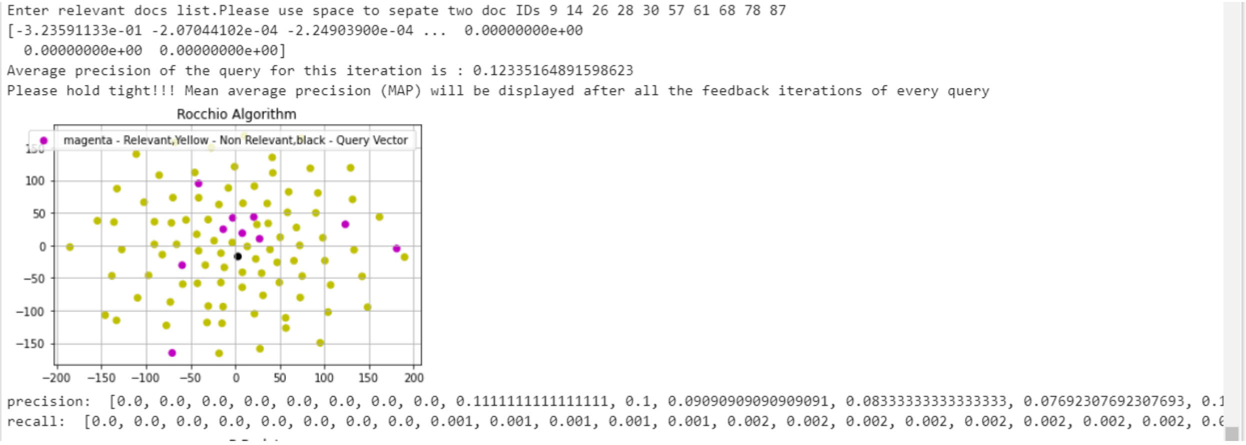
Rocchio Algorithm
150
100
50
0
−50
−100
−150
−200
−100
0
100
200
300
● magenta - Relevant,Yellow - Non Relevant,black - Query Vector

```
Enter your queryScientific tools for preserving rights and body
Enter number of search results you want100
Enter ground truth folder required to plot PR curve in part e:
1:comp.graphics   2:rec.sport.hockey   3:sci.med   4:sci.space   5:talk.politics.misc
Enter a value from 1-5
5
precision:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1111111111111111, 0.1, 0.09090909090909091, 0.08333333333333333, 0.07692307692307693, 0.1
recall:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.0
```

## Query 2 iteration 1



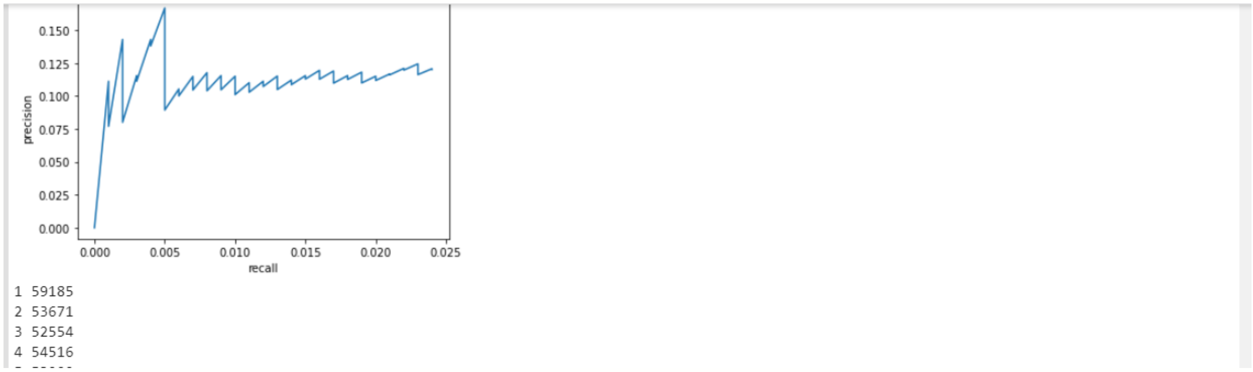
```
1 61335
2 59434
3 58131
4 38879
5 61385
6 38816
7 37920
```



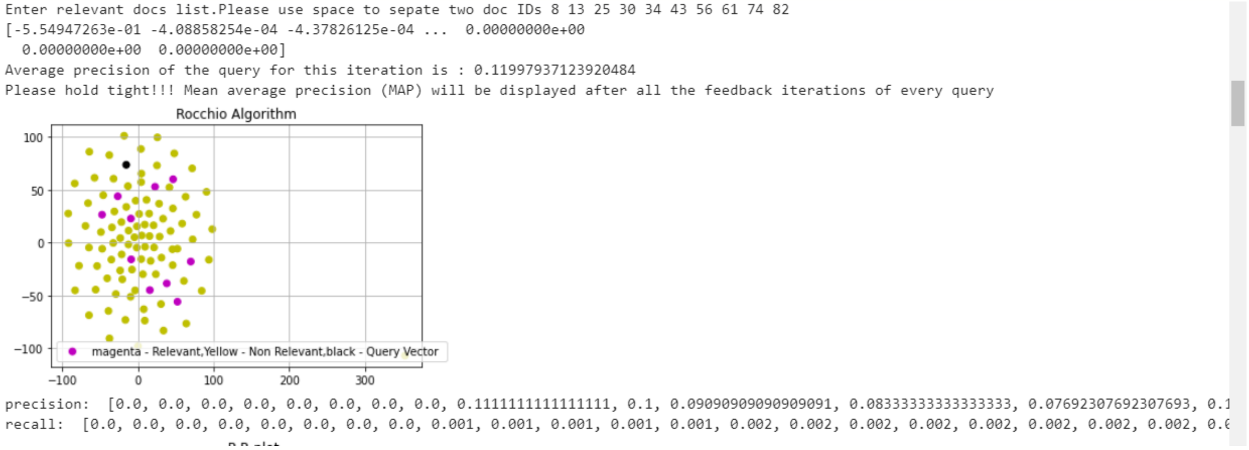
```
Enter relevant docs list.Please use space to sepate two doc IDs 9 14 26 28 30 57 61 68 78 87
[-3.23591133e-01 -2.07044102e-04 -2.24903900e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.12335164891598623
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

```
precision:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1111111111111111, 0.1, 0.09090909090909091, 0.08333333333333333, 0.07692307692307693, 0.1
recall:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.0
```

## Query 2 iteration 2



```
1 59185
2 53671
3 52554
4 54516
```

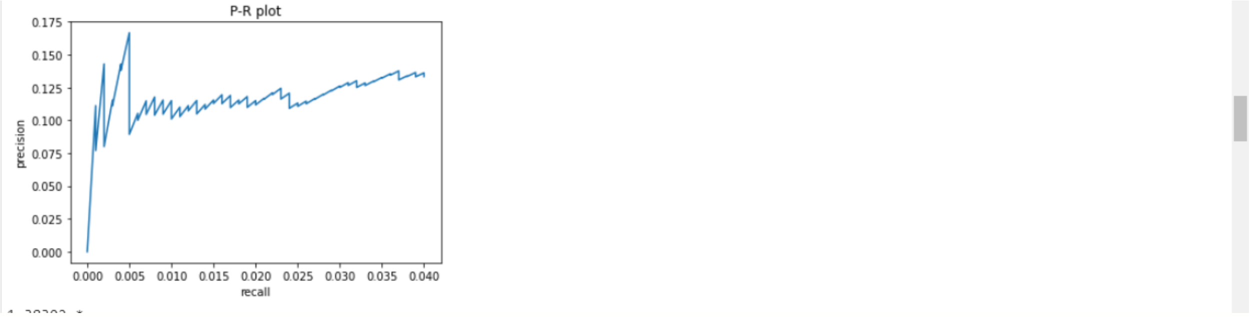


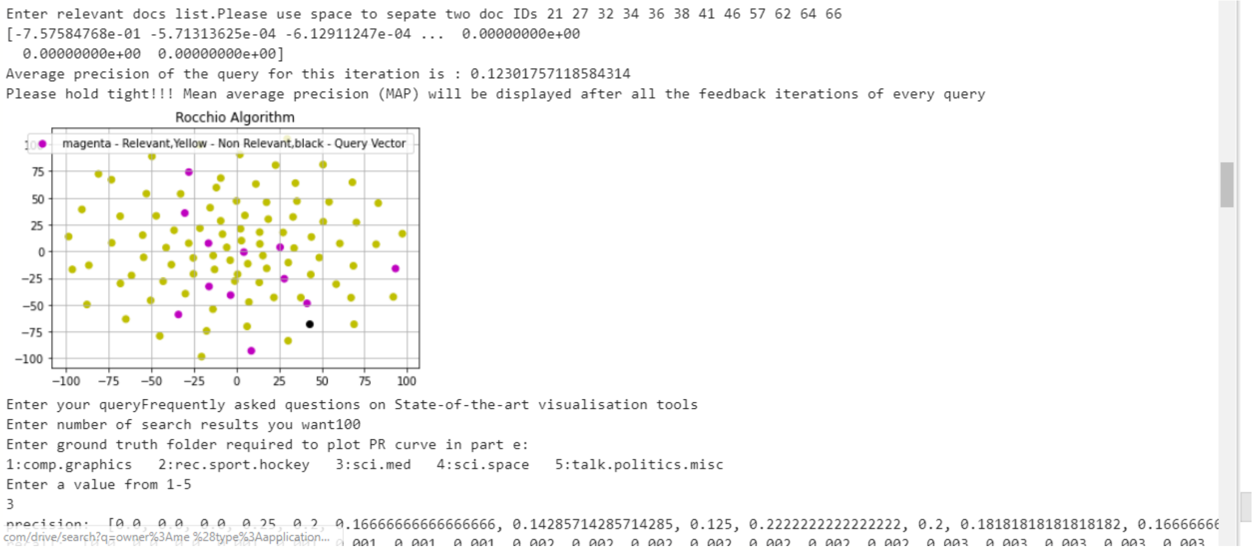
```
Enter relevant docs list.Please use space to sepate two doc IDs 8 13 25 30 34 43 56 61 74 82
[-5.54947263e-01 -4.08858254e-04 -4.37826125e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.11997937123920484
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

Rocchio Algorithm
100
50
0
−50
−100
● magenta - Relevant,Yellow - Non Relevant,black - Query Vector
−100
0
100
200
300

```
precision:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1111111111111111, 0.1, 0.09090909090909091, 0.08333333333333333, 0.07692307692307693, 0.1
recall:  [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.0
```

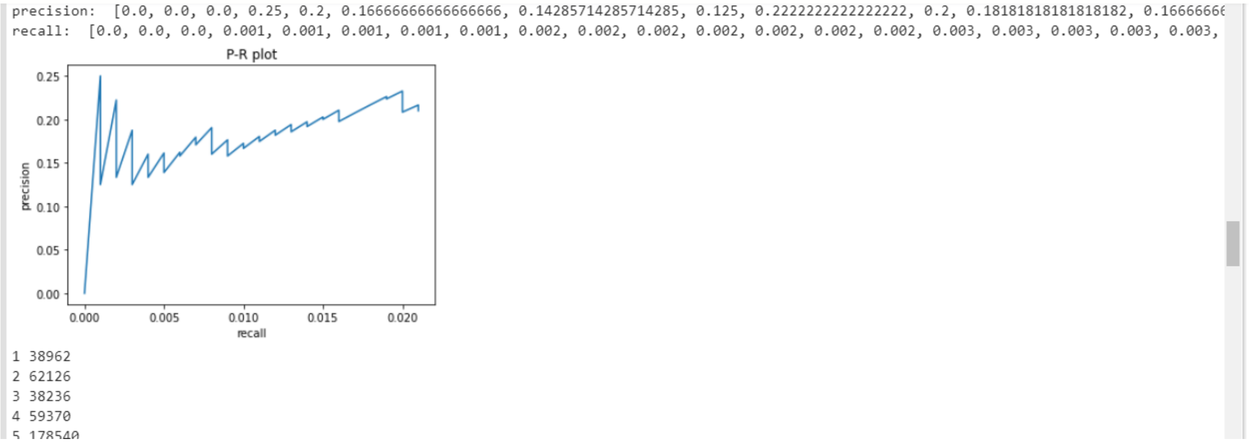
## Query 2 iteration 3



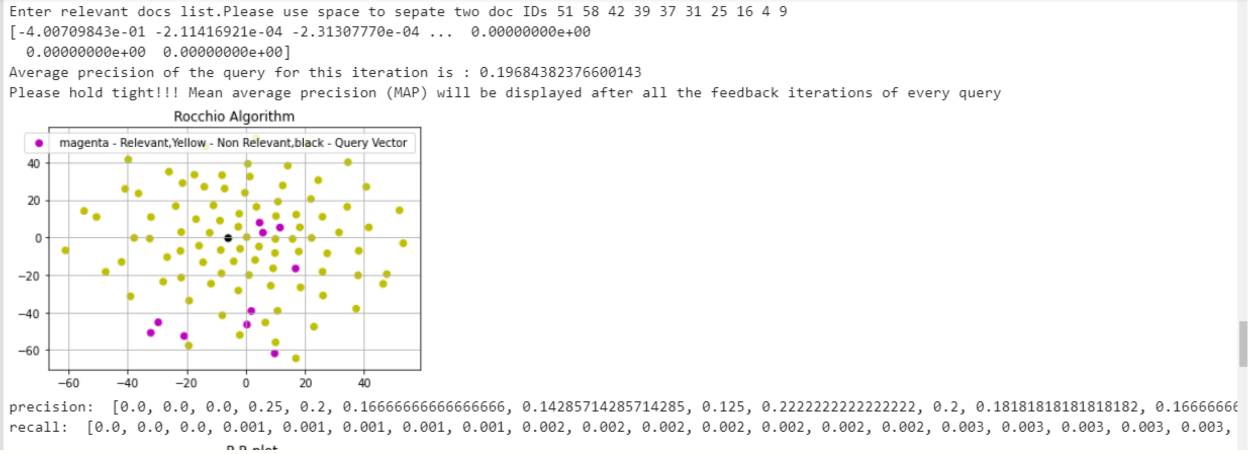

```
Enter relevant docs list.Please use space to sepate two doc IDs 21 27 32 34 36 38 41 46 57 62 64 66
[-7.57584768e-01 -5.71313625e-04 -6.12911247e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.12301757118584314
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

```
Enter your queryFrequently asked questions on State-of-the-art visualisation tools
Enter number of search results you want100
Enter ground truth folder required to plot PR curve in part e:
1:comp.graphics   2:rec.sport.hockey   3:sci.med   4:sci.space   5:talk.politics.misc
Enter a value from 1-5
3
precision:  [0.0, 0.0, 0.0, 0.25, 0.2, 0.16666666666666666, 0.14285714285714285, 0.125, 0.2222222222222222, 0.2, 0.18181818181818182, 0.1666666€
com/drive/search?q=owner%3Ame %28type%3Aapplication...
```

## Query 3 iteration 1


```
precision:  [0.0, 0.0, 0.0, 0.25, 0.2, 0.16666666666666666, 0.14285714285714285, 0.125, 0.2222222222222222, 0.2, 0.18181818181818182, 0.1666666€
recall:  [0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.003, 0.003, 0.003, 0.003, 0.003,
```

```
1 38962
2 62126
3 38236
4 59370
5 178540
```

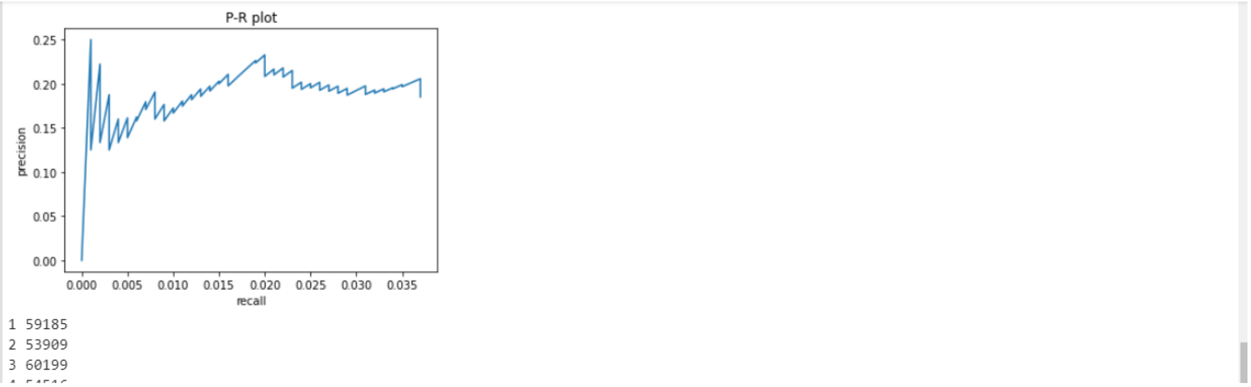


```
Enter relevant docs list.Please use space to sepate two doc IDs 51 58 42 39 37 31 25 16 4 9
[-4.00709843e-01 -2.11416921e-04 -2.31307770e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.19684382376600143
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

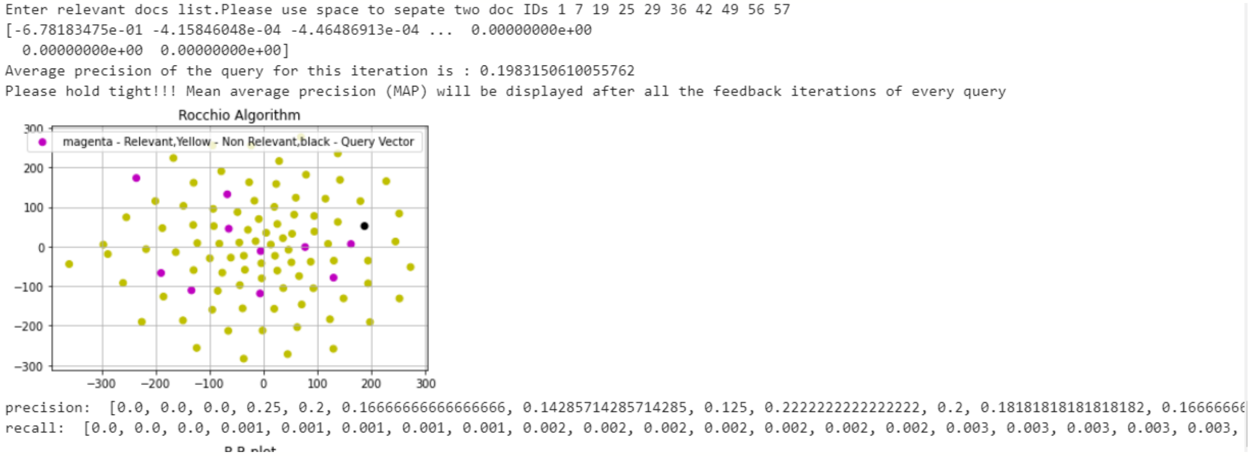
Rocchio Algorithm
magenta - Relevant,Yellow - Non Relevant,black - Query Vector
40
20
0
−20
−40
−60
−60
−40
−20
0
20
40

```
precision:  [0.0, 0.0, 0.0, 0.25, 0.2, 0.16666666666666666, 0.14285714285714285, 0.125, 0.2222222222222222, 0.2, 0.18181818181818182, 0.1666666
recall:  [0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.003, 0.003, 0.003, 0.003, 0.003,
```

## Query 3 iteration 2



```
1 59185
2 53909
3 60199
4 54516
```

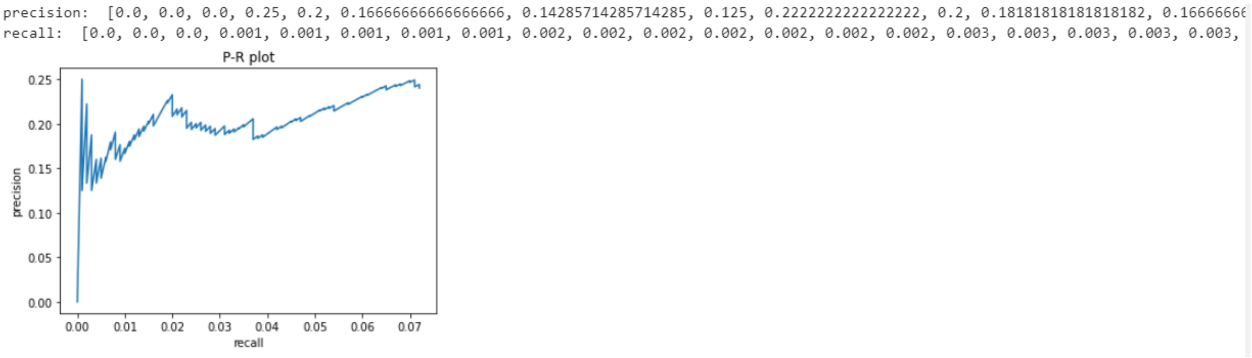


```
Enter relevant docs list.Please use space to sepate two doc IDs 1 7 19 25 29 36 42 49 56 57
[-6.78183475e-01 -4.15846048e-04 -4.46486913e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.1983150610055762
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

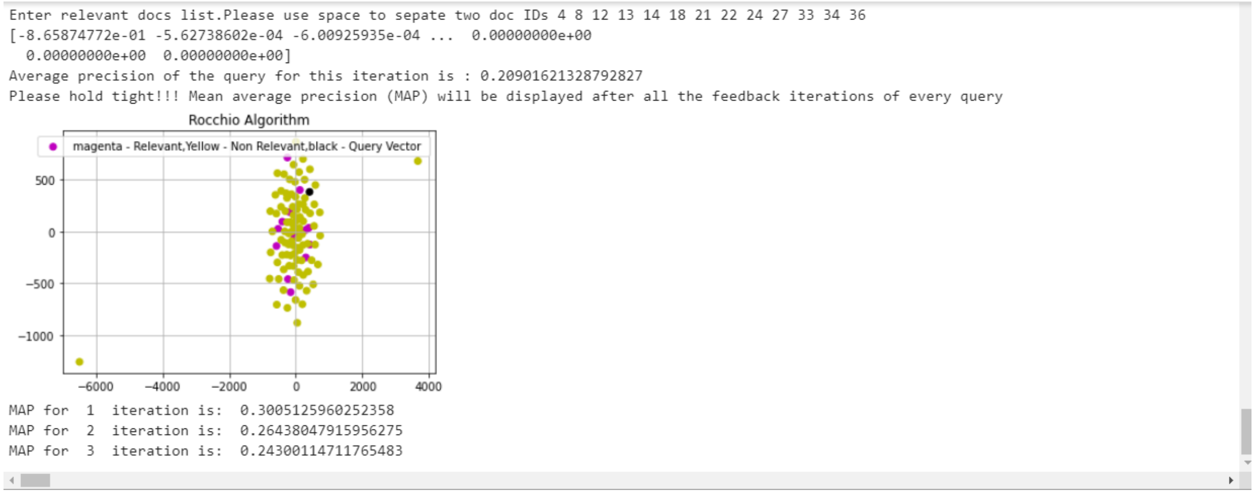
Rocchio Algorithm
magenta - Relevant,Yellow - Non Relevant,black - Query Vector
300
200
100
0
−100
−200
−300
−300
−200
−100
0
100
200
300

```
precision:  [0.0, 0.0, 0.0, 0.25, 0.2, 0.16666666666666666, 0.14285714285714285, 0.125, 0.2222222222222222, 0.2, 0.18181818181818182, 0.1666666
recall:  [0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.003, 0.003, 0.003, 0.003, 0.003,
```

## Query 3 iteration 3



```
precision:  [0.0, 0.0, 0.0, 0.25, 0.2, 0.16666666666666666, 0.14285714285714285, 0.125, 0.2222222222222222, 0.2, 0.18181818181818182, 0.1666666
recall:  [0.0, 0.0, 0.0, 0.001, 0.001, 0.001, 0.001, 0.001, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.002, 0.003, 0.003, 0.003, 0.003, 0.003,
```

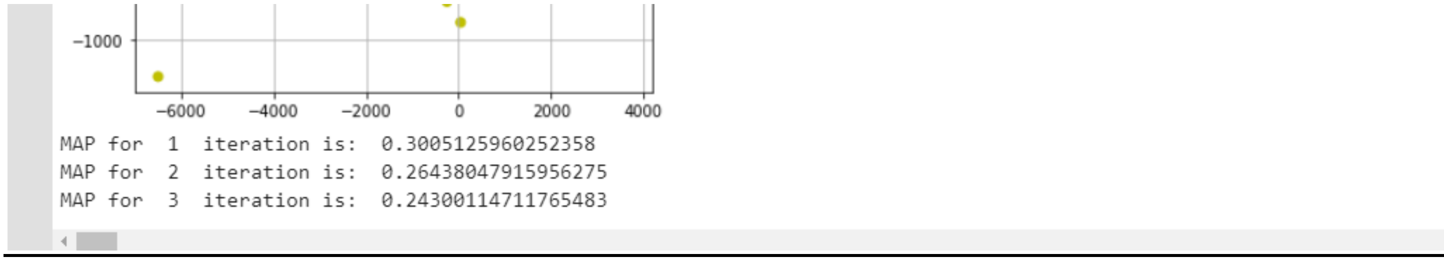


```
Enter relevant docs list.Please use space to sepate two doc IDs 4 8 12 13 14 18 21 22 24 27 33 34 36
[-8.65874772e-01 -5.62738602e-04 -6.00925935e-04 ...  0.00000000e+00
  0.00000000e+00  0.00000000e+00]
Average precision of the query for this iteration is : 0.20901621328792827
Please hold tight!!! Mean average precision (MAP) will be displayed after all the feedback iterations of every query
```

Rocchio Algorithm
● magenta - Relevant,Yellow - Non Relevant,black - Query Vector
500
0
−500
−1000
−6000
−4000
−2000
0
2000
4000

```
MAP for  1  iteration is:  0.3005125960252358
MAP for  2  iteration is:  0.26438047915956275
MAP for  3  iteration is:  0.24300114711765483
```

## MAP



```
MAP for  1  iteration is:  0.3005125960252358
MAP for  2  iteration is:  0.26438047915956275
MAP for  3  iteration is:  0.24300114711765483
```