Assignment 3

Ismael Lopez
Relevance Feedback - Report

Query 1:

Query: " KENNEDY ADMINISTRATION PRESSURE ON NGO DINH DIEM TO STOP SUPPRESSING THE BUDDHISTS . "

Query	Relevant Docs	Exact		Rocchio 1		Rocchio 2		Rocchio 3		Rocchio 4		Rocchio 5
	268	419	N	304	Y	326	Y	326	Y	326	Y	326
	288	268	Y	268	Y	304	Y	304	Y	304	Y	304
	304	406	N	323	Y	308	Y	308	Y	308	Y	308
1	323	288	Y	288	Y	323	Y	323	Y	323	Y	323
	326	304	Y	308	Y	268	Y	268	Y	268	Y	268
	334	323	Y	257	N	288	Y	288	Y	288	Y	288
	308	376	N	326	Y	334	Y	334	Y	334	Y	334
	Precision			0.86		1.00		1.00		1.00		1.00
	Recall			0.86		1.00		1.00		1.00		1.00
	MAP			0.86		0.93		0.95		0.96		0.97

MAP average= (.86 + .93 + .95 + .96 + .97) / 5 = .934

We see that after 2 rounds of feedback, we get all relevant documents.

```
Continue with new query (y/n): y

Top 7 result(s) for the modified query are:

Doc id:  Doc Name:     Score:
326      TEXT 418.txt, 0.3590590630892481
304      TEXT 390.txt, 0.32039579963645115
308      TEXT 396.txt, 0.3200650721633511
323      TEXT 414.txt, 0.30711972374787005
268      TEXT 334.txt, 0.3064681624664226
288      TEXT 363.txt, 0.28264509177177655
334      TEXT 434.txt, 0.18397955750203232

Results found in 0.029361174441874027 seconds.

=== Rocchio Algorithm ===

Iteration: 3
Enter relevant document ids separated by space: 326 304 308 323 268 288 334
Enter non-relevant document ids separated by space:
New query computed in 0.08878847863525152 seconds.
New query terms with weights:
dict_items([('kennedy', 1.3828139380107534), ('administration', 1.347586766422
2134), ('pressure', 1.2283332461979937), ('ngo', 3.2538818002824894), ('dinh',
 3.0395960859967754), ('diem', 9.300326817071735), ('stop', 1.2020043074756444
), ('buddhists', 6.475730549898845), ('december', 0.10714285714285714), ('1960
', 0.21428571428571427), ('help', 0.5357142857142857), ('strike', 0.1071428571
4285714), ('force', 0.9642857142857142), ('attempt', 0.21428571428571427), ('s
tudied', 0.10714285714285714), ('president', 1.7142857142857142), ('minister',
 0.10714285714285714), ('emerging', 0.10714285714285714), ('wants', 0.10714285
714285714), ('support', 0.75), ('crisis', 0.9642857142857142), ('france', 0.32
14285714285714), ('leader', 0.3214285714285714), ('keeping', 0.107142857142857
14), ('terms', 0.10714285714285714), ('government', 6.9642857142857135), ('non
e', 0.10714285714285714), ('cost', 0.10714285714285714), ('poured', 0.21428571
428571427), ('1', 0.21428571428571427), ('presumably', 0.10714285714285714), (
'type', 0.3214285714285714), ('except', 0.10714285714285714), ('national', 0.3
214285714285714), ('committed', 0.10714285714285714), ('eventually', 0.1071428
```
After this, the feedback was the same so the modified query was the same.

Query 2:

Query text: " ELECTION OF PARK CHUNG HEE AS PRESIDENT OF SOUTH KOREA . "

Query	Relevant Docs	Exact		Rocchio 1		Rocchio 2		Rocchio 3		Rocchio 4		Rocchio 5
	94	118	Y	118	Y	118	Y	118	Y	118	Y	118
	118	164	Y	94	Y	164	Y	164	Y	164	Y	164
15	128	128	Y	164	Y	128	Y	128	Y	128	Y	128
	164	422	N	128	Y	94	Y	94	Y	94	Y	94
	424	94	Y	419	N	422	N	419	N	422	N	419
	Precision			0.80		0.80		0.80		0.80		0.80
	Recall			0.80		0.80		0.80		0.80		0.80
	MAP			0.80		0.80		0.80		0.80		0.80

MAP = (.8 + .8 + .8 + .8 + .8) / 5 = .8

We see that despite ruling Document 422 and 419 as non-relevant, we could not produce the last relevant document, number 424.

Assignment 3

Ismael Lopez
Relevance Feedback - Report

```
linux1:~/cs429_spring19_ilopez5
Ω >> python3 -i index.py
TF-IDF Index built in 1.961827022023499 seconds.
Query to search: ' ELECTION OF PARK CHUNG HEE AS PRESIDENT OF SOUTH KOREA . '
Number of (top) results: 5

Top 5 result(s) for the query ' ELECTION OF PARK CHUNG HEE AS PRESIDENT OF SOU
TH KOREA . ' are:

Doc id:  Doc Name:       Score:
118      TEXT 160.txt, 0.2272886492673093
164      TEXT 221.txt, 0.20442797069576232
128      TEXT 176.txt, 0.19312253862380233
422      TEXT 562.txt, 0.1777804026025974
94       TEXT 133.txt, 0.14821796208714125

Results found in 0.0005539432168006897 seconds.

=== Rocchio Algorithm ===

Iteration: 1
Enter relevant document ids separated by space: 118 164 128 94
Enter non-relevant document ids separated by space: 422
New query computed in 0.08071468211710453 seconds.
New query terms with weights:
dict_items([('election', 1.0303719117954557), ('park', 5.980212331696804), ('c
hung', 2.7148703630390236), ('hee', 2.5273703630390236), ('president', 0.73962
12547196624), ('south', 2.817855348496393), ('korea', 3.0177123316968046), ('h
elp', 0.375), ('attempt', 0.1875), ('kennedy', 0.1875), ('minister', 0.1875),
('support', 0.5625), ('crisis', 0.5625), ('dec', 0.1875), ('28', 0.1875), ('le
ader', 0.037500000000000006), ('government', 0.975), ('1', 0.1875), ('build',
0.1875), ('call', 0.375), ('decide', 0.1875), ('withdrawal', 0.1875), ('offici
als', 0.1875), ('decision', 0.1875), ('promised', 0.1875), ('control', 0.375),
```

```
Continue with new query (y/n): y                                    [110/1819]

Top 5 result(s) for the modified query are:

Doc id:  Doc Name:       Score:
118      TEXT 160.txt, 0.43087909004037667
94       TEXT 133.txt, 0.42187276098414345
164      TEXT 221.txt, 0.41514982606175216
128      TEXT 176.txt, 0.4138611283944544
419      TEXT 559.txt, 0.10011080890797862

Results found in 0.007404361851513386 seconds.

=== Rocchio Algorithm ===

Iteration: 2
Enter relevant document ids separated by space: 118 94 164 128
Enter non-relevant document ids separated by space: 419
New query computed in 0.09792223665863276 seconds.
New query terms with weights:
dict_items([('election', 1.1803719117954556), ('park', 7.480212331696804), ('c
hung', 2.8648703630390235), ('hee', 2.6773703630390235), ('president', 0.73962
12547196624), ('south', 3.2678553484963926), ('korea', 3.917712331696804), ('h
elp', 0.375), ('attempt', 0.1875), ('kennedy', 0.1875), ('minister', 0.0375000
00000000006), ('support', 0.5625), ('crisis', 0.5625), ('dec', 0.1875), ('28',
0.1875), ('leader', 0.1875), ('government', 1.125), ('1', 0.1875), ('build',
0.1875), ('call', 0.375), ('decide', 0.1875), ('withdrawal', 0.1875), ('offici
als', 0.1875), ('decision', 0.1875), ('promised', 0.1875), ('control', 0.375),
('defense', -0.11249999999999999), ('run', 0.1875), ('major', -0.262499999999
99996), ('market', 0.1875), ('share', 0.1875), ('hand', 0.1875), ('50', 0.1875
```

```
Continue with new query (y/n): y

Top 5 result(s) for the modified query are:

Doc id:  Doc Name:       Score:
118      TEXT 160.txt, 0.4140121064509327
164      TEXT 221.txt, 0.4011102944498336
128      TEXT 176.txt, 0.393932344102288
94       TEXT 133.txt, 0.393171584148509
422      TEXT 562.txt, 0.20394056925277804

Results found in 0.006022266112267971 seconds.

=== Rocchio Algorithm ===

Iteration: 3
Enter relevant document ids separated by space: 118 164 128 94
Enter non-relevant document ids separated by space: 422
New query computed in 0.09970532730221748 seconds.
New query terms with weights:
dict_items([('election', 1.0303719117954557), ('park', 5.980212331696804), ('c
hung', 2.7148703630390236), ('hee', 2.5273703630390236), ('president', 0.73962
12547196624), ('south', 2.817855348496393), ('korea', 3.0177123316968046), ('h
elp', 0.375), ('attempt', 0.1875), ('kennedy', 0.1875), ('minister', 0.1875),
('support', 0.5625), ('crisis', 0.5625), ('dec', 0.1875), ('28', 0.1875), ('le
ader', 0.037500000000000006), ('government', 0.975), ('1', 0.1875), ('build',
0.1875), ('call', 0.375), ('decide', 0.1875), ('withdrawal', 0.1875), ('offici
als', 0.1875), ('decision', 0.1875), ('promised', 0.1875), ('control', 0.375),
('defense', 0.1875), ('run', 0.1875), ('major', 0.1875), ('market', 0.1875),
```

```
Continue with new query (y/n): y

Top 5 result(s) for the modified query are:

Doc id:  Doc Name:       Score:
118      TEXT 160.txt, 0.43087909004037667
94       TEXT 133.txt, 0.42187276098414345
164      TEXT 221.txt, 0.41514982606175216
128      TEXT 176.txt, 0.4138611283944544
419      TEXT 559.txt, 0.10011080890797862

Results found in 0.006349768489599228 seconds.

=== Rocchio Algorithm ===

Iteration: 4
Enter relevant document ids separated by space: 118 94 164 128
Enter non-relevant document ids separated by space: 419
New query computed in 0.08498838916420937 seconds.
New query terms with weights:
dict_items([('election', 1.1803719117954556), ('park', 7.480212331696804), ('president', 0.73962
12547196624), ('south', 3.2678553484963926), ('korea', 3.917712331696804), ('h
elp', 0.375), ('attempt', 0.1875), ('kennedy', 0.1875), ('minister', 0.0375000
00000000006), ('support', 0.5625), ('crisis', 0.5625), ('dec', 0.1875), ('28',
0.1875), ('leader', 0.1875), ('government', 1.125), ('1', 0.1875), ('build',
0.1875), ('call', 0.375), ('decide', 0.1875), ('withdrawal', 0.1875), ('offici
als', 0.1875), ('decision', 0.1875), ('promised', 0.1875), ('control', 0.375),
```

```
Continue with new query (y/n): y

Top 5 result(s) for the modified query are:

Doc id:  Doc Name:       Score:
118      TEXT 160.txt, 0.4140121064509327
164      TEXT 221.txt, 0.4011102944498336
128      TEXT 176.txt, 0.393932344102288
94       TEXT 133.txt, 0.393171584148509
422      TEXT 562.txt, 0.20394056925277804

Results found in 0.006230795755982399 seconds.

=== Rocchio Algorithm ===

Iteration: 5
Enter relevant document ids separated by space: 118 164 128 94
Enter non-relevant document ids separated by space: 422
New query computed in 0.0954679325222969 seconds.
New query terms with weights:
dict_items([('election', 1.0303719117954557), ('park', 5.980212331696804), ('c
hung', 2.7148703630390236), ('hee', 2.5273703630390236), ('president', 0.73962
12547196624), ('south', 2.817855348496393), ('korea', 3.0177123316968046), ('h
elp', 0.375), ('attempt', 0.1875), ('kennedy', 0.1875), ('minister', 0.1875),
('support', 0.5625), ('crisis', 0.5625), ('dec', 0.1875), ('28', 0.1875), ('le
ader', 0.037500000000000006), ('government', 0.975), ('1', 0.1875), ('build',
0.1875), ('call', 0.375), ('decide', 0.1875), ('withdrawal', 0.1875), ('offici
als', 0.1875), ('decision', 0.1875), ('promised', 0.1875), ('control', 0.375),
('defense', 0.1875), ('run', 0.1875), ('major', 0.1875), ('market', 0.1875),
```

Query 3:

Query Text: "AGREEMENT BY THE UNITED ARAB REPUBLIC AND SAUDI ARABIA TO WITHDRAW THEIR FORCES FROM YEMEN, WHICH INVOLVES OBSERVERS FROM THE UNITED NATIONS EXPEDITIONARY FORCE BEING SENT TO YEMEN . "

Query	Relevant Docs	Exact		Rocchio 1		Rocchio 2		Rocchio 3		Rocchio 4		Rocchio 5
19	99	159	N	99	Y	99	Y	99	Y	99	Y	99
	100	195	Y	195	Y	267	Y	267	Y	267	Y	267
	195	10	N	267	Y	195	Y	195	Y	195	Y	195
	267	99	Y	344	Y	344	Y	344	Y	344	Y	344
	344	86	N	10	N	100	Y	100	Y	100	Y	100
	Precision			0.80		1.00		1.00		1.00		1.00
	Recall			0.80		1.00		1.00		1.00		1.00
	MAP			0.80		0.90		0.93		0.95		0.96

MAP = (.8 + .9 + .93 + .95 + .96) / 5 = .908

We see that after 2 iterations, we produce all relevant documents.

```
linux1:~/cs429_spring19_ilopez5
Ω >> python3 -i index.py
TF-IDF Index built in 2.5889279721304774 seconds.
Query to search: ' AGREEMENT BY THE UNITED ARAB REPUBLIC AND SAUDI ARABIA TO W
ITHDRAW THEIR FORCES FROM YEMEN, WHICH INVOLVES OBSERVERS FROM THE UNITED NATI
ONS EXPEDITIONARY FORCE BEING SENT TO YEMEN . '
Number of (top) results: 5

Top 5 result(s) for the query ' AGREEMENT BY THE UNITED ARAB REPUBLIC AND SAUD
I ARABIA TO WITHDRAW THEIR FORCES FROM YEMEN, WHICH INVOLVES OBSERVERS FROM TH
E UNITED NATIONS EXPEDITIONARY FORCE BEING SENT TO YEMEN . ' are:

Doc id:  Doc Name:      Score:
159       TEXT 215.txt, 0.12610691330141163
195       TEXT 253.txt, 0.1230117834450849
10        TEXT 027.txt, 0.11965433465268312
99        TEXT 138.txt, 0.08699756524652308
86        TEXT 121.txt, 0.08359998998795438

Results found in 0.0006487304344773293 seconds.

=== Rocchio Algorithm ===

Iteration: 1
Enter relevant document ids separated by space: 195 99
Enter non-relevant document ids separated by space: 159 10 86
New query computed in 0.08125180192291737 seconds.
New query terms with weights:
dict_items([('agreement', 0.92428037604708), ('united', 1.7036497822351608), (
'arab', 1.22421911265538), ('republic', 1.1198614503327873), ('saudi', 5.97524
9108319361), ('arabia', 4.7958914459967685), ('withdraw', 1.9982503803830989),
 ('forces', 1.5410105323642753), ('yemen', 9.450800643489265), ('involves', 2.
1492191126553797), ('observers', 1.4623970150682055), ('nations', 1.0037132760
846008), ('expeditionary', 2.72428037604708), ('force', 2.482857418821107), ('
sent', 1.4554883557328981), ('december', 0.375), ('help', 1.075), ('plan', 0.3
75), ('president', 1.2), ('kennedy', 0.375), ('saw', 0.325), ('support', 0.375
Continue with new query (y/n): y

Top 5 result(s) for the modified query are:

Doc id:  Doc Name:      Score:
99        TEXT 138.txt, 0.5405420450937874
195       TEXT 253.txt, 0.4479966961451682
267       TEXT 333.txt, 0.23047114971551008
344       TEXT 459.txt, 0.20170964437417702
10        TEXT 027.txt, 0.19259748626794082

Results found in 0.009074674919247627 seconds.

=== Rocchio Algorithm ===

Iteration: 2
Enter relevant document ids separated by space: 99 195 267 344
Enter non-relevant document ids separated by space: 10
New query computed in 0.08296993654221296 seconds.
New query terms with weights:
dict_items([('agreement', 1.02428037604708), ('united', 2.003649782235161), ('
arab', 1.71171911265538), ('republic', 1.2823614503327871), ('saudi', 5.087749
108319361), ('arabia', 4.095891445996768), ('withdraw', 1.7607503803830988), (
'forces', 1.0660105323642752), ('yemen', 8.638300643489265), ('involves', 2.14
92191126553797), ('observers', 1.5123970150682056), ('nations', 1.291213276084
601), ('expeditionary', 2.3992803760470798), ('force', 1.7828574188211073), ('
sent', 1.605488355732898), ('december', 0.1875), ('help', 0.5625), ('plan', 0.
375), ('president', 1.5375), ('kennedy', 0.1875), ('saw', 0.1875), ('support',
 0.375), ('supply', 0.1875), ('britain', 0.375), ('28', 0.375), ('leader', 0.3
Continue with new query (y/n): y

Top 5 result(s) for the modified query are:

Doc id:  Doc Name:      Score:
99        TEXT 138.txt, 0.4165984823895335
267       TEXT 333.txt, 0.4009791110905055
195       TEXT 253.txt, 0.3991819354822906
344       TEXT 459.txt, 0.37500226212198323
100       TEXT 140.txt, 0.18630057340407055

Results found in 0.008098633959889412 seconds.

=== Rocchio Algorithm ===

Iteration: 3
Enter relevant document ids separated by space: 99 267 195 344 100
Enter non-relevant document ids separated by space:
New query computed in 0.0876356614753604 seconds.
New query terms with weights:
dict_items([('agreement', 1.02428037604708), ('united', 2.003649782235161), ('
arab', 2.79921911265538), ('republic', 1.3948614503327872), ('saudi', 5.800249
108319361), ('arabia', 4.695891445996768), ('withdraw', 1.8732503803830987), (
'forces', 1.1410105323642752), ('yemen', 8.825800643489265), ('involves', 2.14
92191126553797), ('observers', 1.5123970150682056), ('nations', 1.47871327608$
601), ('expeditionary', 2.3242803760470796), ('force', 1.8578574188211072), ('
sent', 2.055488355732982), ('december', 0.15), ('help', 0.44999999999999996), (
('plan', 0.3), ('president', 1.7999999999999998), ('kennedy', 0.15), ('minist
er', 0.15), ('saw', 0.44999999999999996), ('support', 0.3), ('crisis', 0.15),
```

Feedback/Query same after this point

Pseudo Relevance – Query 15:

Query	Relevant Docs	Exact		Rocchio 1		Rocchio 2		Rocchio 3		Rocchio 4		Rocchio 5
Pseudo-15	94	118	-	164	-	164	-	164	-	164	-	164
	118	164	-	128	-	128	-	128	-	128	-	128
	128	128	-	118	-	118	-	118	-	118	-	118
	164	422	-	422	-	422	-	422	-	422	-	422
	424	94	-	94	-	94	-	94	-	94	-	94
	Precision			0.80		0.80		0.80		0.80		0.80
	Recall			0.80		0.80		0.80		0.80		0.80
	MAP			0.80		0.80		0.80		0.80		0.80

MAP = (.8 * 5) / 5 = .8

Now it seems that there was no inherent change, however, if you observe the scores from Exact Top K vs Rocchio (iteration 1), you will see that there was a change, but it was not enough to score the bottom two documents below the documents below them.

```
TF-IDF Index built in 1.736287915147841 seconds.
Query to search: ' ELECTION OF PARK CHUNG HEE AS PRESIDENT OF SOUTH KOREA . '
Number of (top) results: 5

Top 5 result(s) for the query ' ELECTION OF PARK CHUNG HEE AS PRESIDENT OF SOUTH KOREA . ' are:

Doc id:  Doc Name:      Score:
118       TEXT 160.txt, 0.2272886492673093
164       TEXT 221.txt, 0.20442797069576232
128       TEXT 176.txt, 0.19312253862380233
422       TEXT 562.txt, 0.1777804026025974
94        TEXT 133.txt, 0.14821796208714125

Results found in 0.00046596676111221313 seconds.
```

```
=== Rocchio Algorithm ===

Iteration: 1
Enter relevant document ids separated by space:
Enter non-relevant document ids separated by space:
New query computed in 0.08372097741812468 seconds.
Continue with new query (y/n):

Top 5 result(s) for the modified query are:

Doc id:  Doc Name:      Score:
164       TEXT 221.txt, 0.5231161988876982
128       TEXT 176.txt, 0.49251382831961277
118       TEXT 160.txt, 0.443861046192122
422       TEXT 562.txt, 0.11987604599965061
94        TEXT 133.txt, 0.10509122905467223

Results found in 0.0045808907598257065 seconds.
```
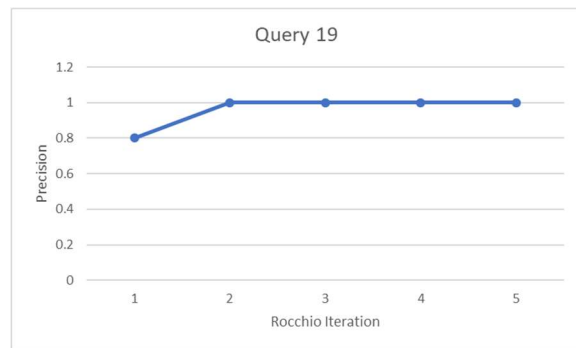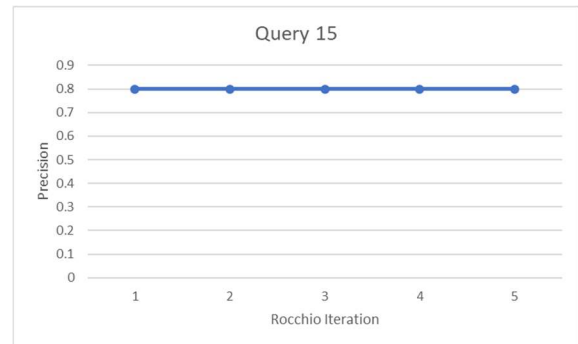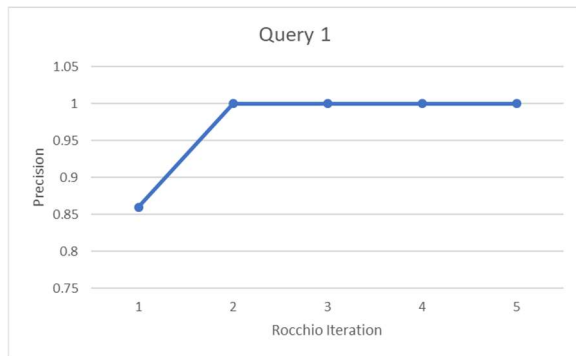
So our Pseudo algorithm of arbitrarily marking the top 3 as *relevant* and the bottom k-3 documents as *nonrelevant* is not as great as we hoped. Consider as well that document 94 is marked as nonrelevant despite being a 'relevant' document from our test bed. So this algorithm will work but not well.
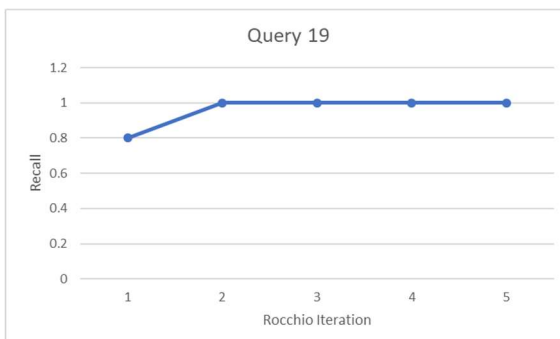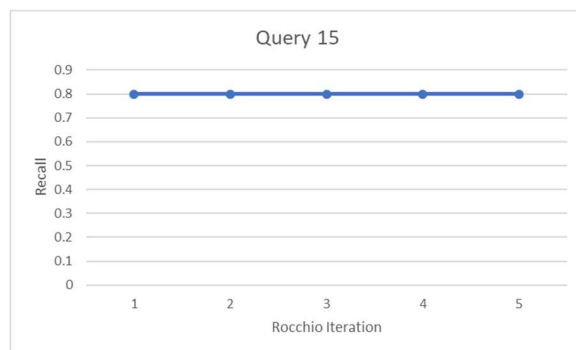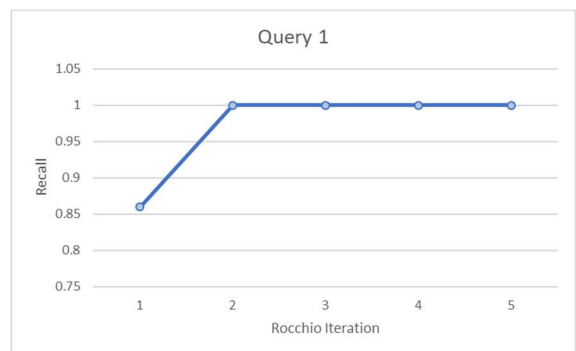
Assignment 3

Ismael Lopez
Relevance Feedback - Report

Precision Graphs:

Recall:

Assignment 3

Ismael Lopez
Relevance Feedback - Report

MAP:

Query 1

Query 15

Query 19