Part A

A sample test run is provided in the output.txt to check the result of Rocchio algorithm described in part A. User have to provide

Part B

For the experimental study section queries used are (Id, query text):-

Query 1 – 68 ['indian', 'fears', 'communist', 'chinese', 'invasion']

Precision[0.25, 0.25, 0.25, 0.25, 0.25]

Recall[0.625, 0.625, 0.625, 0.625, 0.625]

MAP[0.35031897926634764, 0.4083333333333334, 0.5328947368421053, 0.625, 0.4351190476190476]

|  |  |  |
| --- | --- | --- |
| Iteration | Positive docs user feed back | Negative docs user feed back |
| 1 | 70 54 | 553 |
| 2 | 24 25 | 557 |
| 3 | 394 | 557 204 |
| 4 | 96 242 | 550 396 461 |

Query 2- 63 ['president', 'nasser', 'ruling', 'arab', 'union', 'present', 'governing', 'party', 'syria', 'remains', 'control']

Precision[0.3, 0.3, 0.3, 0.3, 0.3]

Recall[0.8181818181818182, 0.8181818181818182, 0.8181818181818182, 0.8181818181818182, 0.8181818181818182]

MAP[0.6112887112887112, 0.6265068265068264, 0.6749916749916749, 0.7521531100478469, 0.5770479520479521]

|  |  |  |
| --- | --- | --- |
| Iteration | Positive docs user feed back | Negative docs user feed back |
| 1 | 398 215 | 60 |
| 2 | 283 268 | 525 459 538 |
| 3 | 163 140 121 | 445 |
| 4 | 104 | 71 |

Query 3- 55 ['suggestion', 'president', 'kennedy', 'nato', 'nuclear', 'missile', 'fleet', 'manned', 'international', 'crews']

Precision[0.26666666666666666, 0.3333333333333333, 0.3333333333333333, 0.3333333333333333, 0.36666666666666664]

Recall[0.6666666666666666, 0.8333333333333334, 0.8333333333333334, 0.8333333333333334, 0.9166666666666666]

MAP[0.5461760461760461, 0.7547822547822548, 0.7855263157894737, 0.7598695286195287, 0.6318602693602694]

|  |  |  |
| --- | --- | --- |
| Iteration | Positive docs user feed back | Negative docs user feed back |
| 1 | 86 126 317 | 399 231 |
| 2 | 17 287 | 181 |
| 3 | 85 71 213 | 88 163 243 |
| 4 | 558 | 365 88 404 |

The above data is also present in output.txt file. The fig1, fig2 and fig3 graphs show the changes in precision, recall and MAP in each iteration for the above queries.

The query weights for terms in each iteration is present in file “Query text and weights.txt” because of the huge size of query after each iteration. The query terms filtering code was initially written based on weights but it does not provide proper results and does not serve the purpose of Rocchio Algorithm.

**Query Drift**

As long as the user’s positive and negative feedback documents are proper, the retrieved documents will be almost similar to the ones provided. But when the user provides a not much relevant document as the positive feedback, the query drifts away from centroid of relevant documents and it can result in decrease in recall in each iteration.

Since in the experimental study only 5 iterations were used there was not much considerable increase in precision and recall. Both increased pretty well in each iteration and since the positive feedbacks were given with care, none of the already retrieved relevant documents were lost in the process. But this can happen in case of wrong feedback or when the relevant documents are not much similar to each other.

**Choosing K value and Query**

The query is chosen randomly from TIME.QUE and match the criteria of having 5 relevant documents in TIME.REL.

If the number of relevant documents is less than 10, number of retrieved documents is kept as 20. For 10 to 15 range, the k value is 30 and for 15 to 20 range the k is made as 40. For all the above cases, the k is kept as twice the size of number of relevant documents.

**Pseudo Feedback System**

The queries used in user feedback system is considered for this method. The query text and weights corresponding to each query in each iteration is specified in “Query text and weights.txt”

The figures in files,

pseudomapfig1.png

pseudomapfig2.png

pseudomapfig3.png

pseudoprefig1.png

pseudoprefig2.png

pseudoprefig3.png

pseudorecfig1.png

pseudorecfig2.png

pseudorecfig3.png

compares the MAP, Precision and Recall between user feedback and pseudo feedback system. The performance of pseudo feedback system is comparatively lower than the user feedback as the system randomly chooses the top 3 documents as positive feedback. The top 3 need not be the relevant document as per query context. So user feedback system is more preferable.