IT458 Assignment 3

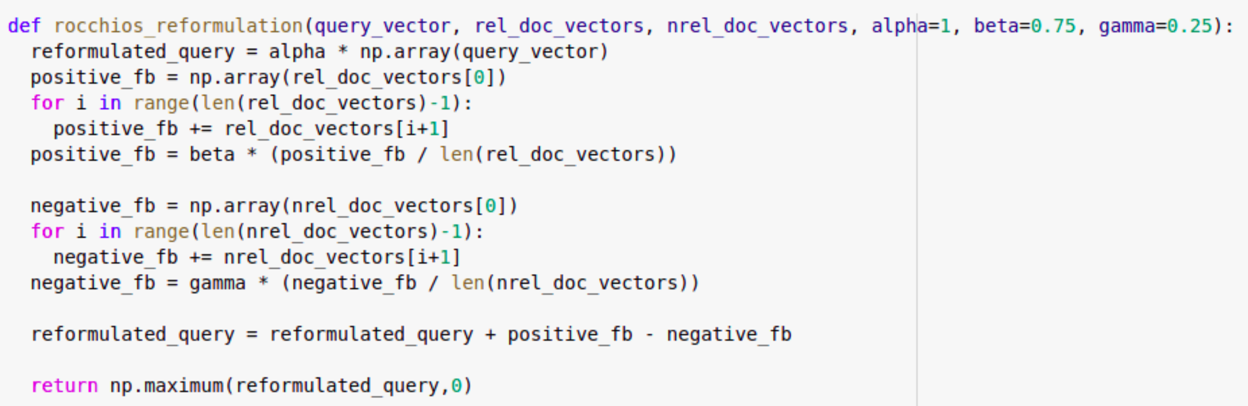
NAME: SUYASH CHINTAWAR
ROLL NO.: 191IT109
TOPIC: ROCCHIO’S ALGORITHM

Note:
1) The colab link has been attached below. After opening the link, if it opens in drive, click on “Open with Google Colaboratory” to view the complete code.

Colab notebook link:
https://colab.research.google.com/drive/14ojEMci33hufMFYXwXKw4sXb0LfDUNPo

Q.Implement Rocchio’s algorithm to incorporate relevance feedback in VSM rankings generated in the previous assignment. Recompute the ranked list using Cosine similarity for each of the reformulated queries to assess the effect of relevance feedback for at least three reformulated queries.

Rocchios algorithm implementation:



```python
def rocchios_reformulation(query_vector, rel_doc_vectors, nrel_doc_vectors, alpha=1, beta=0.75, gamma=0.25):
  reformulated_query = alpha * np.array(query_vector)
  positive_fb = np.array(rel_doc_vectors[0])
  for i in range(len(rel_doc_vectors)-1):
    positive_fb += rel_doc_vectors[i+1]
  positive_fb = beta * (positive_fb / len(rel_doc_vectors))

  negative_fb = np.array(nrel_doc_vectors[0])
  for i in range(len(nrel_doc_vectors)-1):
    negative_fb += nrel_doc_vectors[i+1]
  negative_fb = gamma * (negative_fb / len(nrel_doc_vectors))

  reformulated_query = reformulated_query + positive_fb - negative_fb

  return np.maximum(reformulated_query,0)
```

The algorithm is based on the assumption that most users have a general conception of which documents should be denoted as relevant or non-relevant. Therefore, the user's search query is revised to include an arbitrary percentage of relevant and non-relevant documents as a means of increasing the search engine's recall, and possibly the precision as well. The above snippet shows the rocchio’s algorithm implemented. The values of the three hyper parameters have been set as alpha = 1, beta = 0.75, gamma = 0.25. by default. It can be altered by making changes to the parameter values. If after reformulation, the tf-idf weight becomes less than zero, they are clipped to zero as the minimum value to avoid negative cosine similarities.

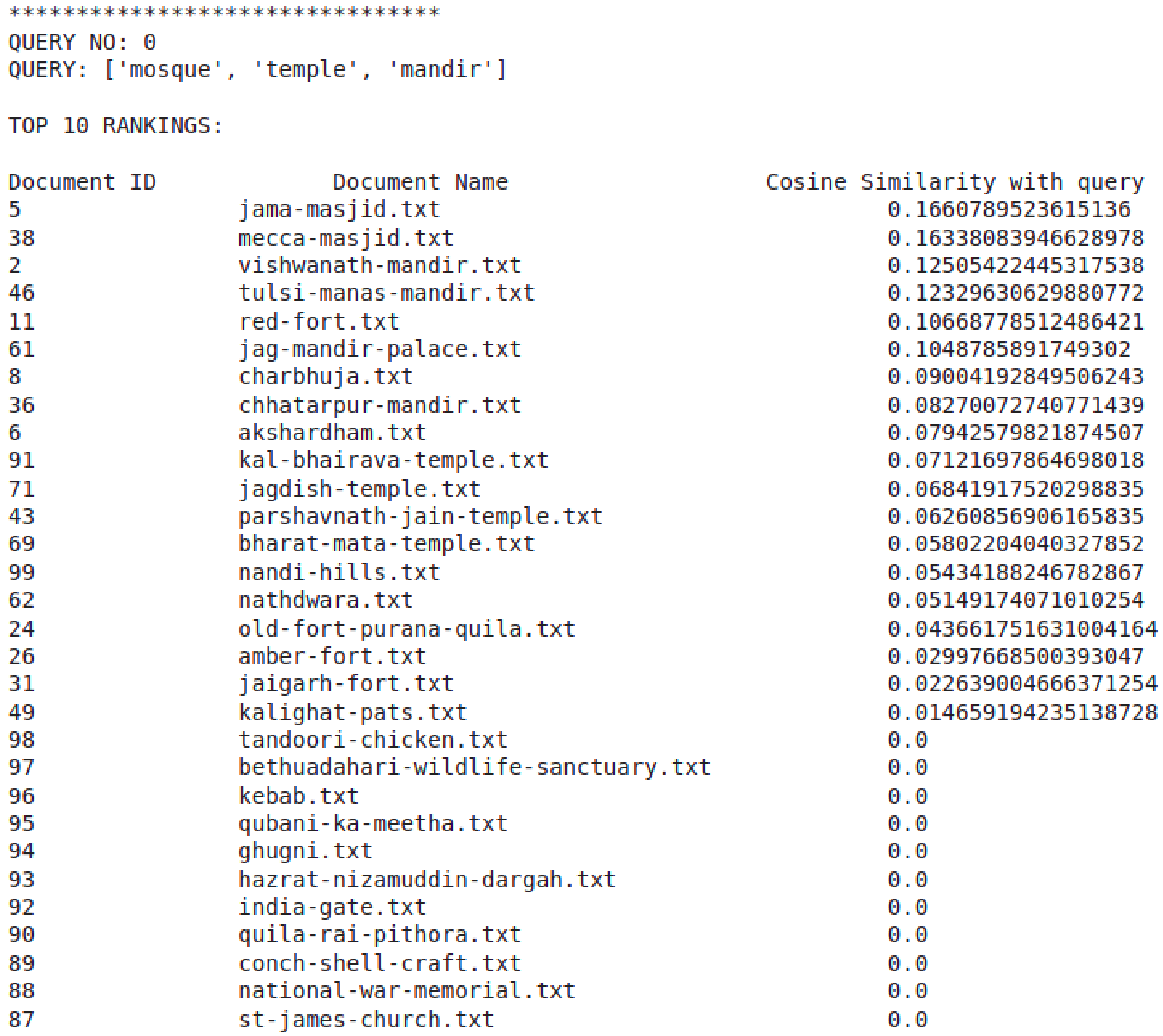
Q. Implement positive feedback mechanism (beta = 1) and analyze the changes in the ranked list for query lengths 3 and 5. How many terms are added and deleted and what is their effect on the original ranking of the documents over each successive reformulation? Clearly demonstrate the effect.

The queries with lengths 3 and 5 used throughout the experiments is

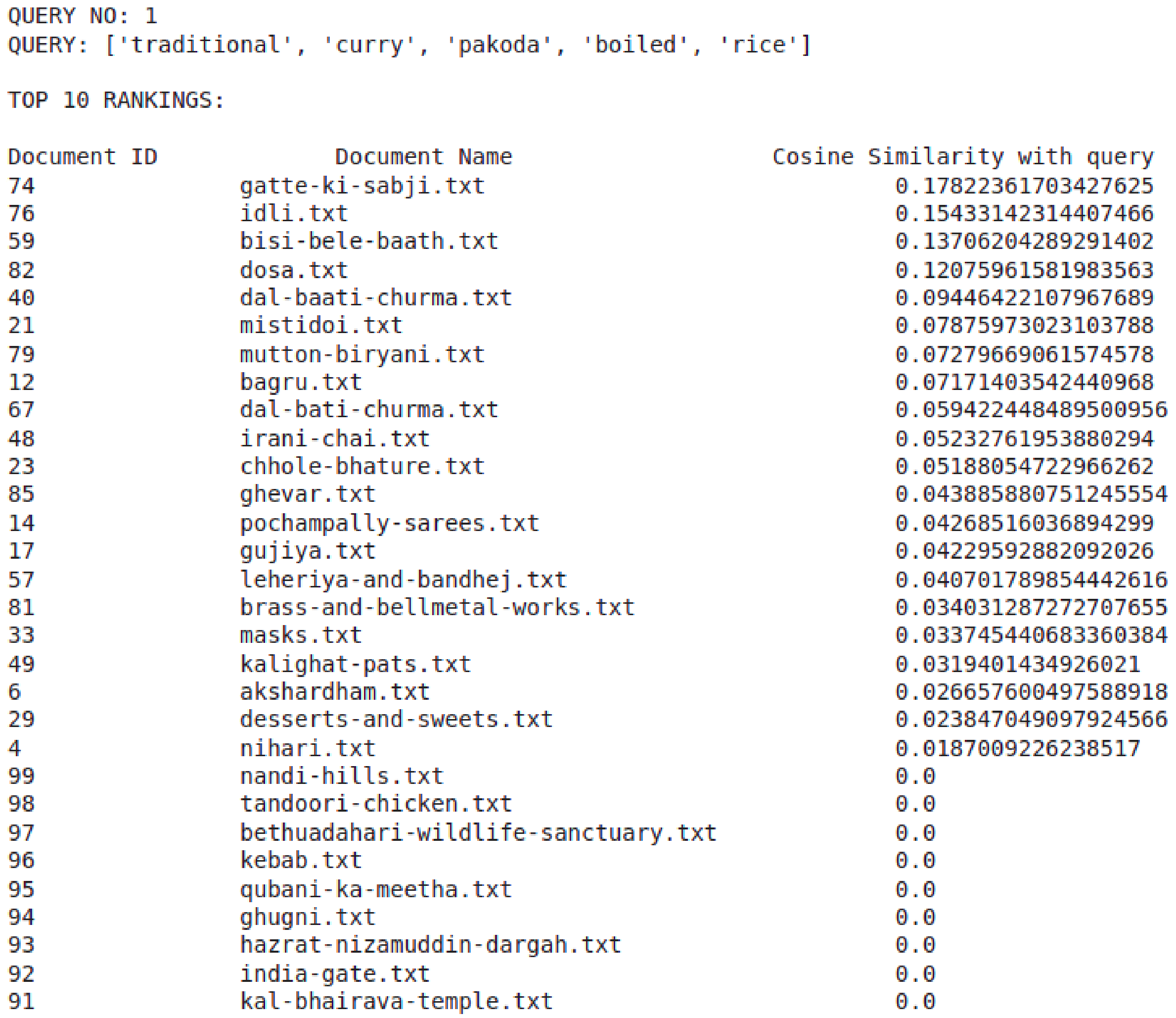
Q0: `['mosque','temple','mandir']`

Q1: `['traditional','curry','pakoda','boiled','rice']`

By using cosine similarity, the ranked list obtained is as follows,



```
********************************
QUERY NO: 0
QUERY: ['mosque', 'temple', 'mandir']

TOP 10 RANKINGS:

Document ID             Document Name                   Cosine Similarity with query
5                jama-masjid.txt                                 0.1660789523615136
38               mecca-masjid.txt                                0.16338083946628978
2                vishwanath-mandir.txt                           0.12505422445317538
46               tulsi-manas-mandir.txt                          0.12329630629880772
11               red-fort.txt                                    0.10668778512486421
61               jag-mandir-palace.txt                           0.1048785891749302
8                charbhuja.txt                                   0.09004192849506243
36               chhatarpur-mandir.txt                           0.08270072740771439
6                akshardham.txt                                  0.07942579821874507
91               kal-bhairava-temple.txt                         0.07121697864698018
71               jagdish-temple.txt                              0.06841917520298835
43               parshavnath-jain-temple.txt                     0.06260856906165835
69               bharat-mata-temple.txt                          0.05802204040327852
99               nandi-hills.txt                                 0.05434188246782867
62               nathdwara.txt                                   0.05149174071010254
24               old-fort-purana-quila.txt                       0.043661751631004164
26               amber-fort.txt                                  0.02997668500393047
31               jaigarh-fort.txt                                0.022639004666371254
49               kalighat-pats.txt                               0.014659194235138728
98               tandoori-chicken.txt                            0.0
97               bethuadahari-wildlife-sanctuary.txt             0.0
96               kebab.txt                                       0.0
95               qubani-ka-meetha.txt                            0.0
94               ghugni.txt                                      0.0
93               hazrat-nizamuddin-dargah.txt                    0.0
92               india-gate.txt                                  0.0
90               quila-rai-pithora.txt                           0.0
89               conch-shell-craft.txt                           0.0
88               national-war-memorial.txt                       0.0
87               st-james-church.txt                             0.0
```

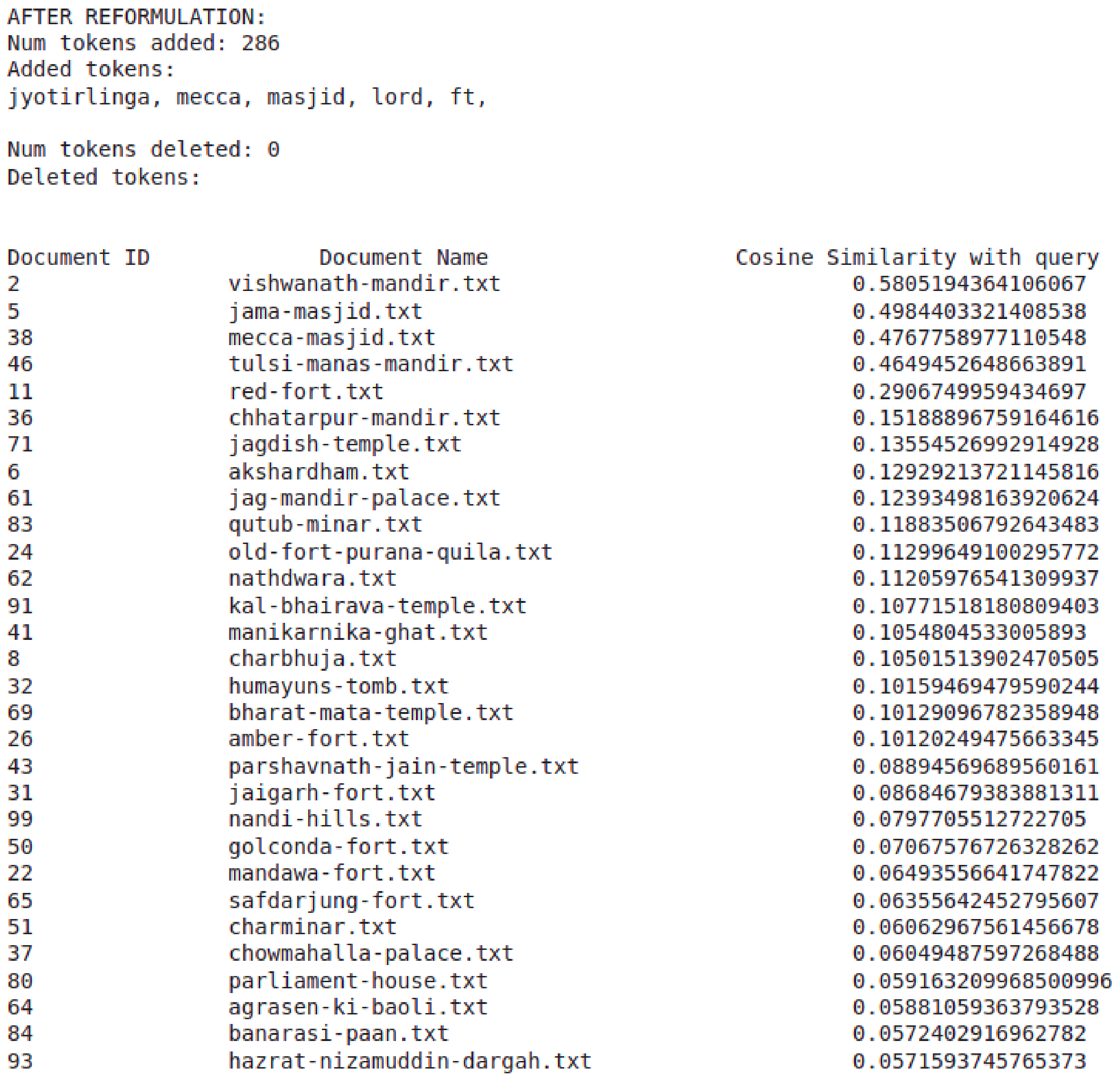


```
QUERY NO: 1
QUERY: ['traditional', 'curry', 'pakoda', 'boiled', 'rice']

TOP 10 RANKINGS:

Document ID            Document Name                    Cosine Similarity with query
74                gatte-ki-sabji.txt                             0.17822361703427625
76                idli.txt                                       0.15433142314407466
59                bisi-bele-baath.txt                            0.13706204289291402
82                dosa.txt                                       0.12075961581983563
40                dal-baati-churma.txt                           0.09446422107967689
21                mistidoi.txt                                   0.07875973023103788
79                mutton-biryani.txt                             0.07279669061574578
12                bagru.txt                                      0.07171403542440968
67                dal-bati-churma.txt                            0.059422448489500956
48                irani-chai.txt                                 0.05232761953880294
23                chhole-bhature.txt                             0.05188054722966262
85                ghevar.txt                                     0.043885880751245554
14                pochampally-sarees.txt                         0.04268516036894299
17                gujiya.txt                                     0.04229592882092026
57                leheriya-and-bandhej.txt                       0.040701789854442616
81                brass-and-bellmetal-works.txt                  0.034031287272707655
33                masks.txt                                      0.033745440683360384
49                kalighat-pats.txt                              0.0319401434926021
6                 akshardham.txt                                 0.026657600497588918
29                desserts-and-sweets.txt                        0.023847049097924566
4                 nihari.txt                                     0.0187009226238517
99                nandi-hills.txt                                0.0
98                tandoori-chicken.txt                           0.0
97                bethuadahari-wildlife-sanctuary.txt            0.0
96                kebab.txt                                      0.0
95                qubani-ka-meetha.txt                           0.0
94                ghugni.txt                                     0.0
93                hazrat-nizamuddin-dargah.txt                   0.0
92                india-gate.txt                                 0.0
91                kal-bhairava-temple.txt                        0.0
```

The above picture shows the top 30 ranked documents obtained using the vector space model for each of the queries. Now it is assumed that top 5% of the documents are relevant and others are non relevant. This value can be altered.

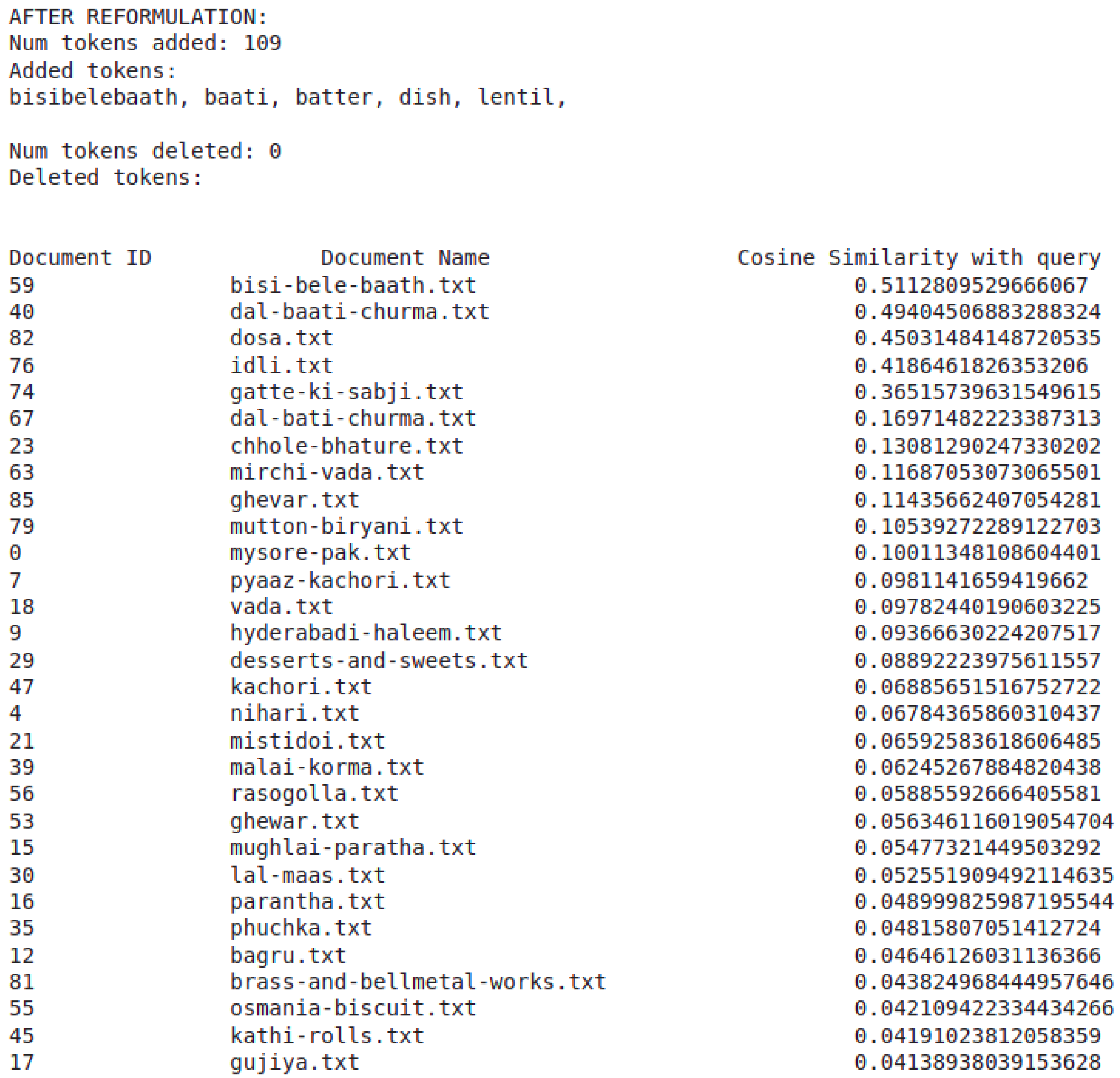
When beta value is 1, i.e. positive feedback mechanism the changes in results for each of the query are as follows,

Query 0:



```
AFTER REFORMULATION:
Num tokens added: 286
Added tokens:
jyotirlinga, mecca, masjid, lord, ft,

Num tokens deleted: 0
Deleted tokens:


Document ID             Document Name                   Cosine Similarity with query
2                vishwanath-mandir.txt                           0.5805194364106067
5                jama-masjid.txt                                 0.4984403321408538
38               mecca-masjid.txt                                0.4767758977110548
46               tulsi-manas-mandir.txt                          0.4649452648663891
11               red-fort.txt                                    0.2906749959434697
36               chhatarpur-mandir.txt                           0.15188896759164616
71               jagdish-temple.txt                              0.13554526992914928
6                akshardham.txt                                  0.12929213721145816
61               jag-mandir-palace.txt                           0.12393498163920624
83               qutub-minar.txt                                 0.11883506792643483
24               old-fort-purana-quila.txt                       0.11299649100295772
62               nathdwara.txt                                   0.11205976541309937
91               kal-bhairava-temple.txt                         0.10771518180809403
41               manikarnika-ghat.txt                            0.1054804533005893
8                charbhuja.txt                                   0.10501513902470505
32               humayuns-tomb.txt                               0.10159469479590244
69               bharat-mata-temple.txt                          0.10129096782358948
26               amber-fort.txt                                  0.10120249475663345
43               parshavnath-jain-temple.txt                     0.08894569689560161
31               jaigarh-fort.txt                                0.08684679383881311
99               nandi-hills.txt                                 0.0797705512722705
50               golconda-fort.txt                               0.07067576726328262
22               mandawa-fort.txt                                0.06493556641747822
65               safdarjung-fort.txt                             0.06355642452795607
51               charminar.txt                                   0.06062967561456678
37               chowmahalla-palace.txt                          0.06049487597268488
80               parliament-house.txt                            0.059163209968500996
64               agrasen-ki-baoli.txt                            0.05881059363793528
84               banarasi-paan.txt                               0.0572402916962782
93               hazrat-nizamuddin-dargah.txt                    0.0571593745765373
```

Query 1:



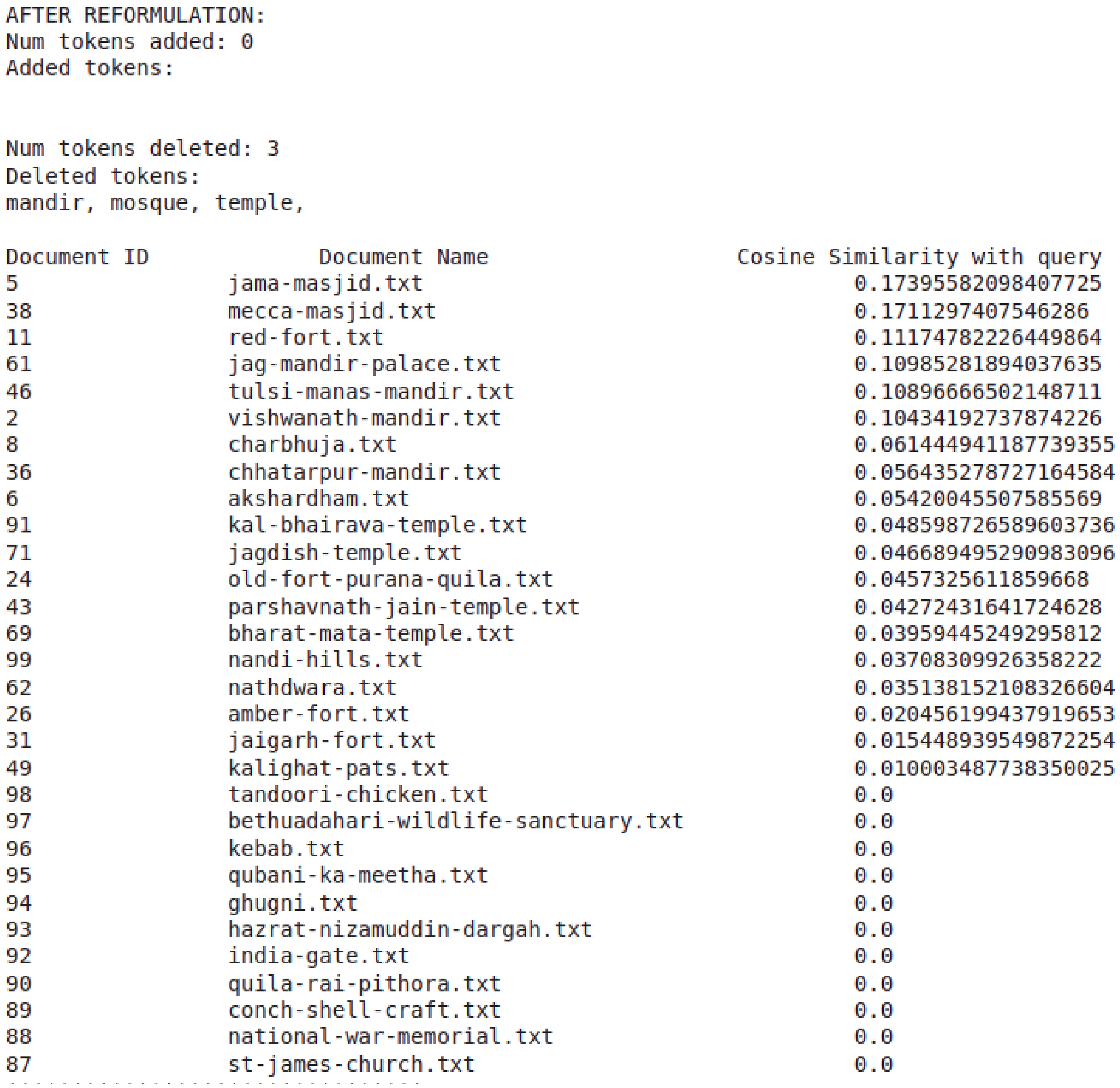
```
AFTER REFORMULATION:
Num tokens added: 109
Added tokens:
bisibelebaath, baati, batter, dish, lentil,

Num tokens deleted: 0
Deleted tokens:


Document ID          Document Name                     Cosine Similarity with query
59                bisi-bele-baath.txt                           0.5112809529666067
40                dal-baati-churma.txt                          0.49404506883288324
82                dosa.txt                                      0.45031484148720535
76                idli.txt                                      0.4186461826353206
74                gatte-ki-sabji.txt                            0.36515739631549615
67                dal-bati-churma.txt                           0.16971482223387313
23                chhole-bhature.txt                            0.13081290247330202
63                mirchi-vada.txt                               0.11687053073065501
85                ghevar.txt                                    0.11435662407054281
79                mutton-biryani.txt                            0.10539272289122703
0                 mysore-pak.txt                                0.10011348108604401
7                 pyaaz-kachori.txt                             0.0981141659419662
18                vada.txt                                      0.09782440190603225
9                 hyderabadi-haleem.txt                         0.09366630224207517
29                desserts-and-sweets.txt                       0.08892223975611557
47                kachori.txt                                   0.06885651516752722
4                 nihari.txt                                    0.06784365860310437
21                mistidoi.txt                                  0.06592583618606485
39                malai-korma.txt                               0.06245267884820438
56                rasogolla.txt                                 0.05885592666405581
53                ghewar.txt                                    0.056346116019054704
15                mughlai-paratha.txt                           0.05477321449503292
30                lal-maas.txt                                  0.052551909492114635
16                parantha.txt                                  0.048999825987195544
35                phuchka.txt                                   0.04815807051412724
12                bagru.txt                                     0.04646126031136366
81                brass-and-bellmetal-works.txt                 0.043824968444957646
55                osmania-biscuit.txt                           0.042109422334434266
45                kathi-rolls.txt                               0.04191023812058359
17                gujiya.txt                                    0.04138938039153628
```

<u>Observations:</u> It is observed that when beta = 1, many tokens are added to the original query. It can be seen for the first query (Q0) that a total of 286 new tokens are added to the query after 3 rocchio's reformulations on the query. Similarly, 109 new tokens are added into the second query after 3 reformulations. It can also be observed that no terms are deleted as this is a positive feedback mechanism. Some of the added tokens are also printed which are pretty semantically related to the tokens in the original query. (eg, lord, mecca, masjid all are tokens belonging to the same category). Due to this the rankings also change and the search space of the VSM is increased giving a large variety of relevant results.

Q. Beta = 0, Gamma = 1,

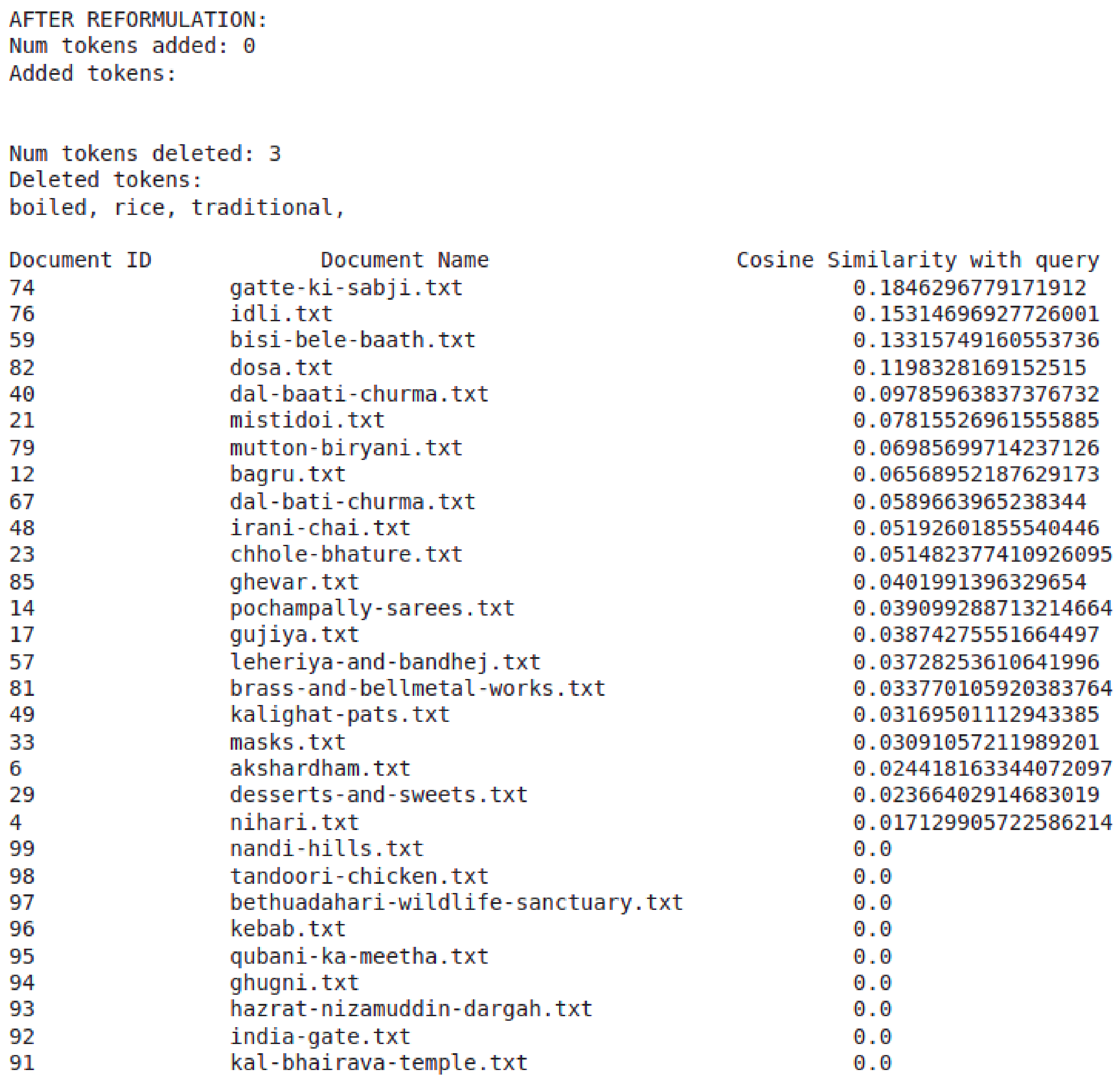
After 3 reformulations , the results obtained for each of the query are,

Query 0:



```
AFTER REFORMULATION:
Num tokens added: 0
Added tokens:


Num tokens deleted: 3
Deleted tokens:
mandir, mosque, temple,

Document ID             Document Name                   Cosine Similarity with query
5                jama-masjid.txt                                 0.17395582098407725
38               mecca-masjid.txt                                0.1711297407546286
11               red-fort.txt                                    0.11174782226449864
61               jag-mandir-palace.txt                           0.10985281894037635
46               tulsi-manas-mandir.txt                          0.10896666502148711
2                vishwanath-mandir.txt                           0.10434192737874226
8                charbhuja.txt                                   0.061444941187739355
36               chhatarpur-mandir.txt                           0.056435278727164584
6                akshardham.txt                                  0.05420045507585569
91               kal-bhairava-temple.txt                         0.048598726589603736
71               jagdish-temple.txt                              0.046689495290983096
24               old-fort-purana-quila.txt                       0.0457325611859668
43               parshavnath-jain-temple.txt                     0.04272431641724628
69               bharat-mata-temple.txt                          0.03959445249295812
99               nandi-hills.txt                                 0.03708309926358222
62               nathdwara.txt                                   0.035138152108326604
26               amber-fort.txt                                  0.020456199437919653
31               jaigarh-fort.txt                                0.015448939549872254
49               kalighat-pats.txt                               0.010003487738350025
98               tandoori-chicken.txt                            0.0
97               bethuadahari-wildlife-sanctuary.txt             0.0
96               kebab.txt                                       0.0
95               qubani-ka-meetha.txt                            0.0
94               ghugni.txt                                      0.0
93               hazrat-nizamuddin-dargah.txt                    0.0
92               india-gate.txt                                  0.0
90               quila-rai-pithora.txt                           0.0
89               conch-shell-craft.txt                           0.0
88               national-war-memorial.txt                       0.0
87               st-james-church.txt                             0.0
```

Query 1:

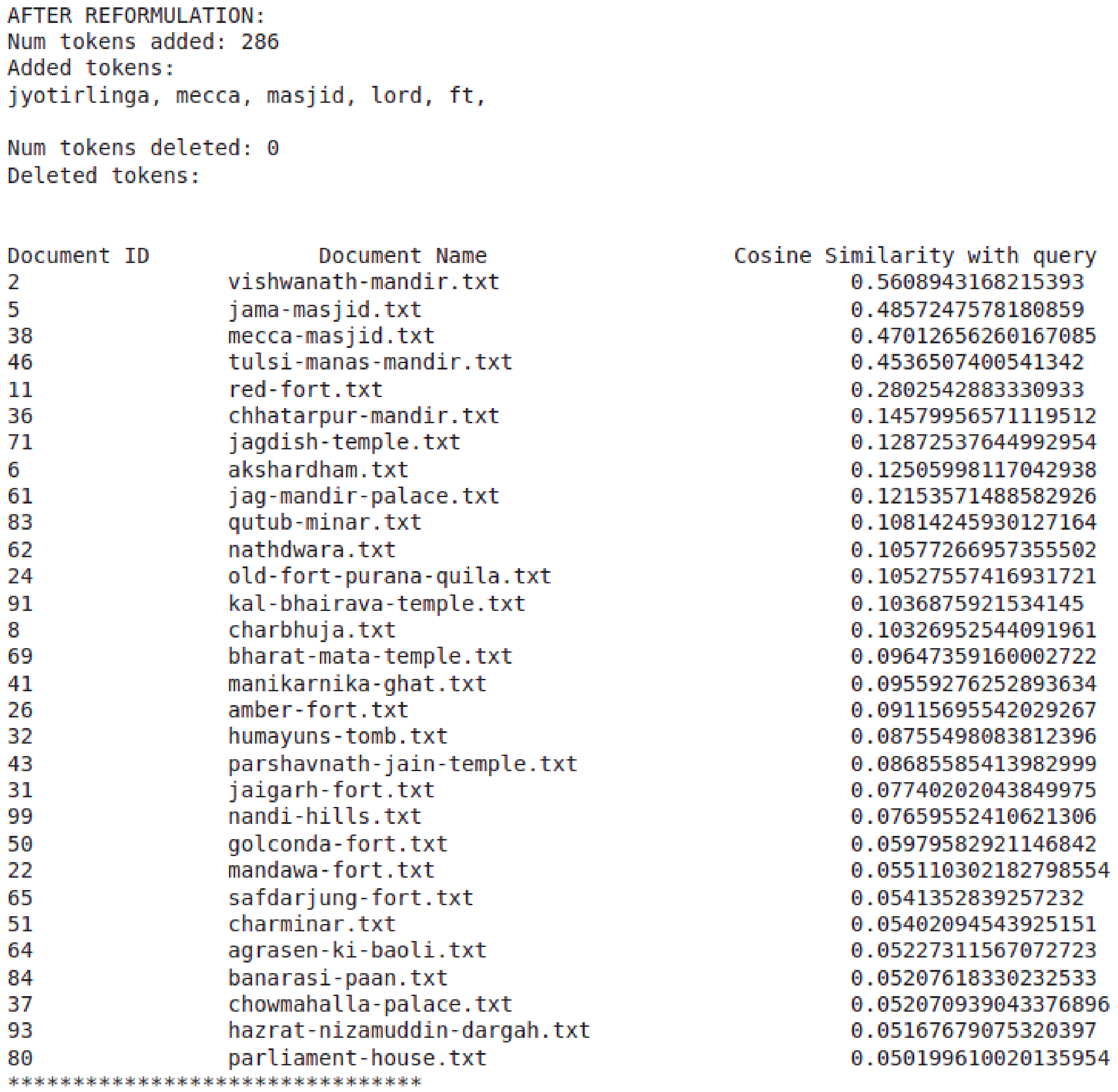


```
AFTER REFORMULATION:
Num tokens added: 0
Added tokens:


Num tokens deleted: 3
Deleted tokens:
boiled, rice, traditional,

Document ID             Document Name                   Cosine Similarity with query
74               gatte-ki-sabji.txt                                 0.1846296779171912
76               idli.txt                                           0.15314696927726001
59               bisi-bele-baath.txt                                0.13315749160553736
82               dosa.txt                                           0.1198328169152515
40               dal-baati-churma.txt                               0.09785963837376732
21               mistidoi.txt                                       0.07815526961555885
79               mutton-biryani.txt                                 0.06985699714237126
12               bagru.txt                                          0.06568952187629173
67               dal-bati-churma.txt                                0.0589663965238344
48               irani-chai.txt                                     0.05192601855540446
23               chhole-bhature.txt                                 0.051482377410926095
85               ghevar.txt                                         0.0401991396329654
14               pochampally-sarees.txt                             0.039099288713214664
17               gujiya.txt                                         0.03874275551664497
57               leheriya-and-bandhej.txt                           0.03728253610641996
81               brass-and-bellmetal-works.txt                      0.033770105920383764
49               kalighat-pats.txt                                  0.03169501112943385
33               masks.txt                                          0.03091057211989201
6                akshardham.txt                                     0.024418163344072097
29               desserts-and-sweets.txt                            0.02366402914683019
4                nihari.txt                                         0.017129905722586214
99               nandi-hills.txt                                    0.0
98               tandoori-chicken.txt                               0.0
97               bethuadahari-wildlife-sanctuary.txt                0.0
96               kebab.txt                                          0.0
95               qubani-ka-meetha.txt                               0.0
94               ghugni.txt                                         0.0
93               hazrat-nizamuddin-dargah.txt                       0.0
92               india-gate.txt                                     0.0
91               kal-bhairava-temple.txt                            0.0
```

Observations: It can be seen that there are no new tokens added as the feedback mechanism only favors negative feedbacks. Some of the terms though do experience a decrease in their original tf-idf weights in the original query. These terms are considered as ‘deleted’. It can be observed that 3 tokens get deleted in both of the queries and no terms are being added.

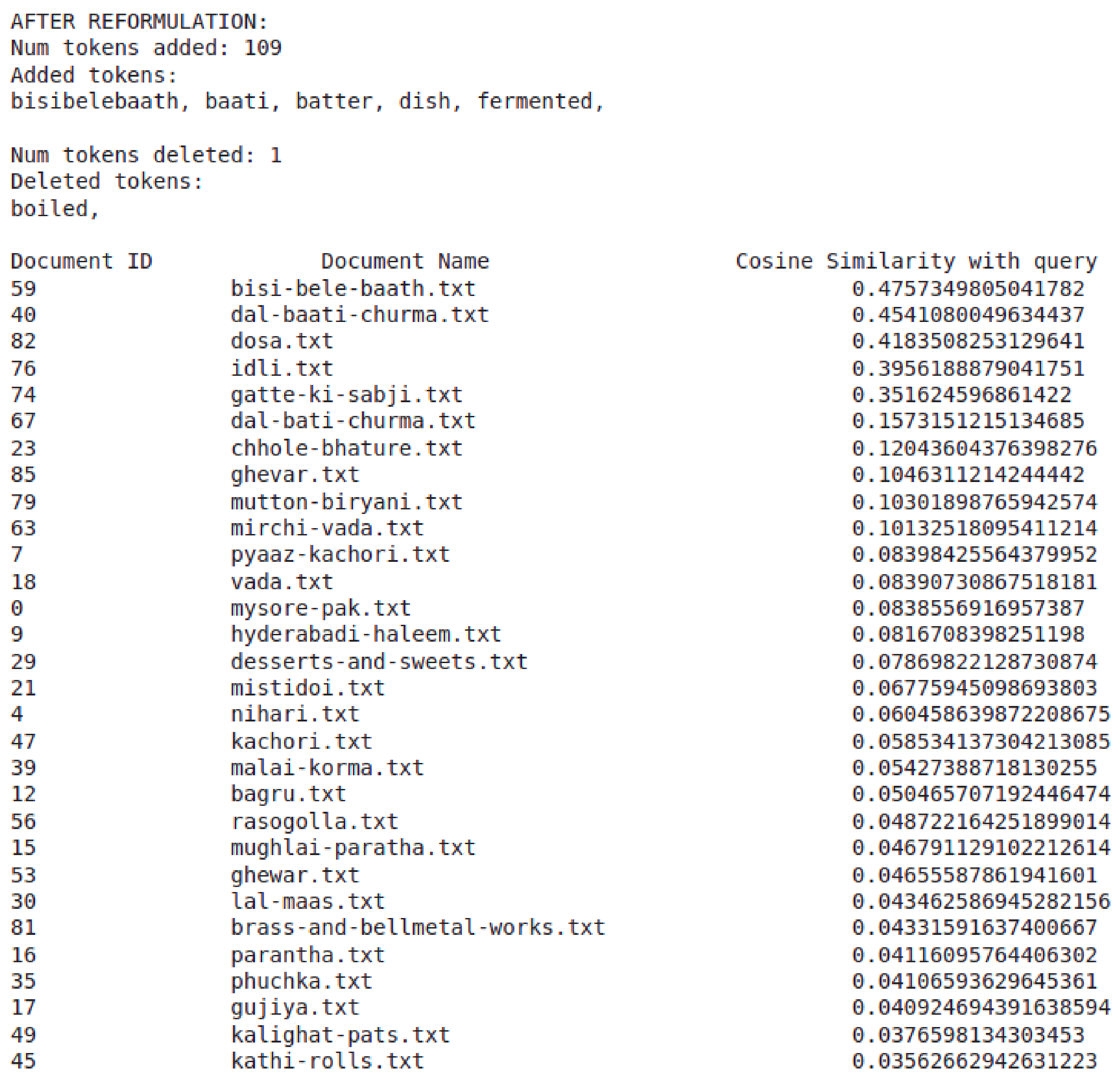
Q3. Beta = 0.75, gamma = 0.25.

Query 0:



```
AFTER REFORMULATION:
Num tokens added: 286
Added tokens:
jyotirlinga, mecca, masjid, lord, ft,

Num tokens deleted: 0
Deleted tokens:


Document ID             Document Name                   Cosine Similarity with query
2                vishwanath-mandir.txt                           0.5608943168215393
5                jama-masjid.txt                                 0.4857247578180859
38               mecca-masjid.txt                                0.47012656260167085
46               tulsi-manas-mandir.txt                          0.4536507400541342
11               red-fort.txt                                    0.2802542883330933
36               chhatarpur-mandir.txt                           0.14579956571119512
71               jagdish-temple.txt                              0.12872537644992954
6                akshardham.txt                                  0.12505998117042938
61               jag-mandir-palace.txt                           0.12153571488582926
83               qutub-minar.txt                                 0.10814245930127164
62               nathdwara.txt                                   0.10577266957355502
24               old-fort-purana-quila.txt                       0.10527557416931721
91               kal-bhairava-temple.txt                         0.1036875921534145
8                charbhuja.txt                                   0.10326952544091961
69               bharat-mata-temple.txt                          0.09647359160002722
41               manikarnika-ghat.txt                            0.09559276252893634
26               amber-fort.txt                                  0.09115695542029267
32               humayuns-tomb.txt                               0.08755498083812396
43               parshavnath-jain-temple.txt                     0.08685585413982999
31               jaigarh-fort.txt                                0.07740202043849975
99               nandi-hills.txt                                 0.07659552410621306
50               golconda-fort.txt                               0.05979582921146842
22               mandawa-fort.txt                                0.055110302182798554
65               safdarjung-fort.txt                             0.0541352839257232
51               charminar.txt                                   0.05402094543925151
64               agrasen-ki-baoli.txt                            0.05227311567072723
84               banarasi-paan.txt                               0.05207618330232533
37               chowmahalla-palace.txt                          0.052070939043376896
93               hazrat-nizamuddin-dargah.txt                    0.05167679075320397
80               parliament-house.txt                            0.050199610020135954
********************************
```

Query 1:



```
AFTER REFORMULATION:
Num tokens added: 109
Added tokens:
bisibelebaath, baati, batter, dish, fermented,

Num tokens deleted: 1
Deleted tokens:
boiled,

Document ID             Document Name                  Cosine Similarity with query
59               bisi-bele-baath.txt                             0.4757349805041782
40               dal-baati-churma.txt                            0.4541080049634437
82               dosa.txt                                        0.4183508253129641
76               idli.txt                                        0.3956188879041751
74               gatte-ki-sabji.txt                              0.351624596861422
67               dal-bati-churma.txt                             0.1573151215134685
23               chhole-bhature.txt                              0.12043604376398276
85               ghevar.txt                                      0.1046311214244442
79               mutton-biryani.txt                              0.10301898765942574
63               mirchi-vada.txt                                 0.10132518095411214
7                pyaaz-kachori.txt                               0.08398425564379952
18               vada.txt                                        0.08390730867518181
0                mysore-pak.txt                                  0.0838556916957387
9                hyderabadi-haleem.txt                           0.0816708398251198
29               desserts-and-sweets.txt                         0.07869822128730874
21               mistidoi.txt                                    0.06775945098693803
4                nihari.txt                                      0.060458639872208675
47               kachori.txt                                     0.058534137304213085
39               malai-korma.txt                                 0.05427388718130255
12               bagru.txt                                       0.050465707192446474
56               rasogolla.txt                                   0.048722164251899014
15               mughlai-paratha.txt                             0.046791129102212614
53               ghewar.txt                                      0.04655587861941601
30               lal-maas.txt                                    0.043462586945282156
81               brass-and-bellmetal-works.txt                   0.04331591637400667
16               parantha.txt                                    0.04116095764406302
35               phuchka.txt                                     0.04106593629645361
17               gujiya.txt                                      0.040924694391638594
49               kalighat-pats.txt                               0.0376598134303453
45               kathi-rolls.txt                                 0.03562662942631223
```

Observations: Here, it can be observed that 286 and 109 terms get added respectively in each of the queries. It can be also observed that one term got deleted in the second query (Q1). The deleted terms are the ones which are not very related to the original query terms. The ranked lists can be seen visually that many related documents are now ranked above in the rankings. Hence, it can be proved that rocchio’s reformulation works well.

THANK YOU