Information Retrieval

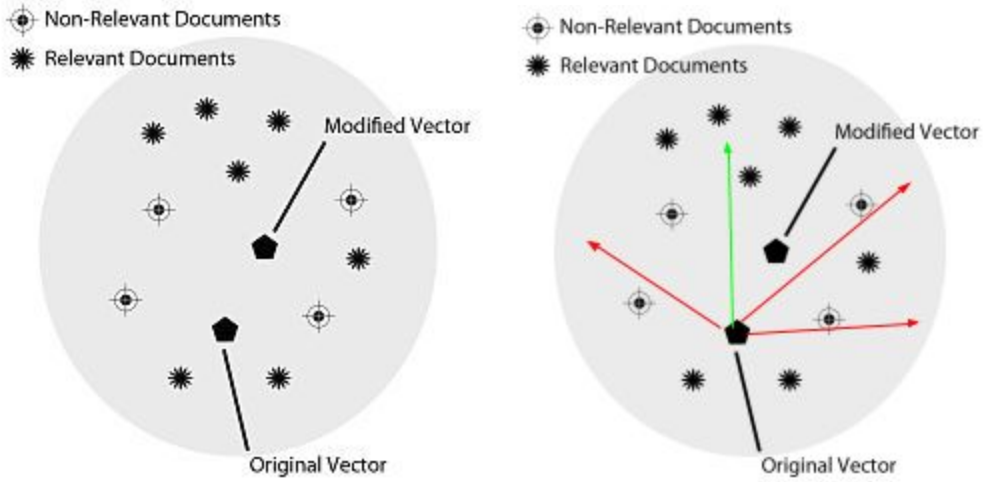
End Semester Exam

Question 1

$$\overrightarrow{Q_m} = (a \cdot \overrightarrow{Q_o}) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)$$

(Source: Wikipedia)

If the **weight for relevant documents (b) is decreased**, if the expected outcome is to have the new query moving away from centroid of relevant documents.



(Source: Wikipedia)

Explanation: The new Query vector Qm (coordinates), will move away from the centroid, if **non relevant documents are favoured** over the relevant as you can see from the image above.

- If we decrease the weight of relevant documents (b) or increase the weight of non relevant documents (c), the new query vector will move away from the centroid in the vector space model (depicted by the **red arrows)**, because we are essentially giving more weight to the non relevant documents and the general direction of these documents lie away from the centroid.
- If we increase the weight of relevant documents (b) or decrease the weight of non relevant documents (c), the new query vector will move towards the centroid in the vector space model (depicted by the **green arrows)**, because we are essentially giving more weight to the relevant documents and the general direction of these documents lie towards the centroid.

Question 2

Rocchio's Relevance Feedback Algorithm

$$\overrightarrow{Q_m} = (a \cdot \overrightarrow{Q_o}) + \left(b \cdot \frac{1}{|D_r|} \cdot \sum_{\overrightarrow{D_j} \in D_r} \overrightarrow{D_j}\right) - \left(c \cdot \frac{1}{|D_{nr}|} \cdot \sum_{\overrightarrow{D_k} \in D_{nr}} \overrightarrow{D_k}\right)$$

(Source: Wikipedia)

- Clearly, the RD (relevant documents) = **{D2, D3, D5, D6, D9}** are given a preference (implying, weight of relevant documents **b is close to 1**) in the resulting query = {**D2, D3, D5, D6, D9,** D11, D13, D14, D15} and

- IRD = {D1, D4, D7, D8, D10} are given no preference, therefore they should be removed from the original query (implying, weight of irrelevant documents **c is 0** since the third term in the equation has a negative sign).

- Q0 results = {D1, **D2, D3,** D4, **D5, D6,** D7, D8, **D9,** D10}
  Also, to remove IRD from original query results, the weight of the original query results in the equation above, **a should be 1** to cancel out the IRD and keep RD.

- Note that to take into account the normalization done in the formula by using the lengths of the RD and IRD vectors.

The equation with weights now look like,

Qm = {D2, D3, D5, D6, D9} + new_documents

Therefore, approximately **a = 1 b = 1 and c = 0**.

Question 3

Query	Tf-idf score (Rank, docID, Score)			BM25 score		
Portable operating system	1	3127	0.100407969844	1	1461	0.851006812757
	2	2246	0.0708559728367	2	2311	0.687930173134
	3	2311	0.064183214871	3	1755	0.622014102817
	4	1461	0.0632713808896	4	2796	0.453938760767
	5	1930	0.0555786634273	5	2735	0.451699246445
	6	1591	0.0550802009696	6	379	0.448616149363
	7	1680	0.0535106101622	7	2405	0.448004964973
	8	1755	0.0479328643103	8	190	0.441278232247
	9	1033	0.0470590258644	9	2689	0.435751294307
	10	3068	0.0448056870588	10	2984	0.429667863582
Parallel algorithm	1	2664	0.0645439161889	1	371	0.744376551823
	2	2714	0.0616798930983	2	199	0.735161851255
	3	2973	0.0598187298652	3	2509	0.654823450718
	4	141	0.0576284965972	4	2430	0.607454389368
	5	1262	0.0556413070594	5	1990	0.58833331617
	6	950	0.0542565137961	6	1560	0.58833331617
	7	1302	0.0537865968241	7	2510	0.568662172959
	8	392	0.0537865968241	8	2090	0.534994249906
	9	2266	0.0525970507947	9	2472	0.499990935839
	10	2685	0.0514978054698	10	1559	0.49403234759
Applied stochastic process	1	1696	0.0898707874579	1	394	0.392657186535
	2	268	0.0808837087122	2	2727	0.385906806263
	3	2065	0.0518931125567	3	141	0.376993238507
	4	2999	0.0510098205023	4	2999	0.368286007248
	5	2727	0.0479110930886	5	268	0.355162757724
	6	293	0.0430100811133	6	597	0.34869728607
	7	927	0.0405523621925	7	392	0.34869728607
	8	3043	0.0386176042221	8	198	0.34869728607
	9	1588	0.0374196128503	9	1085	0.323845936806
	10	3120	0.0357950009045	10	293	0.31693699368
Perform evaluation and model of computer system	1	2318	0.235739430528	1	2318	1.63724603745
	2	2984	0.126055948741	2	2984	1.03064316654
	3	3048	0.103765515637	3	2504	0.895010624175
	4	3070	0.100157984655	4	2311	0.852088854327
	5	1938	0.0989886772492	5	2553	0.851572922419
	6	3089	0.0971103033211	6	2502	0.80739528562
	7	2542	0.0931604199867	7	2743	0.76941173037
	8	1344	0.0925794833052	8	2255	0.73432629124
	9	3136	0.0911233439483	9	3089	0.699609279397
	10	1653	0.0821952703005	10	3136	0.666134254018

Parallel process in information retrieval	1	2288	0.106731788073	1	2278	0.796434597543
	2	2278	0.103726491703	2	141	0.75998176833
	3	891	0.0918639442258	3	392	0.702939875336
	4	141	0.0881432999334	4	651	0.678469319036
	5	1457	0.0847858962943	5	275	0.678469319036
	6	1830	0.083495831255	6	2070	0.654159939246
	7	392	0.0822670799379	7	1085	0.649802186093
	8	651	0.0771451285634	8	1830	0.641047395313
	9	275	0.0771451285634	9	1251	0.601104918492
	10	2070	0.0747343432958	10	239	0.55203179303

Extra Credit

I have updated the Assignment 1 and 3 code to use the Query Relevance feedback with the choice of parameters **a = 1, b = 1, c = 0**. Therefore, any **irrelevant documents will be removed** and **relevant documents** will be used to find new results!

Output