Q1. In the Rocchio feedback formula for query expansion, which parameter would be assigned a lower value or higher value if the expected outcome is to have the new query moving away from centroid of relevant documents? Give justification for your answer? (1)

Q2. For a user query q, ae search engine returns 10 documents,

D= {D1, D2, D3, D4, D5, D6, D7, D8, D9, D10}

The user constitutes a set of relevant documents and irrelevant documents from D.

RD = {D2, D3, D5, D6, D9}, IRD = {D1, D4, D7, D8, D10}

Using above relevance feedback (Rocchio algorithm) a new query q1 is formed which when submitted to the same search engine returns the following 10 documents?

={D2, D3, D5, D6, D9, D11, D13, D14, D15}

What are the approximate values of the three parameters. (4)

Q3. Implement the BM 25 retrieval model in a programming language of your choice, (Refer slides 9 and 10 in attached PDF).

There is a corpus.txt that has the following format

* Each document id begins with a #. The next and the following lines are the document contents that are already stemmed.

You need to build a search engine assuming that no relevance information is available

1. Take the corpus, build an inverted index (only index strings and alphanumeric token)
2. Compute tf-idf score for each query mentioned below
3. Compute BM25 score for each query mentioned below
4. Assume k=1 and b=0.5.
5. Prepare a table with the top 10 documents in the sorted order in the following format

|  |  |  |
| --- | --- | --- |
| Query | Tf-idf score | BM 25 score |
| Portable operating system | Rank1: document-id: score  Rank2: document-d: score  .  .  .  Rank 10: document-id:score | -Same - |
| Parallel algorithm | -Same - | -Same - |
| Applied stochastic process | -Same - | -Same - |
| Perform evaluation and model of computer system | -Same - | -Same - |
| Parallel process in information retrieval | -Same - | -Same - |

Note that, you need to stem the above queries because the document contents are already stemmed. (20)