Assignment 3

Information Retrieval

- Mayank Chauhan, MT18008

Q1 . Relevance Feedback

Rocchio Algorithm

$$q_{opt} = \alpha q_0 + \beta \frac{1}{|C_+|} \sum_{d \varepsilon C_+} d - \gamma \frac{1}{|C_-|} \sum_{d \varepsilon C_-} d$$

Where the weighting parameters are set as $\alpha = 1,\ \beta = 0.75$ and $\gamma = 0.15$

T-SNE Visualization on an example query -

q_0 *= LightWave3D is part of a suite argument against automobiles*

Table 1. Documents retrieved for the given query

Rank	Relevance Feedback 1	Relevance Feedback 2	Relevance Feedback 3	Final Result
1	8	1014 (relevant)	1014 (relevant)	1014
2	1014 (relevant)	8	1558 (relevant)	1558
3	837	1558 (relevant)	1469 (relevant)	1469
4	590	1324	1324 (relevant)	1324
5	908	1469	8	1343
6	973	837	1400	8
7	1389	1737	1590	1400
8	1263	1333	1077	1643
9	452	337	1151	1004
10	1224	1224	775	1852

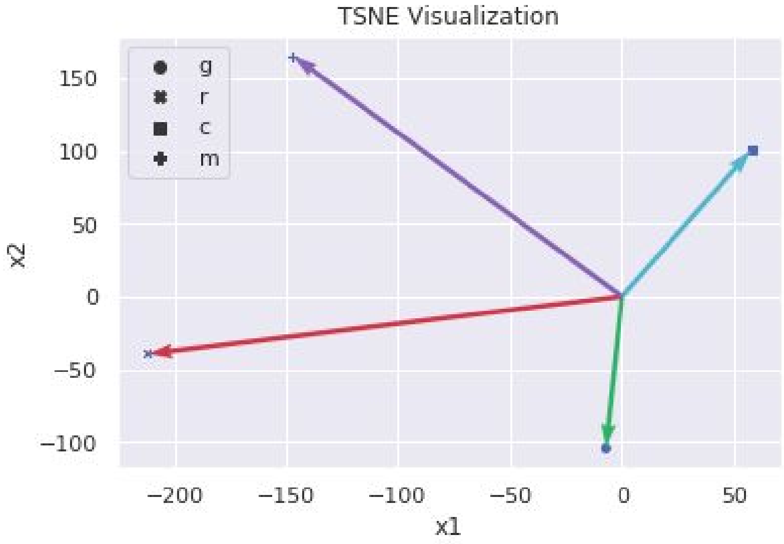


Fig 1. First relevance feedback.

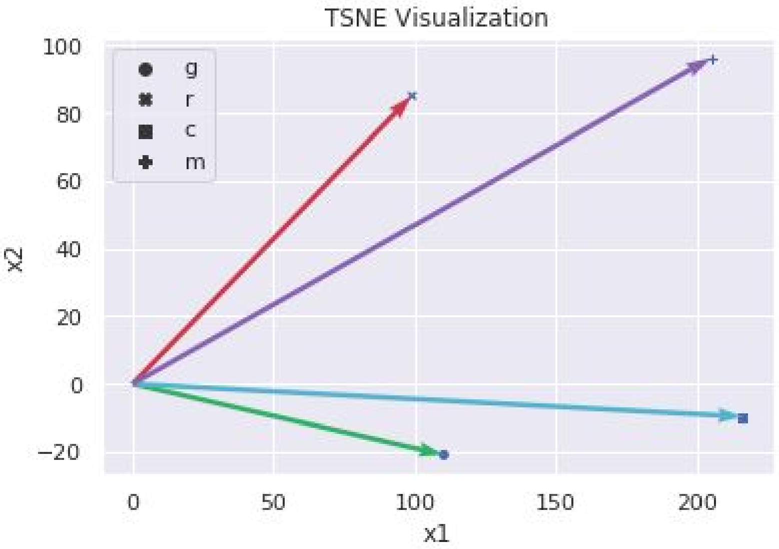


Fig 2. Second relevance feedback.

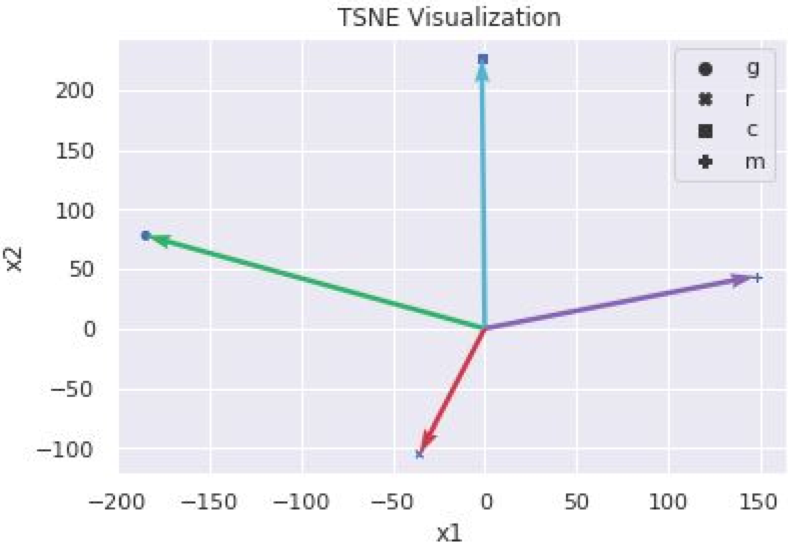


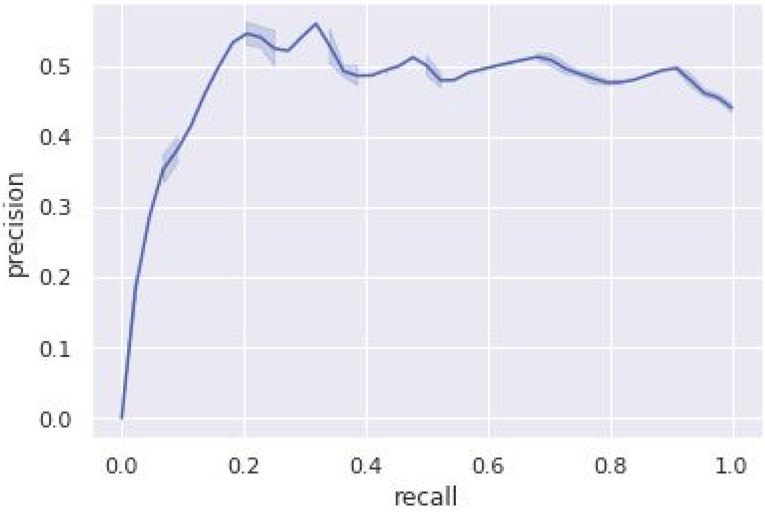
Fig 3. Third relevance feedback.

Analysis

Green (g) = Optimal query
Red (r) = Old query
Cyan (c) = Centroid of relevant documents
Magenta (m) = Centroid of non-relevant documents

From the visualizations we can see, the optimal query has maximum cosine similarity with relevant documents and minimum cosine similarity with non-relevant documents. Documents with id 0 to 1000 are from graphics folder and 1001 to 2000 are from motorcycles folder. And, the final results have most of the documents with id more than 1000 because in the feedback only documents with id more than 1000 were assigned as relevant.

Q2. Precision Recall Curve


precision
recall