Assignment 3 : Rocchio Algorithm

Submitted by : Nikita Bafna

Rocchio Algorithm comparison:

	β = 0.2	**0.4**	**0.6**	**0.8**	**1**
γ = 0	0.3786	0.3813	0.3790	0.3718	0.3713
0.2	0.3781	0.3769	0.3788	**0.3823**	0.3798
0.4	0.3778	0.3762	0.3787	0.3795	0.3762
0.6	0.3742	0.3716	0.3736	0.3784	0.3757
0.8	0.3738	0.3725	0.3716	0.3703	0.3716
1	0.3707	0.3704	0.3717	0.3716	0.3703

- The formula for Rocchio feedback algorithm is given by:

$$\vec{q}_m = \alpha\vec{q}_0 + \beta\frac{1}{|D_r|}\sum_{\vec{d}_j \in D_r}\vec{d}_j - \gamma\frac{1}{|D_{nr}|}\sum_{\vec{d}_j \in D_{nr}}\vec{d}_j$$

- The above values are the F1 scores calculated using below formula:

F1 score : 2(Precision * Recall)/(Precision + Recall)*

- To calculate the Precision and Recall, I Have considered the values @10 for both in the TrecEval results.

- From the table, it can be seen that the algorithm performs the best at below settings:

β = 0.8 and ϒ=0.2

- The reason behind this is that, Beta is multiplied with the positive feedback(relevant documents vector) in the formula. And the algorithm gives more importance to positive feedback. Hence, with bigger value of Beta ie 0.8 we can a better F1 score.

- Gamma, is associated with the negative feedback(non-relevant documents vector) and the algorithm gives less weightage to negative values. Hence at less value of gamma ie 0.2, Rochhio algorithm works with best results.

- We then compared Rochhio algorithm with value Alpha = 1, Beta = 0.8 and gamma = 0.2 with Vector Space model. Below table for comparison:

Evaluation metric	Best Parameter	Vector Space Model
F1 Score	0.3823	0.1675
P@5	0.7760	0.2960
P@10	0.5520	0.3020
P@20	0.3820	0.2600
P@100	0.1840	0.1648
Recall@5	0.2292	0.0539
Recall@10	0.2827	0.0960
Recall@20	0.3230	0.1416
Recall@100	0.4897	0.3578
MAP	0.3942	0.1975
NDCG@5	0.8485	0.3116
NDCG@10	0.7111	0.3183
NDCG@20	0.6072	0.3043
NDCG@100	0.5494	0.3213

From the comparison above, we can say that Rochhio algorithm with parameters of bets = 0.8 and gamma = 0.2 performed better than Vector Space model.

- The basic idea of the Vector Space (VS) model is to represent both a document and a query as a vector in a high-dimensional space where each dimension corresponds to a term.
- The main assumption is that if document vector V1 is closer to the query vector than another document vector V2, then the document represented by V1 is more relevant than the one represented by V2.
- That is, relevance is modeled through similarity between a document and a query.
- Whereas, in Rocchio we simple construct the new query vector by moving the original query vector closer to the centroid vector of the positive/relevant document vectors and farther away from the negative centroid.
- Based on correct settings of alpha, beta and gamma, corresponding to the weight on the original query vector, the positive centroid and the negative centroid.
- In practice, negative examples are often not very useful, so in Rocchio it may involve just moving the query vector closer to the positive centroid.
- In order to avoid overfitting to the relevant examples (often a small sample), we generally need to put a relatively high weight on the original query.
- The relative weight of the original query vs. information extracted from feedback examples often affects feedback performance significantly.

TASK2: Choose the feedback terms

- For the second task, I am using the logic of IDF ie Inverse Document Frequency to find the subset of relevant words.
- IDF helps in filtering out the noisy words such as ‘the, of, in’ etc.
- Once we compute the IDF score for each term in the document, we can easily find the top terms with the highest value of IDF scores.

The formula to calculate the IDF is :

IDF: log(1+N/k(t)) where,
N is total number of documents
K(t) = total number of documents that have the term t

Below table is the comparison table between the best parameters of Alpha, beta and gamma for Rocchio and IDF model.

We can see that the IDF model worked somewhat similar to Rocchio feedback. However, Rocchio feedback giving boost to the weights of positive documents gives better results.

Evaluation metric	Best Parameter	IDF Model
F1 Score	0.3823	0.3648
P@5	0.7760	0.8160
P@10	0.5520	0.5340
P@20	0.3820	0.3580
P@100	0.1840	0.1434
Recall@5	0.2292	0.2375
Recall@10	0.2827	0.2770
Recall@20	0.3230	0.3098
Recall@100	0.4897	0.4144
MAP	0.3942	0.3442
NDCG@5	0.8485	0.8768
NDCG@10	0.7111	0.7017
NDCG@20	0.6072	0.5896
NDCG@100	0.5494	0.4947