COL764
Information Retrieval And Web Search

Assignment 2
Document Reranking Task

Manoj Kumar
2018CS50411

I have implemented the Pseudo-relevance Feedback with Rocchio's method for reranking the documents. The implementation details are as follows:

1. First I create a dictionary 'allwords' which consists of all words and total no. Of documents in which each word in present for atleast one time.

2. Then I calculated the Term frequency of each document using this formulae:
   Term Frequency: which measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. Thus, the term frequency is often divided by the document length (aka. the total number of terms in the document) as a way of normalization:
   TF(t) = (Number of times term t appears in a document) / (Total number of terms in the document).

3. I also stored the total no. Of words in each documents.

4. Now for each document, we also calculate the TF-IDF by multiplying their TF and IDF of each word (like bitwise multiplication).
   IDF: Inverse Document Frequency, which measures how important a term is. While computing TF, all terms are considered equally important. However it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus we need to weigh down the frequent terms while scale up the rare ones, by computing the following:

   IDF(t) = log_e(Total number of documents / Number of documents with term t in it).

5. Now we calculate the tf idf vectors for each query.

6. Now, we apply the Rocchio methods for the modified tfIdf of each query, using this formula:

D_r: set of relevant documents retrieved
D_n: set of non–relevant documents retrieved
α, β, γ : tuning parameters

$$\overrightarrow{q_m} = \alpha \overrightarrow{q_0} + \beta \frac{1}{|D_r|} \sum_{\overrightarrow{d_j} \in D_r} \overrightarrow{d_j} - \gamma \frac{1}{|D_n|} \sum_{\overrightarrow{d_k} \in D_n} \overrightarrow{d_k}$$

- Introduced in Gerard Salton's SMART system [1970s]
- α, β, γ: control the balance between the trust in judged document set vs. the original query
- Typically, γ < β - positive feedback is more valuable than negative feedback

7. Here we take alpha = 1, beta = 0.7 and gamma = 0.1 for optimal results.

8. Using the above formula, we calculate q_m vector for each query.

9. Now for each query and their top 100 relevant documents, we take the cosine of their TfIdf and q_m values of the queries.

10. We use these cosine values for the relevance of a document for each query and rerank top 100 relevant documents for each query.

Manoj Kumar
2018CS50411
4th Year Dual
Computer Science and Engineering
IIT Delhi