Report Information Retrieval

Assignment 3 [Vinit Raj, 19CS10065]

Part 3A(Relevance Feedback)

Preprocessing:

After reading the relevant files, preprocessing similar to part 2A is done from the previous assignment to get the ranked result, ground truth dictionary. A major design change was done in tf-idf representation, working with the whole tf_idf vector was slow, but since it is sparse we can store only the non zero weights with index, hence making relevance calculations faster.

Implementations:

Relevance Feedback:

The function for relevance feedback calculation takes in ranked_list, ground_truth dictionary, doct_tf_idf vectors as dictionary and K.
The alpha, beta, gamma values are not used here because it can easily be multiplied later on if we keep the positive and negative feedback separately and that is what we do.
We calculate the sum of positive feedback vectors(vector representation of relevant docs) and negative feedback(vector representation of non-relevant docs) separately.
Docs with feedback == 2 are taken into the positive sum, others in negative and finally return the normalized vectors separately. In a similar fashion pseudo relevance is calculated by just taking the first 10 cord-ids.