INFORMATION RETRIEVAL
Assignment 3 Report

Made By: Pranav Rajput
Roll Number: 19CS30036

Contribution partwise : 3A (Implementation and design formulation)
3B (code modification and debugging)
Rocchio Result Report : Written

DETAILS
(3A)
Design Formulation:
Contributed to the design for the program was decided upon as group to be vectors that would be realized as dictionaries and then the rocchio's algorithm formula,

$$\vec{Q}_m = a\,\vec{Q}_o + b\,\frac{1}{|D_r|}\sum_{\vec{D}_j \in D_r}\vec{D}_j - c\,\frac{1}{|D_{nr}|}\sum_{\vec{D}_k \in D_{nr}}\vec{D}_k$$

Where a, b, and c are alpha, beta, and gamma, the constants that are varied across 3 sets of values in the question in hand.

Implementation:
Wrote the main function and implemented the csv file creation and modification code based on the mAP and NDCG values (inside main). Wrote the initial draft for the compute_tf_idf function which was then modified by my teammates. Also contributed to the ranked_list implementation for both the relevance and pseudo relevance methods, where a list of lists was used as a parameter.

Debugged the get_ranks function (removing the recurring 0.0 values for the mean average precision and NDCG values for both the relevance feedback and pseudo relevance method).

Debugged the pseudo_relevance feedback function: error removal to the areas related to the conversion of vectors to lists (mainly the positive and negative feedback vector related errors)

(3B)
Debugging
With the computation model being switched to the faster method of computation via the use of dictionaries, the initial draft of the code written to extract the 5 tokens/words with the largest tf-idf values across the documents was modified for the changes related to the dictionary method as discussed in part 2.

(Rocchio Report): Calculated the average values of Normalised Discounted Cumulative Gain and the mean average precision values for the top-20 documents and the average was calculated over the three set of values of **α**, **β**, and **γ** and the results were compared across the 3 evaluation methods considered for the evaluation of results, the lnc-ltc weighing scheme, the relevance feedback method and the pseudo relevance method. Theoretical expectations inferred from lectures were also included in the report to explain the results.