Seung Yoo

Sky274

Assignment 1 report

for the query "Einstein Rosen", documents that contain "Einstein" but not "Rosen" should get 0.5 for this fraction, whereas documents that contain both terms should get a 1.0. This "query count" score should be combined with the existing cosine similarity to produce a final hybrid score that also considers the TF-IDF of the overlapping terms. I first found the total number of tokens the query has. They for every document that has the token inside, I gave it a value one. If the document had two tokens inside, it received a score of 2. In the end, I divided the score the document received from the total number of tokens in the query. I added this divided value onto the final score.