COMP6714 PROJ P2
Z5190746 Alfred LU

1. Find out the the upper bound (α_{max}) of threshold α **such that**
   A. $\forall \alpha < \alpha_{max}$, the algorithm (NB: the modified version of ADCS 2013 algorithm) will *always* return the correct top-k answers for Q;
   B. $\forall \alpha \geq \alpha_{max}$, the algorithm *may* return the wrong top-k answers for Q in some cases.
2. Prove that your answer is correct in a rigorous way.

Answer:

Since the top k answer is already given, means we have the top k score and all selected document IDs sorted descending by the score. For the given query terms, extract all inverted lists with them. After this, find all top k docID in the extracted lists, among all this (docID, score) tuple, get the max score value $\text{maxScore}_{\text{term}}$ for each term lists. At last add all these $\text{maxScore}_{\text{term}}$ together, this will be the new alpha.

invertedList: {[term](DocID, score)} Dict
topK(Q): [(score, DocID)] List
selectedDoc = (DcoID | DocID∈topK(Q))
scoreList = ((term, DocID, score) | term, DocID∈invertedList{[term](DocID, score)})
localUpper = max(scoreList) (key = score)
$\text{alpha}_{\text{max}}$ = ∑localUpper

Proof: This new alpha is strictly less than or equal to the true Upper Bound of **Q**, and also bigger than the **min{score(D;Q) | D∈top-k(Q)}**, it is actually the upper bound of certain documents, which are already proved to be the correct Top K answer.

In the new code, the alpha is mainly used ***as the activation threshold of pivot choosing, the function of limiting the minimum score in Ans list is removed***. For the activation threshold, the code is **"if tem_s_lim > alpha"**, the max value **tem_s_lim** can achieve is the true upper bound of query terms, as long as the alpha is smaller than both the **true upper bound and the local upper bound(which is proved to be $\text{alpha}_{\text{max}}$),** this activation threshold will always be met.

1: alpha < $\text{alpha}_{\text{max}}$
tem_s_lim∈(0, trueUpper)
alpha∈(-inf, $\text{alpha}_{\text{max}}$)
this code will always be met for document that need to be scored.

2: trueUpperBound > alpha > $\text{alpha}_{\text{max}}$
tem_s_lim∈(0, trueUpper)
it is possible that the tem_s_lim is smaller than the alpha because alpha it is bigger than localUpper, in this condition, this document will not be correctly scored, thus missing a needed document at the end.

3: alpha > trueUpperBound
Under this condition, the threshold will never be met. Thus no document will be take into consideration. Resulting in empty answer.