**Implementation details**

Task 2 is implemented using vector space model.

**Calculation of term frequency :**

Since relevance does not increase proportionally with term frequency following was used

1 + log10(tft,d)

If the term frequency is 1 the log will be 0 hence 1 is added

**Calculation of inverse document frequency:**

Rare terms are more informative than frequent terms.

Hence to give high weight to the terms that rarely we use inverse document frequency.

Idf = log10(N/df)

Where N is total number of documents in corpus and df is number of documents in which the term is present



**Calculating weight for query**

Wq = tf \* idf

= (1 + log10(tft,d) ) \* log10(N/df)

**Calculating weight for document**

Wd = tf

= 1 + log10(tft,d)

We don’t take idf into consideration as idf for same is used in query weighting and also increases efficiency for compressed indices

**The weight for query and document is cosine normalized by multiplying with**



Since the weight for query and document for a term is normalized.

We multiple normalized query weight and normalized document of term. We then add all the score for all the terms in document from query to get final score of the document.

The approach was used as it provided the better results and also most recommended

<https://www.youtube.com/watch?v=E3shpvJUZ84>

<http://nlp.stanford.edu/IR-book/html/htmledition/queries-as-vectors-1.html>

<https://en.wikipedia.org/wiki/Vector_space_model>

<http://www.ics.uci.edu/~djp3/classes/2008_09_26_CS221/Lectures/Lecture26.pdf>

<https://www.kth.se/social/files/56c4d886f2765449d27d5407/lecture05.pdf>

<https://www.cl.cam.ac.uk/teaching/1314/InfoRtrv/lecture4.pdf>

CMS text book page 261-267

MRS text book page 153

Other approaches which were referred and not chosen.

<https://courses.cs.washington.edu/courses/cse573/12sp/lectures/17-ir.pdf>

<https://pt.coursera.org/learn/big-data-management/lecture/0GMs1/vector-space-model>

<http://www.tfidf.com/>

<https://janav.wordpress.com/2013/10/27/tf-idf-and-cosine-similarity/>