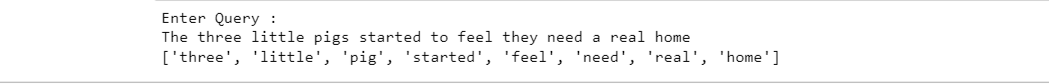
IR Assignment 2 Analysis

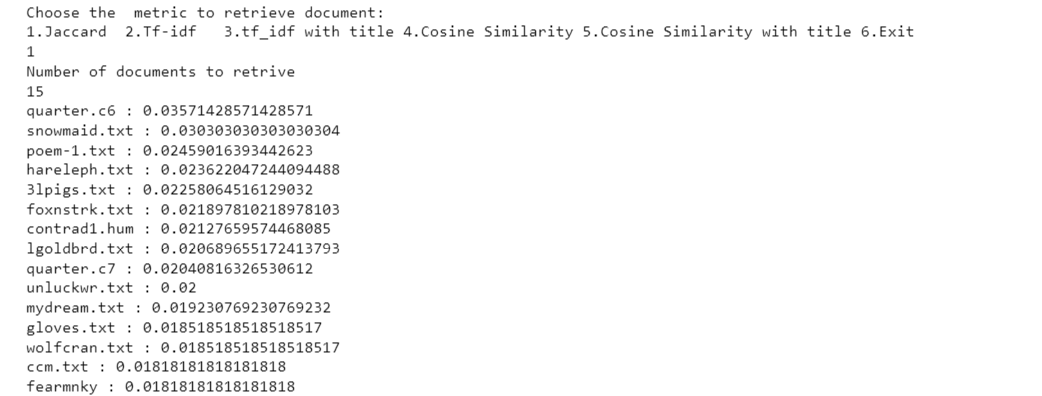
-Shalini Bhardwaj MT19045

Q1

Analysis on various metrics is done on following query:



1.jacarad Coefficient



The above query is from 3lpigs.txt and 3lpigs.txt is on 5th position.

Pros

Jaccard coefficient is easy to implement.

Cons

Jaccard coefficient also does not consider context. Jaccard coefficient does not provide special score to the document in which term exists in title.

It doesn’t consider term frequency.

Rare terms are more informative than frequent terms. Jaccard does not consider this information.

It gives more false positive results.

2.TF-idf

Tf idf gives more score if term is more in one document and present in less number of documents. Various variants applied are:

1.(with tf raw count)

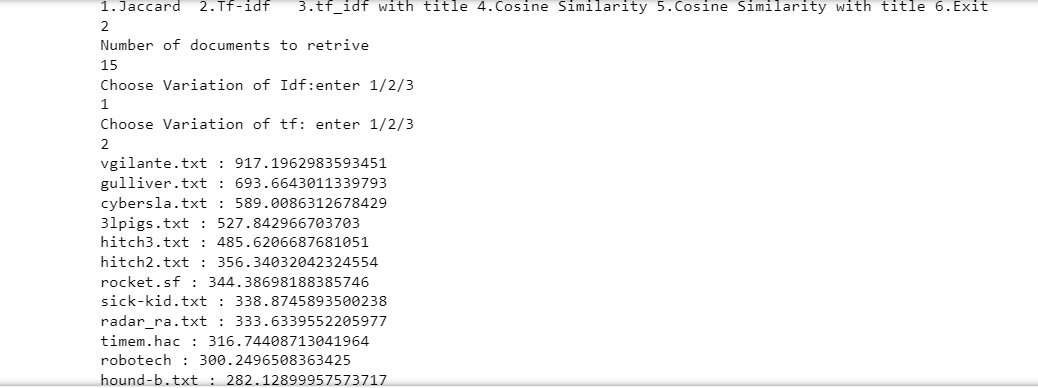
Idf is set to Number of total docs/frequency of term in docs

As tf is measured as just frequency/ count of words. This raw variant of tf does not consider the length of document .It tends to give more weightage to documents with large length. Therefore, count with relation to length of that document must be considered.

As here the document where this line belongs to comes at 1st position from 4th as biaseness towards .

Therefore false positive increase.

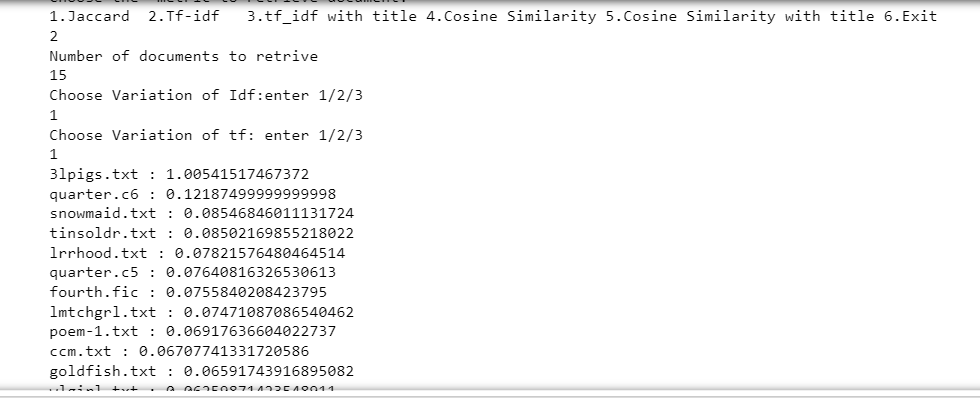
Here if valid query is present in document will small length then too it can be ranked later.



2. tf with normalization with doc length

Here tf is calculated as count/length of document.

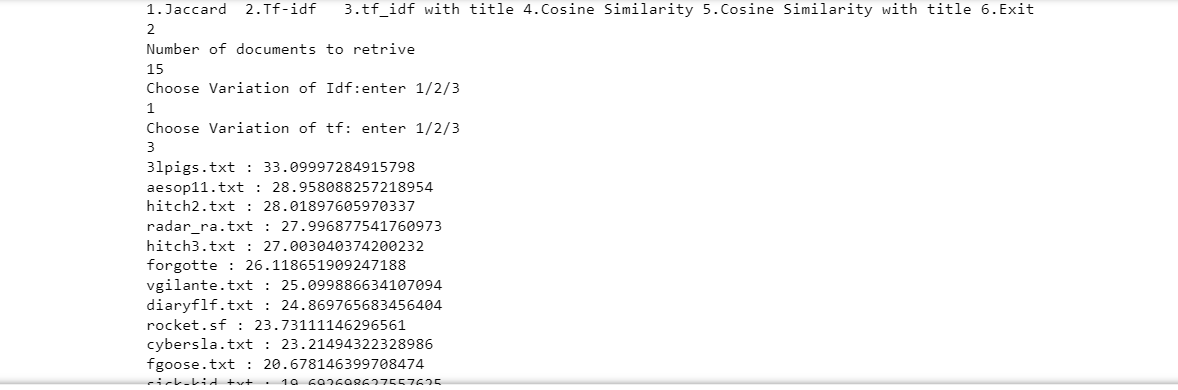
Here we can see document arrives at first .Normalization with length of document does not give false rank to larger documents.



3. tf with log scaling

Tf =1+log(1+count)

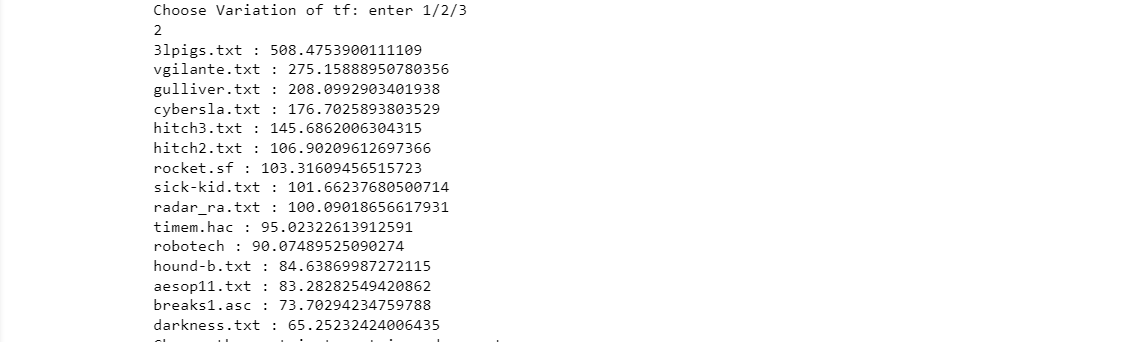
Here scaling is done by applying log function.We can see that 3lpigs document is present at first position. Hence better than raw count variant.



4. tf idf with title

To compare I have taken raw count variant with title weightage.Result shows if title is given more weightage then count tf also performs better than before.3lpigs comes at first position.

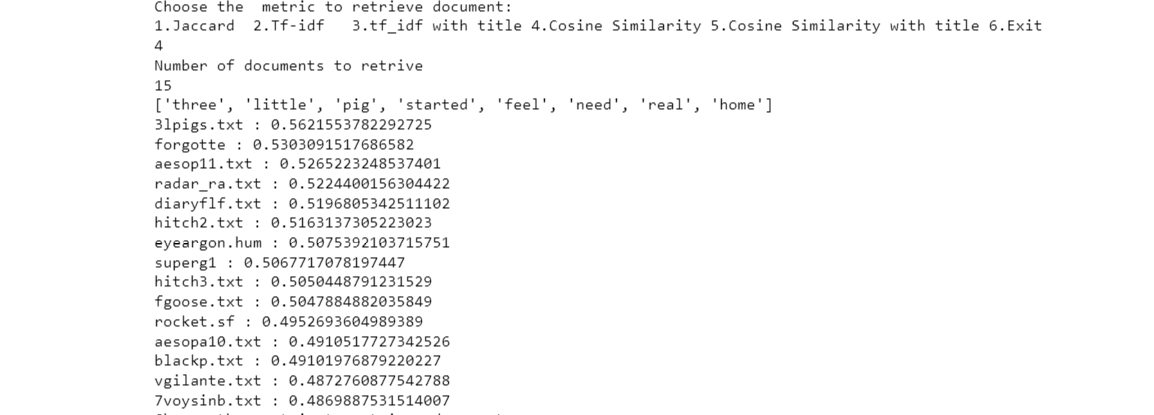
But in this case if query is relevant to doc1 but not present in doc1 title and some of the query terms are present another documents title then it will give more weightage to other document.



5.Cosine similarity

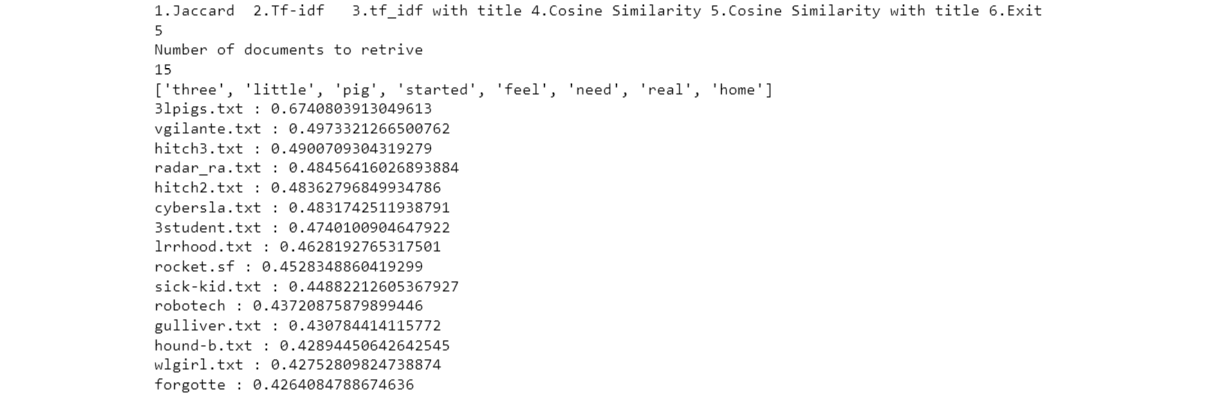
Cosine similarity works best as it take into consideration query vector which improvises the result of ranking. It considers the term frequency unlike jaccard.

It is somewhat difficult to implement.



6. Cosine with title

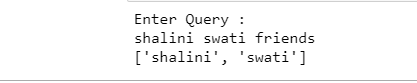
It could be seen than score of doc 3lpigs mainly increases as compared to when title was not considered.Doc vgliante comes at upper rank because some of query terms mightbe matching title of vgliante.



Ques2

Edit distance

Query on which analysis is done:



Here shalini and swati are terms which are not in dictionary. Therefore for them edit distance from all dictionary words is taken out. As user input he needs top 5 closest words therefore those are displayed.

