IR A2 ANALYSIS FILE

Question 1:

Part 1: -

The Jaccard score is not a good measure for document retrieval as it does not consider the count of the term in the document and only takes into account whether the term is present in the document. Usually the less-frequent terms in a document are more informative about the information present in a document. This is not taken care of in Jaccard score and that is why tf-idf is used.

Also, there is a bias based on the length of the document.

Part 2: -

On implementing different versions of tf-idf it is analyzed that the variant double norm 0.5 tf and inverse document frequency max performs best.

The pros and cons of different tf techniques are:

1. Binary:

   Pros- It has no pros as compared to other techniques

   Cons- It does not take into consideration the term frequency of the terms. So, in case the less frequent term is more informative about the document then that case cannot be handled

2. Raw Count:

   Pros- Easy to calculate

   Cons- Here we are using tf to find the query document matching score. But actually, the relevance of a document with respect to a query does not increase in the same proportion as the count of a term. So, it is not a good tf to use.

3. Term Frequency:

   Pros- Here the term frequency is normalized by dividing it with the total no. of tokens in the document. On doing this it is taken care of the fact that the term does not become more relevant in larger documents as the larger documents generally contain some terms repetitively.

   Cons- It is dependent on the document size.

4. Log Normalization:

   Pros- It takes into account the term frequencies in the document and also applies smoothening by doing log(1+f(t,d)) so a term which is not present in the document has value = 0. Also it takes care of the fact that relevance of a document does not increase in the same proportion of the counts by taking log.

Cons- It is a bit biased towards the documents which are larger in size as it uses term frequencies only.

5. Double normalization 0.5:

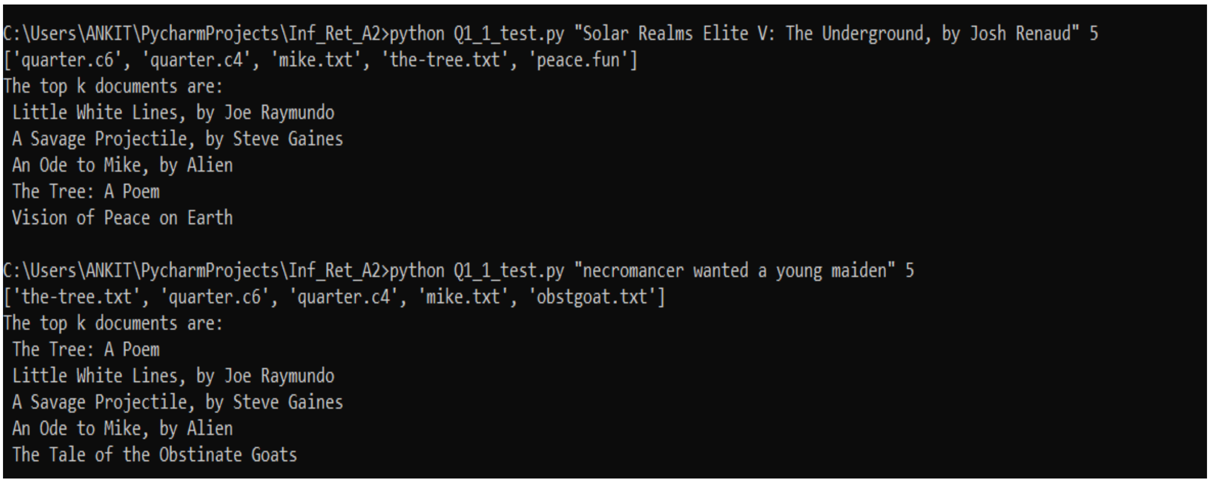
   Pros- Now the values are brought in the range of 0 to 1. This method tries to avoid the bias towards longer documents by dividing the term frequency by the term frequency of the term which occurs most in the document.

   Also, if a new document is A’ is created by appending document A to itself. According to this method the 2 documents A’ and A are equivalent in relevance as the tf weights assigned to the terms are same for A’ and A.

   Cons- A document may contain one term with the maximum frequency among all terms in the document but the term does not tell anything about the content of that document.

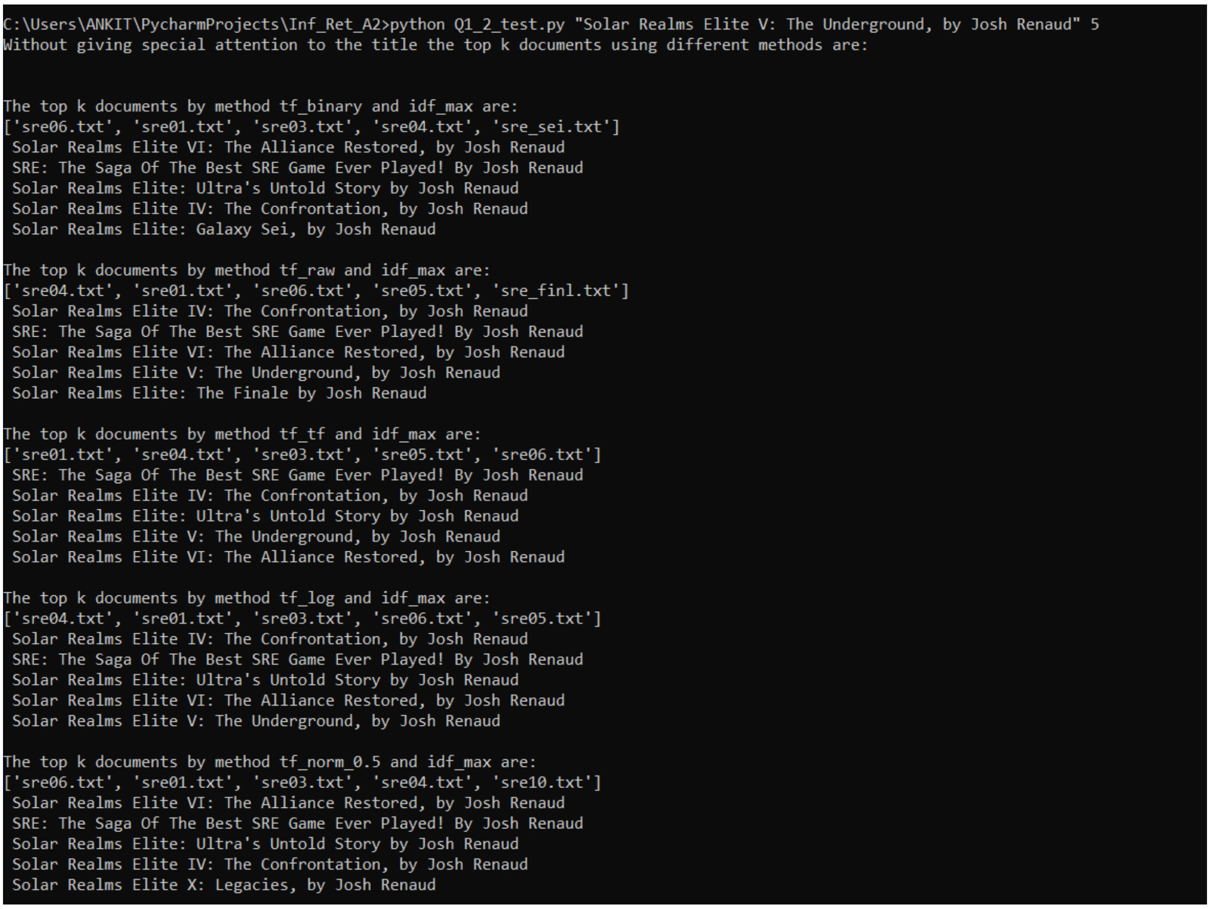
Analysis:

1. The documents are retrieved for the query based on Jaccard co-efficient.
   First in the query the title of a document is given and as can be seen that document is not present in the top 5 docs retrieved.
   Second a phrase is taken from the file knuckle.txt but that doc is not retrieved.
   So, Jaccard is not a good measure compared to tf-idf.



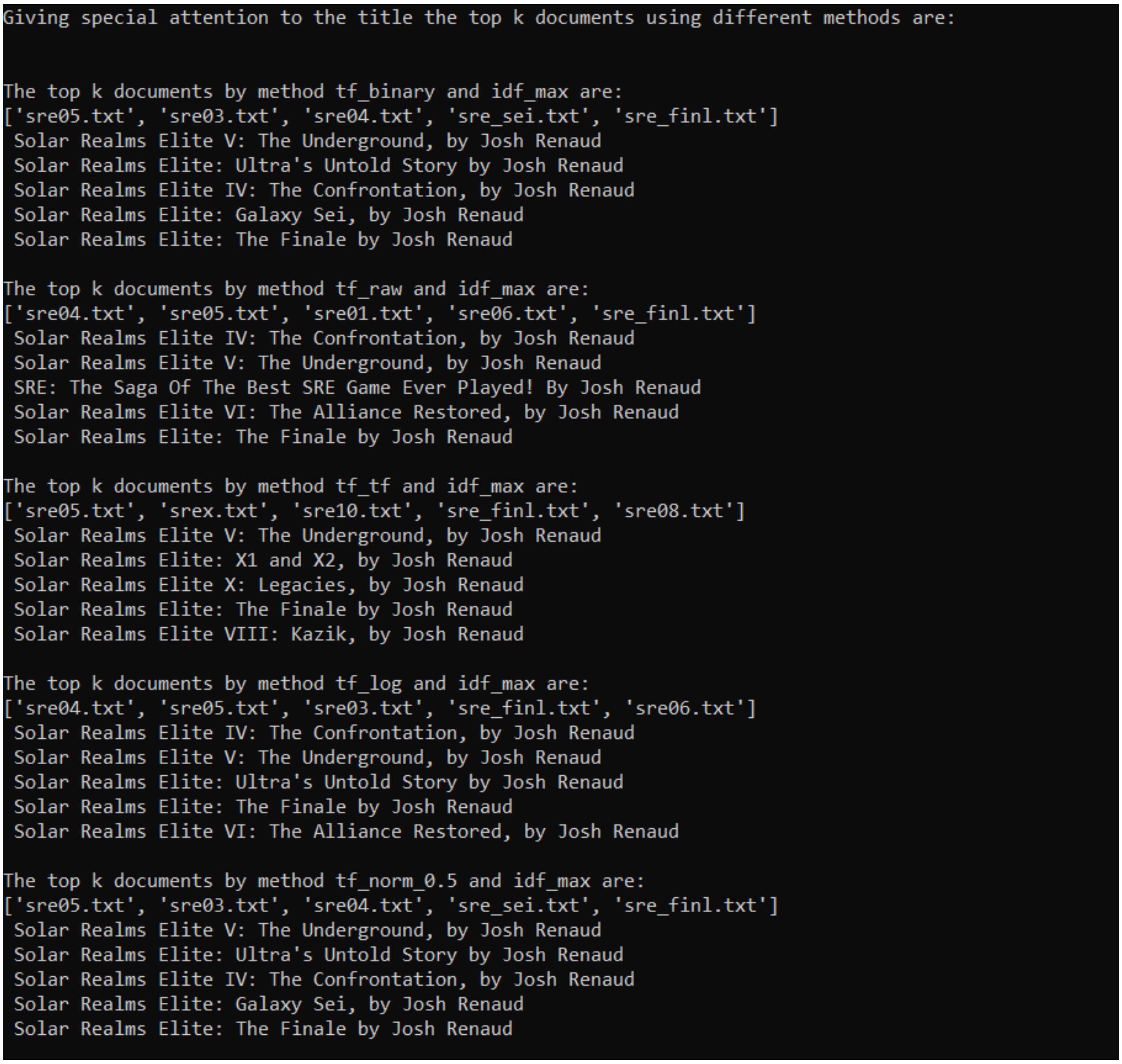
```
C:\Users\ANKIT\PycharmProjects\Inf_Ret_A2>python Q1_1_test.py "Solar Realms Elite V: The Underground, by Josh Renaud" 5
['quarter.c6', 'quarter.c4', 'mike.txt', 'the-tree.txt', 'peace.fun']
The top k documents are:
 Little White Lines, by Joe Raymundo
 A Savage Projectile, by Steve Gaines
 An Ode to Mike, by Alien
 The Tree: A Poem
 Vision of Peace on Earth

C:\Users\ANKIT\PycharmProjects\Inf_Ret_A2>python Q1_1_test.py "necromancer wanted a young maiden" 5
['the-tree.txt', 'quarter.c6', 'quarter.c4', 'mike.txt', 'obstgoat.txt']
The top k documents are:
 The Tree: A Poem
 Little White Lines, by Joe Raymundo
 A Savage Projectile, by Steve Gaines
 An Ode to Mike, by Alien
 The Tale of the Obstinate Goats
```

2. Here for part 2 in the query the title of a document is given and the top 5 docs are retrieved using different tf-idf methods without giving special attention to the titles. As can be seen here none of the lists conatin the queried document on the first position.



```
C:\Users\ANKIT\PycharmProjects\Inf_Ret_A2>python Q1_2_test.py "Solar Realms Elite V: The Underground, by Josh Renaud" 5
Without giving special attention to the title the top k documents using different methods are:


The top k documents by method tf_binary and idf_max are:
['sre06.txt', 'sre01.txt', 'sre03.txt', 'sre04.txt', 'sre_sei.txt']
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite: Galaxy Sei, by Josh Renaud

The top k documents by method tf_raw and idf_max are:
['sre04.txt', 'sre01.txt', 'sre06.txt', 'sre05.txt', 'sre_finl.txt']
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud

The top k documents by method tf_tf and idf_max are:
['sre01.txt', 'sre04.txt', 'sre03.txt', 'sre05.txt', 'sre06.txt']
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud

The top k documents by method tf_log and idf_max are:
['sre04.txt', 'sre01.txt', 'sre03.txt', 'sre06.txt', 'sre05.txt']
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud
 Solar Realms Elite V: The Underground, by Josh Renaud

The top k documents by method tf_norm_0.5 and idf_max are:
['sre06.txt', 'sre01.txt', 'sre03.txt', 'sre04.txt', 'sre10.txt']
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite X: Legacies, by Josh Renaud
```

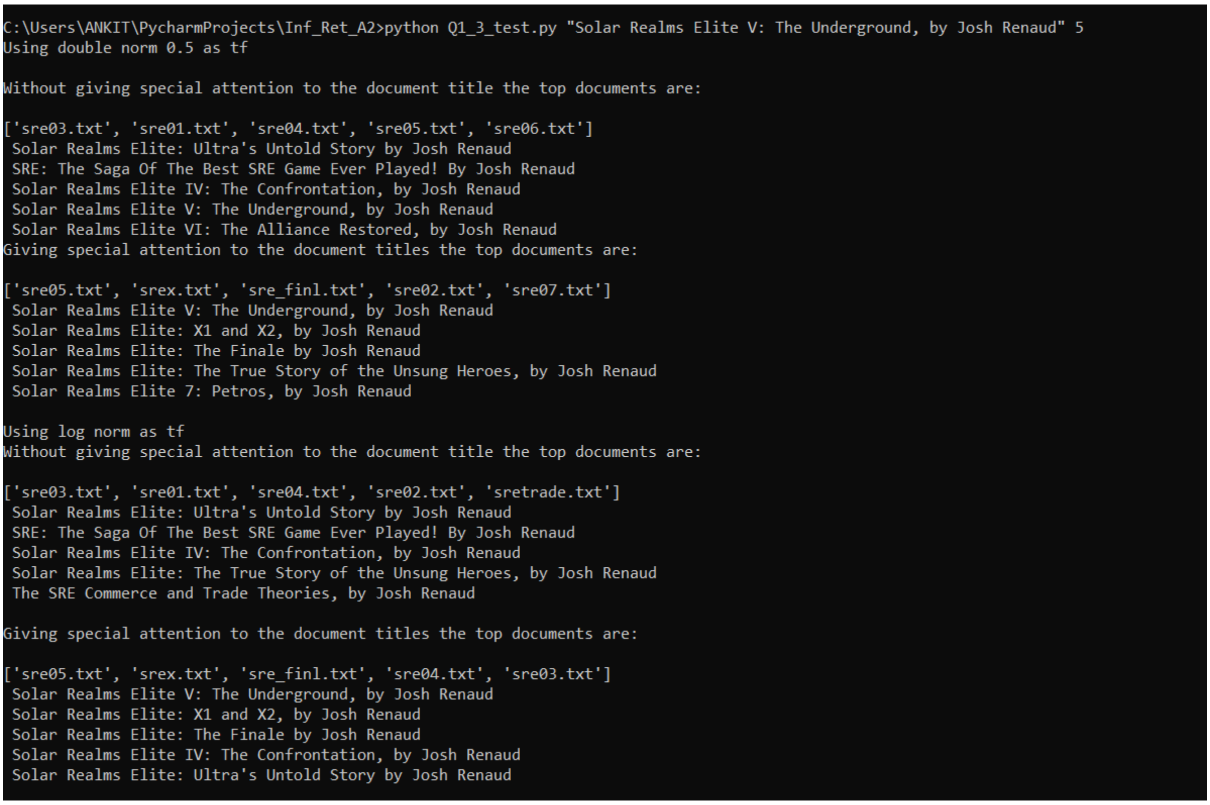
Now giving special attention to titles the docs retrieved are:



```
Giving special attention to the title the top k documents using different methods are:


The top k documents by method tf_binary and idf_max are:
['sre05.txt', 'sre03.txt', 'sre04.txt', 'sre_sei.txt', 'sre_finl.txt']
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite: Galaxy Sei, by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud

The top k documents by method tf_raw and idf_max are:
['sre04.txt', 'sre05.txt', 'sre01.txt', 'sre06.txt', 'sre_finl.txt']
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite V: The Underground, by Josh Renaud
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud

The top k documents by method tf_tf and idf_max are:
['sre05.txt', 'srex.txt', 'sre10.txt', 'sre_finl.txt', 'sre08.txt']
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite: X1 and X2, by Josh Renaud
 Solar Realms Elite X: Legacies, by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud
 Solar Realms Elite VIII: Kazik, by Josh Renaud

The top k documents by method tf_log and idf_max are:
['sre04.txt', 'sre05.txt', 'sre03.txt', 'sre_finl.txt', 'sre06.txt']
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud

The top k documents by method tf_norm_0.5 and idf_max are:
['sre05.txt', 'sre03.txt', 'sre04.txt', 'sre_sei.txt', 'sre_finl.txt']
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite: Galaxy Sei, by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud
```

As can be seen that tf-norm-0.5 method contains the queried doc on the top and also as discussed previously in the pros and cons this method is the best method.

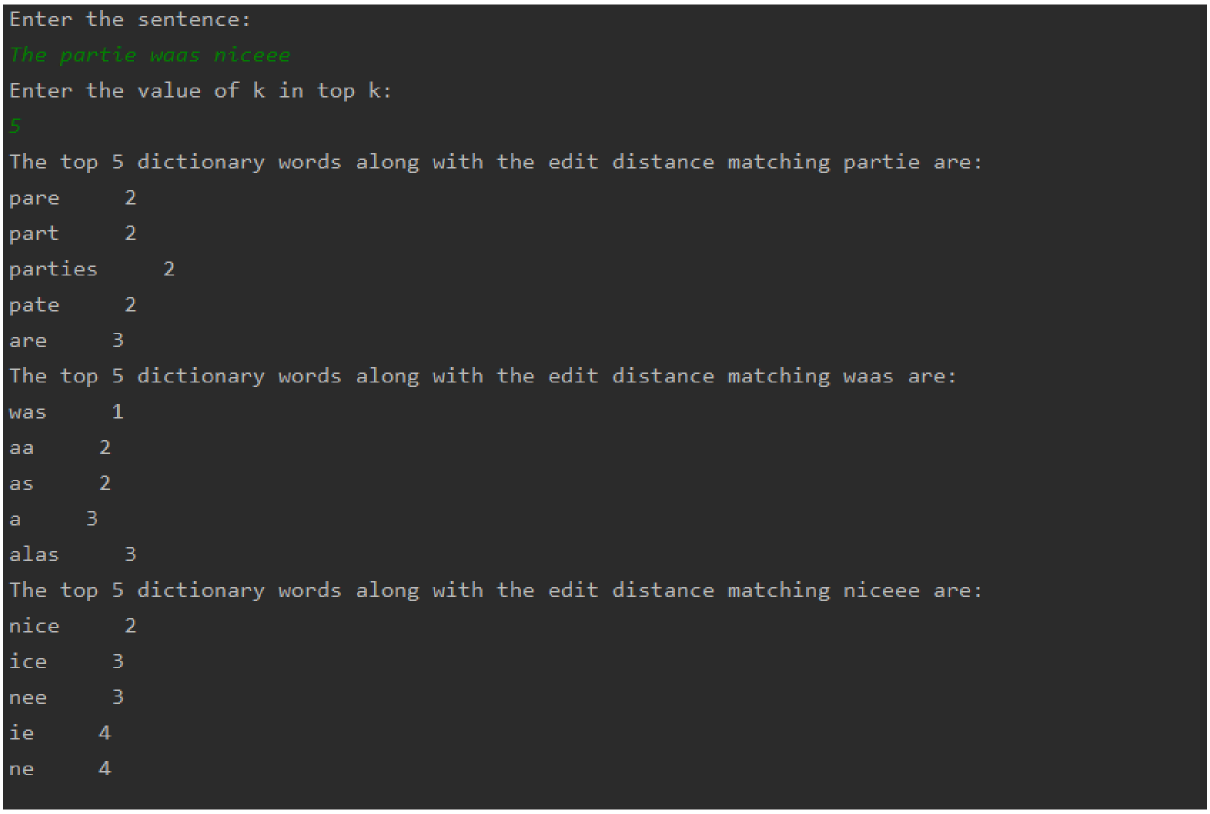
3. Cosine sim



```
C:\Users\ANKIT\PycharmProjects\Inf_Ret_A2>python Q1_3_test.py "Solar Realms Elite V: The Underground, by Josh Renaud" 5
Using double norm 0.5 as tf

Without giving special attention to the document title the top documents are:

['sre03.txt', 'sre01.txt', 'sre04.txt', 'sre05.txt', 'sre06.txt']
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite VI: The Alliance Restored, by Josh Renaud
Giving special attention to the document titles the top documents are:

['sre05.txt', 'srex.txt', 'sre_finl.txt', 'sre02.txt', 'sre07.txt']
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite: X1 and X2, by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud
 Solar Realms Elite: The True Story of the Unsung Heroes, by Josh Renaud
 Solar Realms Elite 7: Petros, by Josh Renaud

Using log norm as tf
Without giving special attention to the document title the top documents are:

['sre03.txt', 'sre01.txt', 'sre04.txt', 'sre02.txt', 'sretrade.txt']
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
 SRE: The Saga Of The Best SRE Game Ever Played! By Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite: The True Story of the Unsung Heroes, by Josh Renaud
 The SRE Commerce and Trade Theories, by Josh Renaud

Giving special attention to the document titles the top documents are:

['sre05.txt', 'srex.txt', 'sre_finl.txt', 'sre04.txt', 'sre03.txt']
 Solar Realms Elite V: The Underground, by Josh Renaud
 Solar Realms Elite: X1 and X2, by Josh Renaud
 Solar Realms Elite: The Finale by Josh Renaud
 Solar Realms Elite IV: The Confrontation, by Josh Renaud
 Solar Realms Elite: Ultra's Untold Story by Josh Renaud
```

On giving special attention to title we get the queried doc in the first rank while in case we do not give special attention to the title then also we get the doc but not on the first rank.

Q2.) Edit distance



```
Enter the sentence:
The partie waas niceee
Enter the value of k in top k:
5
The top 5 dictionary words along with the edit distance matching partie are:
pare     2
part     2
parties     2
pate     2
are     3
The top 5 dictionary words along with the edit distance matching waas are:
was     1
aa     2
as     2
a     3
alas     3
The top 5 dictionary words along with the edit distance matching niceee are:
nice     2
ice     3
nee     3
ie     4
ne     4
```