IR – FINAL PROJECT

Name : - E.P.S.Santhosh

Roll Number :- S20180010053

Project Topic :-News .

Implementation :-

Step -1) Read the data set .

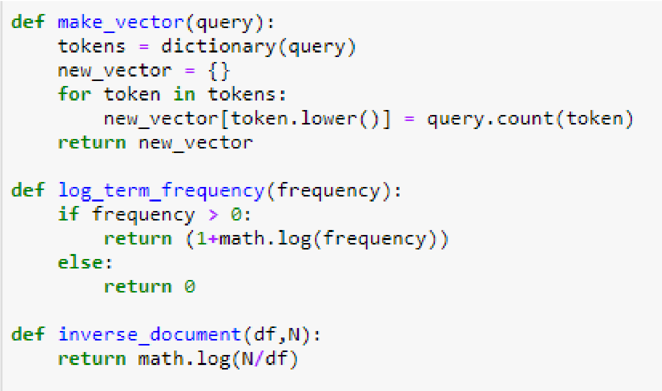
Step -2) Create an inverted index for the given documents from scratch without any inbuilt libraries .

Step -3) I used pickle library in order to save the created inverted index and save the time .

1) Task -1 :-

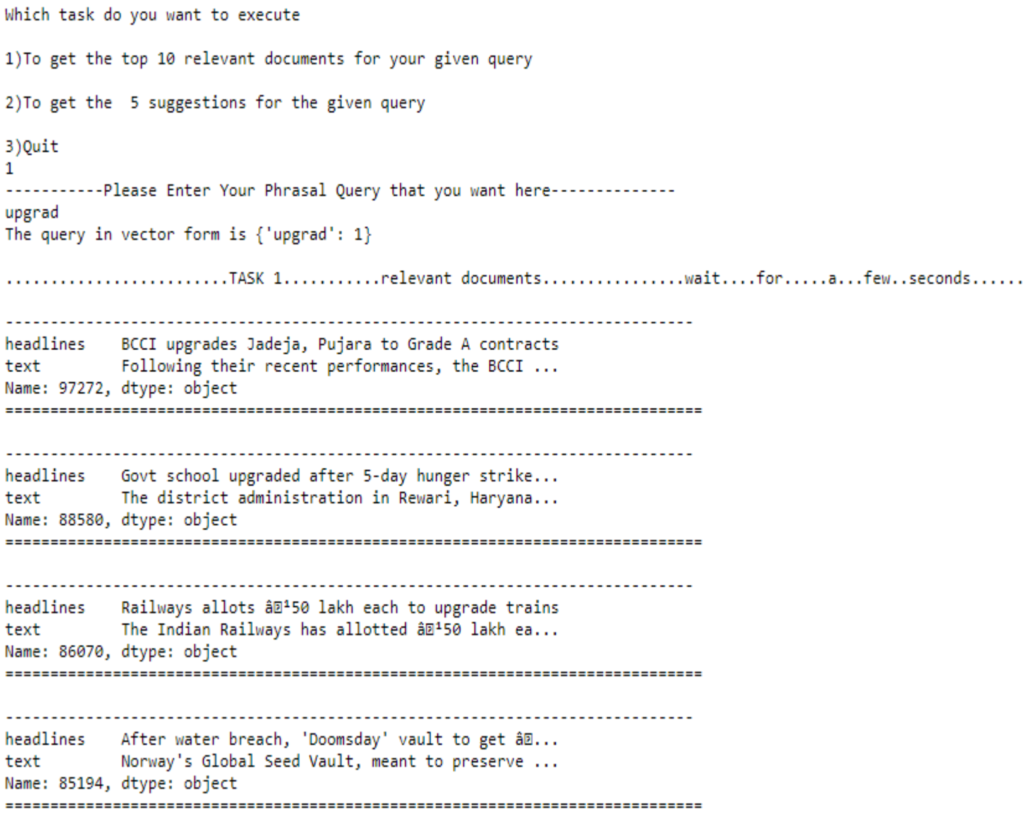
Ranked Retrival for the given query

- Used the vector space model for representing a documents and query .
- Query will be represented as the vector with each dimension as each word .
- When the query is given we will calculate the similarity score between the each document with the query as shown below.



```
def make_vector(query):
    tokens = dictionary(query)
    new_vector = {}
    for token in tokens:
        new_vector[token.lower()] = query.count(token)
    return new_vector

def log_term_frequency(frequency):
    if frequency > 0:
        return (1+math.log(frequency))
    else:
        return 0

def inverse_document(df,N):
    return math.log(N/df)
```

- I multiplied the Tf-idf scores obtained from between each common term between query and document and got the resultant score .
- Now I collected the score and based on the score I sorted them .
- Now I printed the top-10 relevant documents with the high score .

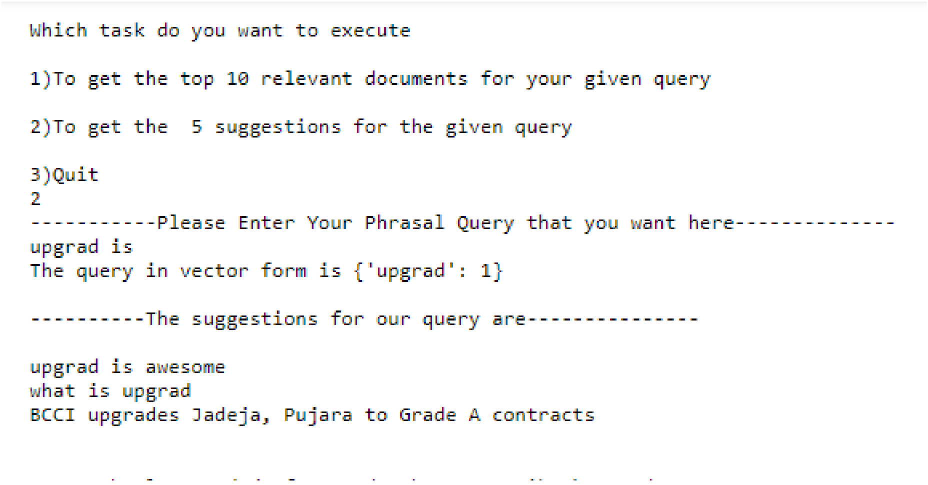


```
Which task do you want to execute

1)To get the top 10 relevant documents for your given query

2)To get the  5 suggestions for the given query

3)Quit
1
----------Please Enter Your Phrasal Query that you want here-------------
upgrad
The query in vector form is {'upgrad': 1}

.........................TASK 1...........relevant documents................wait....for.....a...few..seconds......

-----------------------------------------------------------------------------
headlines    BCCI upgrades Jadeja, Pujara to Grade A contracts
text         Following their recent performances, the BCCI ...
Name: 97272, dtype: object
==============================================================================

-----------------------------------------------------------------------------
headlines    Govt school upgraded after 5-day hunger strike...
text         The district administration in Rewari, Haryana...
Name: 88580, dtype: object
==============================================================================

-----------------------------------------------------------------------------
headlines    Railways allots â▯¹50 lakh each to upgrade trains
text         The Indian Railways has allotted â▯¹50 lakh ea...
Name: 86070, dtype: object
==============================================================================

-----------------------------------------------------------------------------
headlines    After water breach, 'Doomsday' vault to get â▯...
text         Norway's Global Seed Vault, meant to preserve ...
Name: 85194, dtype: object
==============================================================================
```

2)Task-2 :-

User Suggestions

- Used the vector space model for representing a documents and query .
- Query will be represented as the vector with each dimension as each word .
- We have used previous queries used by the user in our search engine and used query expansion for getting the suggestions .



```
Which task do you want to execute

1)To get the top 10 relevant documents for your given query

2)To get the  5 suggestions for the given query

3)Quit
2
-----------Please Enter Your Phrasal Query that you want here--------------
upgrad is
The query in vector form is {'upgrad': 1}

----------The suggestions for our query are---------------

upgrad is awesome
what is upgrad
BCCI upgrades Jadeja, Pujara to Grade A contracts
```

3)Succesfully implemented Task -1 and Task-2 as shown above .