# IR Assignment -2

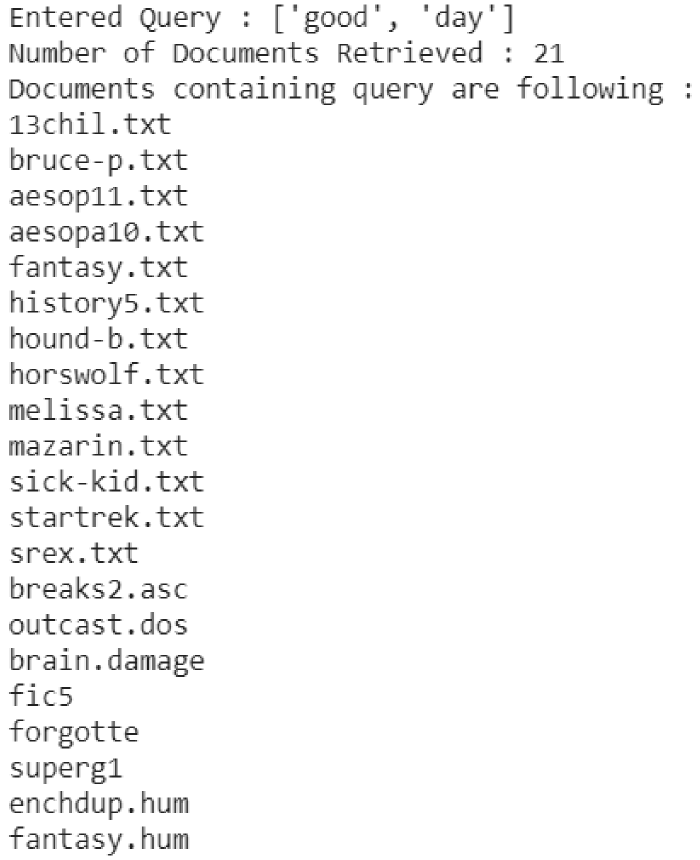
# Group Number 22

# Analysis

## Ans 1)

**Query:** "GOOD DAYS"

**Output :** following are the results obtained



```
Entered Query : ['good', 'day']
Number of Documents Retrieved : 21
Documents containing query are following :
13chil.txt
bruce-p.txt
aesop11.txt
aesopa10.txt
fantasy.txt
history5.txt
hound-b.txt
horswolf.txt
melissa.txt
mazarin.txt
sick-kid.txt
startrek.txt
srex.txt
breaks2.asc
outcast.dos
brain.damage
fic5
forgotte
superg1
enchdup.hum
fantasy.hum
```

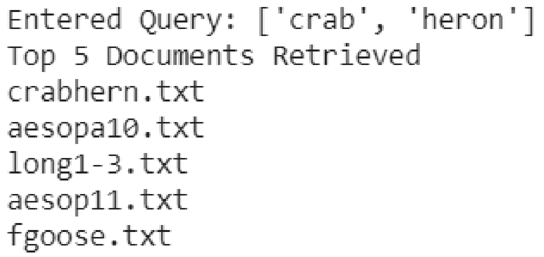
Observation: To evaluate the correctness of the Output, Manually searching from the above documents containing “good days” was done. The Documents retrieved did contain the words either Good days or good-day.

## Ans 2)

## JACCARD :

**Query** : "THE CRAB AND THE HERON"

**Output:** Following are the top 5 documents retrieved-

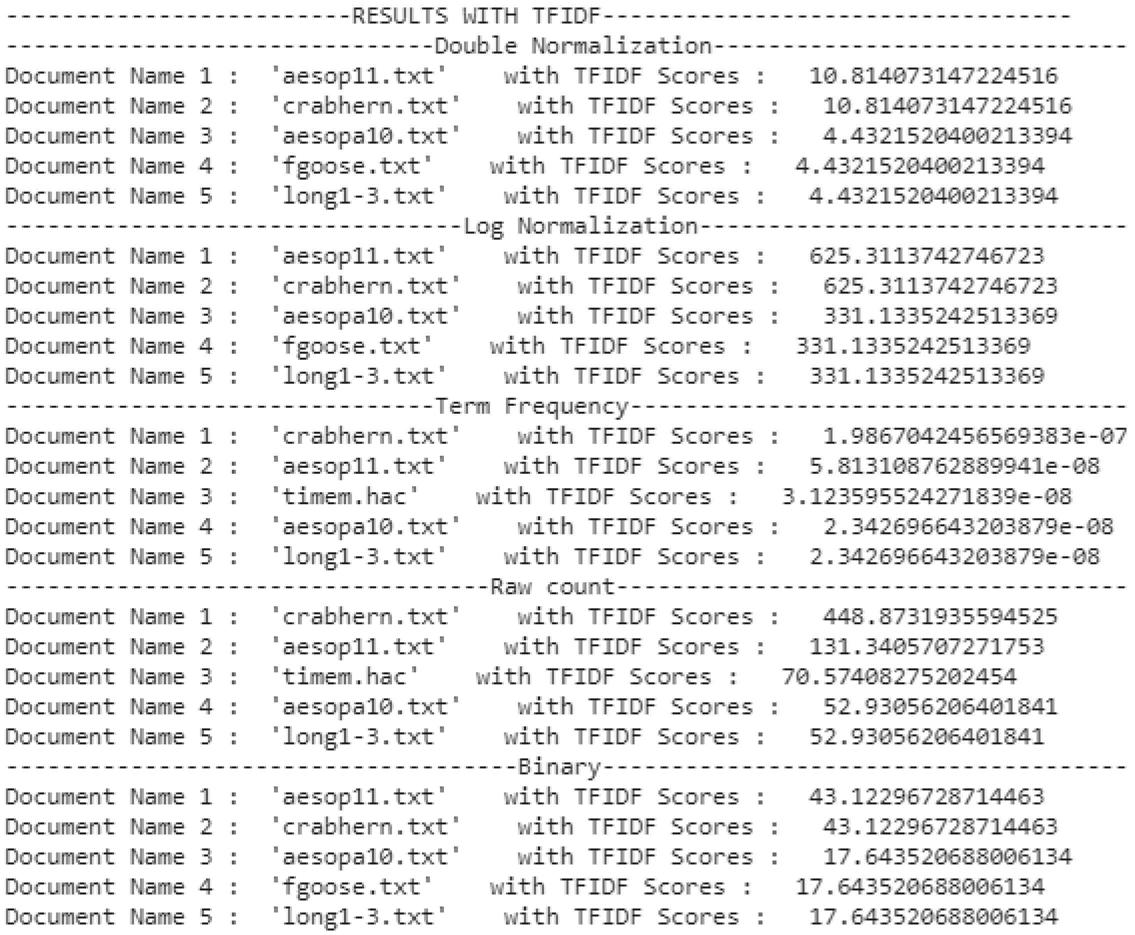


```
Entered Query: ['crab', 'heron']
Top 5 Documents Retrieved
crabhern.txt
aesopa10.txt
long1-3.txt
aesop11.txt
fgoose.txt
```

## TFIDF :

**Query :** "THE CRAB AND THE HERON"

**Output :** following are the results obtained



```
-------------------------RESULTS WITH TFIDF---------------------------------
-------------------------------Double Normalization-----------------------------
Document Name 1 :  'aesop11.txt'    with TFIDF Scores :   10.814073147224516
Document Name 2 :  'crabhern.txt'    with TFIDF Scores :   10.814073147224516
Document Name 3 :  'aesopa10.txt'    with TFIDF Scores :   4.4321520400213394
Document Name 4 :  'fgoose.txt'    with TFIDF Scores :   4.4321520400213394
Document Name 5 :  'long1-3.txt'    with TFIDF Scores :   4.4321520400213394
---------------------------------Log Normalization------------------------------
Document Name 1 :  'aesop11.txt'    with TFIDF Scores :   625.3113742746723
Document Name 2 :  'crabhern.txt'    with TFIDF Scores :   625.3113742746723
Document Name 3 :  'aesopa10.txt'    with TFIDF Scores :   331.1335242513369
Document Name 4 :  'fgoose.txt'    with TFIDF Scores :   331.1335242513369
Document Name 5 :  'long1-3.txt'    with TFIDF Scores :   331.1335242513369
-------------------------------Term Frequency-----------------------------------
Document Name 1 :  'crabhern.txt'    with TFIDF Scores :   1.9867042456569383e-07
Document Name 2 :  'aesop11.txt'    with TFIDF Scores :   5.813108762889941e-08
Document Name 3 :  'timem.hac'    with TFIDF Scores :   3.123595524271839e-08
Document Name 4 :  'aesopa10.txt'    with TFIDF Scores :   2.342696643203879e-08
Document Name 5 :  'long1-3.txt'    with TFIDF Scores :   2.342696643203879e-08
-----------------------------------Raw count------------------------------------
Document Name 1 :  'crabhern.txt'    with TFIDF Scores :   448.8731935594525
Document Name 2 :  'aesop11.txt'    with TFIDF Scores :   131.3405707271753
Document Name 3 :  'timem.hac'    with TFIDF Scores :   70.57408275202454
Document Name 4 :  'aesopa10.txt'    with TFIDF Scores :   52.93056206401841
Document Name 5 :  'long1-3.txt'    with TFIDF Scores :   52.93056206401841
-------------------------------------Binary-------------------------------------
Document Name 1 :  'aesop11.txt'    with TFIDF Scores :   43.12296728714463
Document Name 2 :  'crabhern.txt'    with TFIDF Scores :   43.12296728714463
Document Name 3 :  'aesopa10.txt'    with TFIDF Scores :   17.643520688006134
Document Name 4 :  'fgoose.txt'    with TFIDF Scores :   17.643520688006134
Document Name 5 :  'long1-3.txt'    with TFIDF Scores :   17.643520688006134
```

Observation : Out of 5 files most relevant files retrieved in all matrices 3 files are common in all five kinds of matrix and that are aesop11,fgoose,long1-3.

When manually checked these files are actually most relevant due to the count of words present in query. All the 3 similarities gave similar results.

Advantages

1. Binary : for this tf scheme only presence and absence of word is considered , thus it is the simplest one to be computed.

2. Raw Count : In this scheme raw count of words in documents is determined.Disadvantage of binary scheme is overcomed in this scheme. Frequency of words is taken into account thus comparatively more relevant documents are retrieved.

3. Term Frequency : In this scheme disadvantage of above scheme is overcomed.Frequency of words is normalized using length of the document thus reducing the biases caused by the length of the documents.

4. Long Normalization : in this scheme raw count is normalized by taking log of count and 1 to it. A significant smaller number is obtained , thus reducing the computational power.

5. Double Normalization : this scheme involves both normalization using length of the doc and the long normalization. Advantage in this scheme is that it reduces the computation power and also considers the term frequency that is normalized using document length.

Disadvantages

6. Binary : It does not consider the frequency of term which is quite an important factor in determining the most relevant documents.

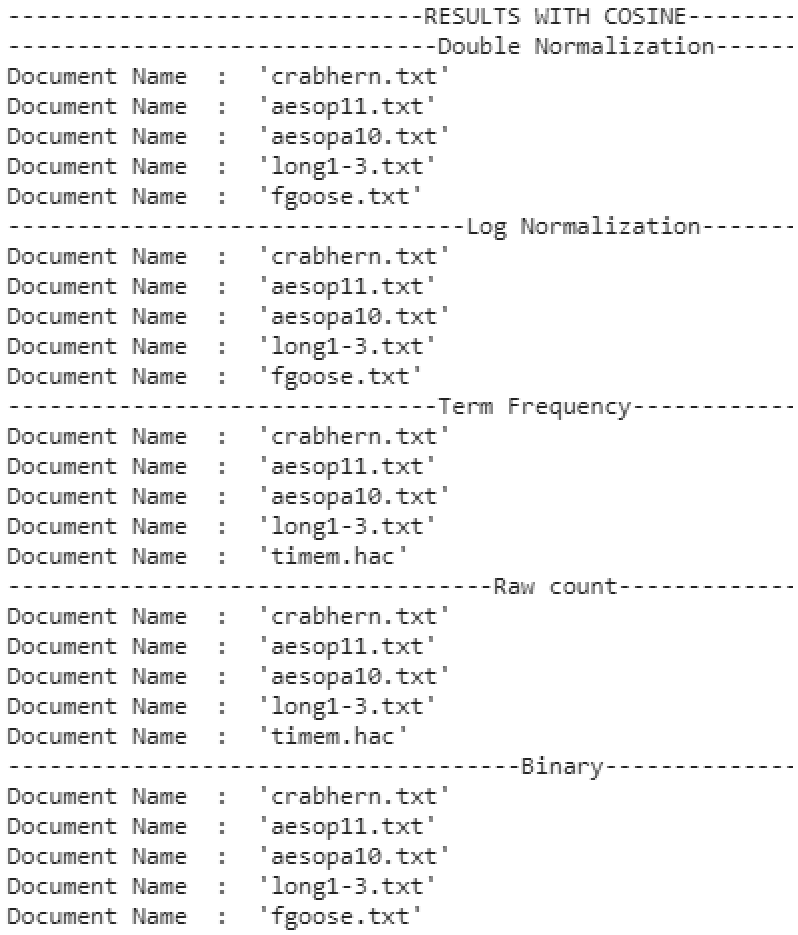
7. Raw Count : In this scheme larger documents are more favoured than comparatively shorter ones, which should not be the case. Also for very large documents huge(that would in turn require more storage space ) numbers need to be stored.

8. Term Frequency : Numbers obtained are still huge causing requirement of more storage space

## COSINE SIMILARITY :

**Query :** "THE CRAB AND THE HERON"

**Output :** following are the results obtained



```
------------------------------RESULTS WITH COSINE--------
-------------------------------Double Normalization------
Document Name  :  'crabhern.txt'
Document Name  :  'aesop11.txt'
Document Name  :  'aesopa10.txt'
Document Name  :  'long1-3.txt'
Document Name  :  'fgoose.txt'
---------------------------------Log Normalization-------
Document Name  :  'crabhern.txt'
Document Name  :  'aesop11.txt'
Document Name  :  'aesopa10.txt'
Document Name  :  'long1-3.txt'
Document Name  :  'fgoose.txt'
-------------------------------Term Frequency------------
Document Name  :  'crabhern.txt'
Document Name  :  'aesop11.txt'
Document Name  :  'aesopa10.txt'
Document Name  :  'long1-3.txt'
Document Name  :  'timem.hac'
-----------------------------------Raw count-------------
Document Name  :  'crabhern.txt'
Document Name  :  'aesop11.txt'
Document Name  :  'aesopa10.txt'
Document Name  :  'long1-3.txt'
Document Name  :  'timem.hac'
-------------------------------------Binary--------------
Document Name  :  'crabhern.txt'
Document Name  :  'aesop11.txt'
Document Name  :  'aesopa10.txt'
Document Name  :  'long1-3.txt'
Document Name  :  'fgoose.txt'
```

## Ans 3)

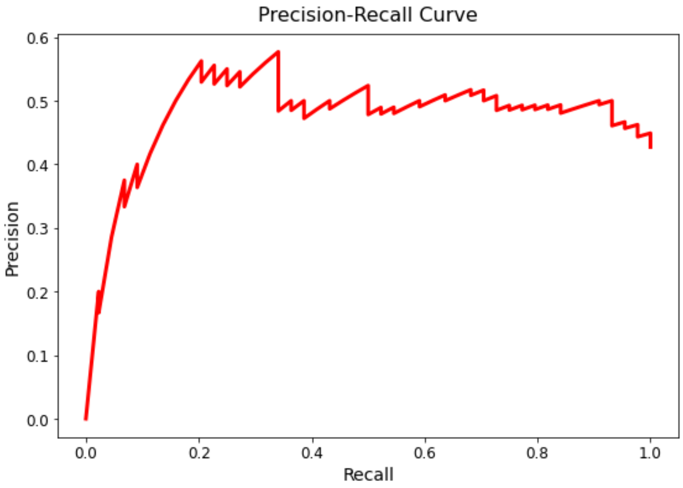
### Outputs -

No of query url pair with query id 4 are: **103**

Normalized Discounted Cumulative Gain
For whole dataset: **0.578**
At rank 50: **0.356**

Maximum Discounted Cumulative Gain: **28.988**

Number of files that could be made: **59! * 26! * 17! * 1!**



### Analysis –

The precision-recall curve is a distinctive sawtooth shape.