ANALYSIS

Submitted By :

Group No.-3

Atul	atul17032@iiitd.ac.in
Prashant Jain	prashant18253@iiitd.ac.in
Akanksha Pandey	akanksha20048@iiitd.ac.in
Shivani Mishra	shivani20062@iiitd.ac.in

Pros and cons of Scoring Scheme:

1. **Jaccard Coefficient:**

   Pros:

   - If a word comes more than one time in a document, it will not affect the Jaccard coefficient. So it works better for the analysis where duplicate data does not matter.
   - It can also be used to identify mirror sites

   Cons:

   - Rare terms in a collection are more informative than frequent terms. Jaccard doesn't consider this information.

2. **Tfidf:**

   Pros:

   - It is very easy to compute
   - It has some basic metric to extract the most descriptive terms in a document
   - It can easily compute the similarity between 2 documents or 1 query over each document

Cons:

- Tfidf is just based on the Bag of words model, therefore it can not measure the semantic information of the words in the sentence.
- It measures the presence and absence of words in the document without knowing any meaning.

3. **Cosine Similarity:**

Pros:

- Cosine similarity is one of the most widely used and powerful similarity measures in Data Science.
- It is used in multiple applications such as finding similar documents in NLP, information retrieval, finding similar sequences to DNA in bioinformatics, detecting plagiarism and many more.
- There are also computational benefits associated with this, as sum(x*y) is cheaper to compute for sparse data.
- Can be used for plagiarism checks.

Cons:

- The difference in rating scale between different users is not taken into account.
- Cosine similarity is less optimal under spaces of lower dimension.

Outputs:

Question 01 :

```
Enter query good day
['good', 'day']
total number of documents are  21
The documents are
13chil.txt
aesopa10.txt
aesop11.txt
brain.damage
bruce-p.txt
breaks2.asc
enchdup.hum
fantasy.hum
fic5
fantasy.txt
forgotte
hound-b.txt
history5.txt
horswolf.txt
melissa.txt
mazarin.txt
outcast.dos
sick-kid.txt
startrek.txt
superg1
srex.txt
```

Question 02 :

a. Jaccard Coefficient

```
Input the query100 west by 50 north


peace.fun
(210, 0.02)
snowmaid.txt
(205, 0.0078125)
prince.art
(373, 0.006060606060606061)
campfire.txt
(391, 0.00546448087431694)
glimpse1.txt
(421, 0.004975124378109453)
```

b. Tfidf

Input query : good day

a. Binary

```
sre04.txt
(467, 1.2118993284791508)
srex.txt
(466, 1.2118993284791508)
sre_finl.txt
(464, 1.2118993284791508)
sre_sei.txt
(458, 1.2118993284791508)
sre_feqh.txt
(457, 1.2118993284791508)
```

b. Raw Count

```
gulliver.txt
(90, 122.97113137888769)
vgilante.txt
(449, 91.20558784272181)
hound-b.txt
(334, 80.82635511721014)
outcast.dos
(128, 71.77894036182423)
aesop11.txt
(309, 65.24416442627367)
```

c. Term Frequency

```
contrad1.hum
(14, 0.022910863821364 78)
blossom.pom
(151, 0.022813865696745084)
blasters.fic
(386, 0.017257620352738225)
horswolf.txt
(333, 0.016453252812124054)
clevdonk.txt
(379, 0.01643940322265455)
```

d. Log Normalization

```
gulliver.txt
(90, 8.08114582285416)
hound-b.txt
(334, 7.368314819268528)
vgilante.txt
(449, 7.238925621790946)
outcast.dos
(128, 7.162279898208043)
aesop11.txt
(309, 6.935602984126474)
```

e. Double Normalization

```
pepsi.degenerat
(414, 0.9786994645831164)
pepdegener.txt
(399, 0.9786994645831164)
brain.damage
(221, 0.9734518090085775)
7voysinb.txt
(185, 0.9219851728757058)
history5.txt
(370, 0.9045049964921101)
```

c. Cosine Similarity

Input Query: King Kong

BINARY MODEL

```
mario.txt
19.lws
lionwar.txt
lionmosq.txt
monkking.txt
```

RAW_COUNT MODEL

```
fable.txt
hop-frog.poe
monkking.txt
6ablemen.txt
pussboot.txt
```

TERM FREQUENCY MODEL

```
fable.txt
hop-frog.poe
monkking.txt
6ablemen.txt
pussboot.txt
```

LOG-NORMALISATION MODEL

```
monkking.txt
pussboot.txt
mario.txt
6ablemen.txt
lpeargrl.txt
```

DOUBLE-NORMALISATION MODEL

```
monkking.txt
mario.txt
pussboot.txt
lionwar.txt
19.lws
```

Question 03 :

Max DCG is: 20.989750804831452

nDCG at 50 : 0.35210427403248856

nDCG for document : 0.5979226516897828

Total 19893497375938370599826047614905329896936840170566570588205180312704857992695193482412686565431050240000000000000000000000 files can be made.