# Information Retrieval

# Assignment 2

**Dibyendu Roy Chaudhuri**
**MT19034**

# Q1) CLI tool

**Total number of words-** 55156

**Tf-Idf versions used:**

1. I have used 5 types of formula to calculate Tf-
    a. Binary weight
    b. Raw frequency weight
    c. Term frequency
    d. Log normalization
    e. Double normalization K

**a) Binary weight:**

**Pros:**

- Easy to compute

**Cons:**

- Give the same weightage to every word present in the document
- Does not take care of the frequency of a word in a document

**b) Raw Frequency weight:**

**Pros:**

- Easy to compute
- Take care about the frequency of a word in a document

**Cons:**

- Does not take care of the relative frequency of a word in a document

c) **Term Frequency:**

**Pros:**

- Take about the relative frequency of a word in a document.

**Cons:**

- High computations are needed

d) **Log normalization:**

**Pros:**

- Take care about the frequency of a word in a document

**Cons:**

- Does not take about the relative frequency of a word
- High computations are needed

e) **Double normalization:**

**Pros:**

- Take care about the frequency of a word in a document

**Cons:**

- Does not take about the relative frequency of a word
- High computations are needed

**Which method is best and why?**

According to my observations, Term frequency is the best method to calculate the Tf-Idf score of a word present in a document. It takes care of the relative frequency of a word present in a document which seemed to be a good method to give weight to a particular word. It helps us to give less to stop words like- and, the, a and etc and also give high weightage to rarely occurring words which can dominate the subject of a particular document.

**Attention to title:**

Zone indexing is used to give extra attention to the title. The value of g-weight is shifted from (0.6 to 1.0) to observe variations in result. With increased values of g-weight, we can mostly classify the documents using the title only. Cosine similarity scores of documents are becoming dependent upon words that are present in the title part only.

Without attention to the title, documents are being depending upon the Tf-Idf scores of the words present in that document.

## Output 1 (without attention to title) using term frequency

Enter the query- As I approached my outer door, I was amazed to see a key in it. For an instant I imagined that I had left my own there, but on feeling in my pocket I found that it was all right. The only duplicate which existed, so far as I knew, was that which belonged to my servant, Bannister -- a man who has looked after my room for ten years, and whose honesty is absolutely above suspicion. I found that the key was indeed his, that he had entered my room to know if I wanted tea, and that he had very carelessly left the key in the door when he came out. His visit to my room must have been within a very few minutes of my leaving it. His forgetfulness about the key would have mattered little upon any other occasion, but on this one day it has produced the most deplorable consequences.

How many document should be retrieved? 3

Jaccard Coefficient result: top 5 documents are-

1 ) Dataset/Stories and Fictions\toilet.s
2 ) Dataset/Stories and Fictions\wanderer.fun
3 ) Dataset/Stories and Fictions\gloves.txt

Tf-Idf based document retrieval result: top 5 documents are-

1 ) Dataset/Stories and Fictions\the-tree.txt
2 ) Dataset/Stories and Fictions\mydream.txt
3 ) Dataset/Stories and Fictions\greedog.txt

Cosine similarity result: top 5 documents are-

1 ) Dataset/Stories and Fictions\quarter.c10
2 ) Dataset/Stories and Fictions\3student.txt
3 ) Dataset/Stories and Fictions\lionmane.txt

**This query is taken from 3student.txt. Here cosine similarity gives the accurate answer in the top-2 document.**

## Output 2 (without attention to title) using double normalization

Enter the query- As I approached my outer door, I was amazed to see a key in it. For an instant I imagined that I had left my own there, but on feeling in my pocket I found that it was all right. The only duplicate which existed, so far as I knew, was that which belonged to my servant, Bannister -- a man who has looked after my room for ten years, and whose honesty is absolutely above suspicion. I found that the key was indeed his, that he had entered my room to know if I wanted tea, and that he had very carelessly left the key in the door when he came out. His visit to my room must have been within a very few minutes of my leaving it. His forgetfulness about the key would have mattered little upon any other occasion, but on this one day it has produced the most deplorable consequences.

How many document should be retrieved? 3

Jaccard Coefficient result: top 3 documents are-

1 ) Dataset/Stories and Fictions\toilet.s
2 ) Dataset/Stories and Fictions\wanderer.fun
3 ) Dataset/Stories and Fictions\gloves.txt

Tf-Idf based document retrieval result: top 3 documents are-

1 ) Dataset/Stories and Fictions\3student.txt
2 ) Dataset/Stories and Fictions\gulliver.txt
3 ) Dataset/Stories and Fictions\darkness.txt

Cosine similarity result: top 3 documents are-

1 ) Dataset/Stories and Fictions\shrdfarm.txt
2 ) Dataset/Stories and Fictions\toilet.s
3 ) Dataset/Stories and Fictions\fish.txt

**This query is taken from 3student.txt. Here** Tf-Idf based document retrieval **gives the accurate answer in the top-1 document.**

## Output 3 (with attention to title g-weight =0.07 ) using Term frequency

Enter the query- The Early Days of a High-Tech Start-up are Magic (November 18, 1991) by M. Peshota

How many document should be retrieved? 4

Jaccard Coefficient result: top 4 documents are-

1 ) Dataset/Stories and Fictions\write
2 ) Dataset/Stories and Fictions\quarter.c6
3 ) Dataset/Stories and Fictions\quarter.c4
4 ) Dataset/Stories and Fictions\quarter.c16

Tf-Idf based document retrieval result: top 4 documents are-

1 ) Dataset/Stories and Fictions\rid.txt
2 ) Dataset/Stories and Fictions\17.lws
3 ) Dataset/Stories and Fictions\domain.poe
4 ) Dataset/Stories and Fictions\quarter.c4

Cosine similarity result: top 4 documents are-

1 ) Dataset/Stories and Fictions\17.lws
2 ) Dataset/Stories and Fictions\rid.txt
3 ) Dataset/Stories and Fictions\eyeargon.hum
4 ) Dataset/Stories and Fictions\gay

## Output 4 (without attention to title ) using Term frequency

Enter the query- The Early Days of a High-Tech Start-up are Magic (November 18, 1991) by M. Peshota

How many document should be retrieved? 4

Jaccard Coefficient result: top 4 documents are-

1 ) Dataset/Stories and Fictions\write
2 ) Dataset/Stories and Fictions\quarter.c6
3 ) Dataset/Stories and Fictions\quarter.c4
4 ) Dataset/Stories and Fictions\quarter.c16

Tf-Idf based document retrieval result:  top  4  documents are-

1 )  Dataset/Stories and Fictions\rid.txt
2 )  Dataset/Stories and Fictions\thewave
3 )  Dataset/Stories and Fictions\domain.poe
4 )  Dataset/Stories and Fictions\fearmnky

Cosine similarity result:  top  4  documents are-

1 )  Dataset/Stories and Fictions\thewave
2 )  Dataset/Stories and Fictions\100west.txt
3 )  Dataset/Stories and Fictions\17.lws
4 )  Dataset/Stories and Fictions\mindwar

**Actual documents are highlighted in green color.**

# Q2)  Edit Distance-based word correction

**Total number of words-**  65197

## Output 1

Enter the query-  The Early Days of a High-Tech Start-up are Magic (November 18, 1991) by M. Peshota

How many words should be suggested?  4

100%|██████████| 65197/65197 [00:05<00:00, 12721.10it/s]

tech  Suggestions-

[(2, 'ch'), (2, 'eh'), (2, 'te'), (2, 'teach')]

100%|██████████| 65197/65197 [00:02<00:00, 22945.08it/s]

18  Suggestions-

[(4, 'a'), (4, 'b'), (4, 'c'), (4, 'd')]

100%|██████████| 65197/65197 [00:05<00:00, 12302.98it/s]

1991 Suggestions-

[(6, 'a'), (6, 'b'), (6, 'c'), (6, 'd')]

100%|██████████| 65197/65197 [00:08<00:00, 7425.72it/s]

peshota Suggestions-

[(3, 'peso'), (3, 'pest'), (3, 'shot'), (4, 'eta')]

## Output 2

Enter the query- Dreamworld, by Zaphod Beeblebrox

How many words should be suggested? 3

100%|██████████| 65197/65197 [00:12<00:00, 5369.15it/s]

dreamworld Suggestions-

[(4, 'reword'), (5, 'drawl'), (5, 'dread')]

100%|██████████| 65197/65197 [00:07<00:00, 8524.36it/s]

zaphod Suggestions-

[(3, 'hod'), (3, 'phd'), (3, 'pod')]

100%|██████████| 65197/65197 [00:12<00:00, 5356.82it/s]

beeblebrox Suggestions-

[(6, 'beer'), (7, 'bee'), (7, 'beeper')]