Assignment Report on

TF-IDF Score and Ranking of 100 Wikipedia articles

Adikar Bharath N V S (171IT202)

Under the Guidance of,

Dr. Sowmya Kamath S.
Department of Information Technology, NITK Surathkal

Date of Submission: 4th October 2020

in partial fulfillment for the award of the degree
of
Bachelor of Technology
In
Information Technology
At

Department of Information Technology
National Institute of Technology Karnataka, Surathkal
October 2020

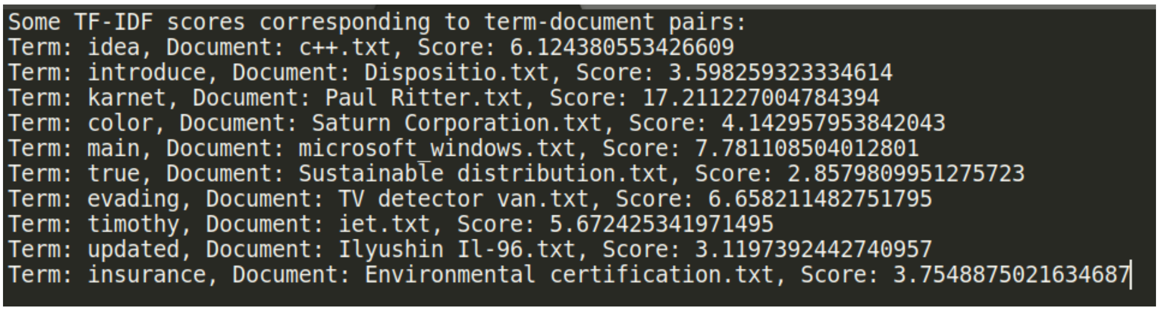
Table of Contents

Chapter 1

Results

1.1 Term-Weight Modelling using TF-IDF scoring

The TF-IDF score of a term-document pair is the product of Term Frequncy as well as Inverse Document Frequency. Illustrated below are TF-IDF score examples:



```
Some TF-IDF scores corresponding to term-document pairs:
Term: idea, Document: c++.txt, Score: 6.124380553426609
Term: introduce, Document: Dispositio.txt, Score: 3.598259323334614
Term: karnet, Document: Paul Ritter.txt, Score: 17.211227004784394
Term: color, Document: Saturn Corporation.txt, Score: 4.142957953842043
Term: main, Document: microsoft_windows.txt, Score: 7.781108504012801
Term: true, Document: Sustainable distribution.txt, Score: 2.8579809951275723
Term: evading, Document: TV detector van.txt, Score: 6.658211482751795
Term: timothy, Document: iet.txt, Score: 5.672425341971495
Term: updated, Document: Ilyushin Il-96.txt, Score: 3.1197392442740957
Term: insurance, Document: Environmental certification.txt, Score: 3.7548875021634687
```

Figure 1.1: Examples of TF-IDF scores

1.2 Variants in TF-IDF

The TF-IDF score of a term-document pair is affected by how TF and IDF are calculated. The following formulae are used:

1.2.1 TF

- Log Normalization: 1 + log2(f(i,j))
- Double Normalization K: K + (1 - K) * (f(i,j) / max_i(f(i,j)))

f(i,j) denotes frequency of term i in document j. max_i denotes maximum over all i terms.

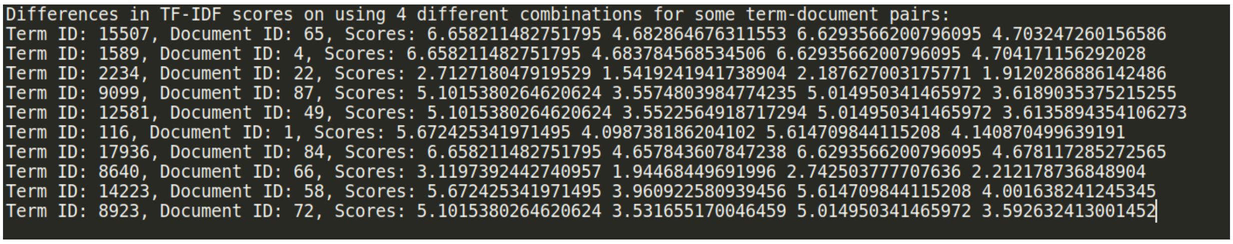
1.2.2 IDF

- Inverse Frequency Smooth: log2(1 + (N/n(i)))
- Probabilistic Inverse Frequency: log2((N - n(i) / n(i)))

N denotes number of documents. n(i) denotes number of documents the term i is present in.

Illustrated below are examples of TF-IDF scores using the above four combinations in the order:

1. Log Normalization + Inverse Frequency Smooth
2. Double Normalization 0.7 + Probabilistic Inverse Frequency
3. Log Normalization + Probabilistic Inverse Frequency
4. Double Normalization 0.7 + Inverse Frequency Smooth



```
Differences in TF-IDF scores on using 4 different combinations for some term-document pairs:
Term ID: 15507, Document ID: 65, Scores: 6.658211482751795 4.682864676311553 6.6293566200796095 4.703247260156586
Term ID: 1589, Document ID: 4, Scores: 6.658211482751795 4.683784568534506 6.6293566200796095 4.704171156292028
Term ID: 2234, Document ID: 22, Scores: 2.712718047919529 1.5419241941738904 2.187627003175771 1.9120286886142486
Term ID: 9099, Document ID: 87, Scores: 5.1015380264620624 3.5574803984774235 5.014950341465972 3.6189035375215255
Term ID: 12581, Document ID: 49, Scores: 5.1015380264620624 3.5522564918717294 5.014950341465972 3.6135894354106273
Term ID: 116, Document ID: 1, Scores: 5.672425341971495 4.098738186204102 5.614709844115208 4.140870499639191
Term ID: 17936, Document ID: 84, Scores: 6.658211482751795 4.657843607847238 6.6293566200796095 4.678117285272565
Term ID: 8640, Document ID: 66, Scores: 3.1197392442740957 1.94468449691996 2.742503777707636 2.212178736848904
Term ID: 14223, Document ID: 58, Scores: 5.672425341971495 3.960922580939456 5.614709844115208 4.001638241245345
Term ID: 8923, Document ID: 72, Scores: 5.1015380264620624 3.531655170046459 5.014950341465972 3.592632413001452
```

Figure 1.2: Differences between 4 different TF-IDF scoring types

1.3 Generating the Ranking using TF-IDF scores

To rank these documents according to their TF-IDF scores, they are represented as a Vector Space Model collection. Then a query is generated with a fixed number of terms, with the scores and terms itself randomly chosen. Each document is then compared with the query using Cosine Similarity as the closeness measure.

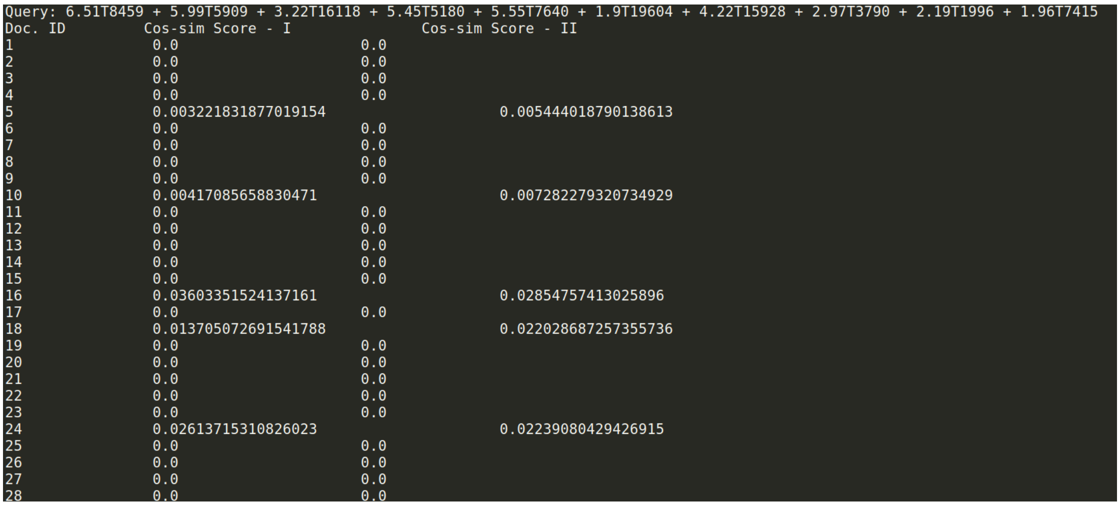
The formula is:

$$\cos(\alpha) = \frac{\sum_{j=1}^{t} w_{qj} w_{ij}}{\sqrt{\sum_{j=1}^{t} (w_{ij})^2 \sum_{j=1}^{t} (w_{qj})^2}}$$

Figure 1.3: Formula for Cosine Similarity

wqj denotes weight of term j in query q, and wij for document i.

Illustrated below are the Cos-sim scores of the documents against a query using the first two TF-IDF scoring types from section 1.2.

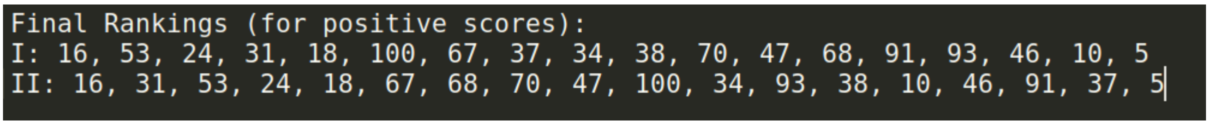


```
Query: 6.51T8459 + 5.99T5909 + 3.22T16118 + 5.45T5180 + 5.55T7640 + 1.9T19604 + 4.22T15928 + 2.97T3790 + 2.19T1996 + 1.96T7415
Doc. ID         Cos-sim Score - I              Cos-sim Score - II
1                0.0                     0.0
2                0.0                     0.0
3                0.0                     0.0
4                0.0                     0.0
5                0.003221831877019154                    0.005444018790138613
6                0.0                     0.0
7                0.0                     0.0
8                0.0                     0.0
9                0.0                     0.0
10               0.00417085658830471                     0.007282279320734929
11               0.0                     0.0
12               0.0                     0.0
13               0.0                     0.0
14               0.0                     0.0
15               0.0                     0.0
16               0.03603351524137161                     0.02854757413025896
17               0.0                     0.0
18               0.013705072691541788                    0.022028687257355736
19               0.0                     0.0
20               0.0                     0.0
21               0.0                     0.0
22               0.0                     0.0
23               0.0                     0.0
24               0.02613715310826023                     0.02239080429426915
25               0.0                     0.0
26               0.0                     0.0
27               0.0                     0.0
28               0.0                     0.0
```

Figure 1.4: Cos-sim scores of the documents against an example query

1.4 Differences in Ranking of TF-IDF Variants

To compare the rankings of TF-IDF scoring types, the documents are arranged according to the decreasing order of cos-sim scores, giving a final ranked list. Illustrated below are the final ranked lists:



```
Final Rankings (for positive scores):
I: 16, 53, 24, 31, 18, 100, 67, 37, 34, 38, 70, 47, 68, 91, 93, 46, 10, 5
II: 16, 31, 53, 24, 18, 67, 68, 70, 47, 100, 34, 93, 38, 10, 46, 91, 37, 5
```

Figure 1.5: Ranking the documents

The ranked lists are observed to be different for different TF-IDF scoring types.