

Assignment Report on

COMPARISON OF BEST-MATCH MODELS

Adikar Bharath N V S (171IT202)

Under the Guidance of,

Dr. Sowmya Kamath S.

Department of Information Technology, NITK Surathkal

Date of Submission: 31st 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

1	Results	1
1.1	Original BM Models	1
1.1.1	Preliminary Calculations	1
1.1.2	Main Formulae	2
1.1.3	Output	3
1.1.4	Observations	3
1.2	Simplified BM Models	3
1.2.1	Main Formulae	4
1.2.2	Output	5
1.2.3	Observations	5
1.3	BM25 Modelling	6
1.3.1	Preliminary Calculations	6
1.3.2	Main Formulae	6
1.3.3	Output	7
1.3.4	Observations	8

Chapter 1

Results

The query was chosen at random, both in term and frequency selection.

```
Query Q:  
(levett, 1), (stoic, 1), (internet, 1), (girlfriend, 1), (kalayaan, 3),
```

Figure 1.1: Query Q

1.1 Original BM Models

For calculating values using BM1, BM11, and BM15, the following formulae were used:

1.1.1 Preliminary Calculations

$$\mathcal{F}_{i,j} = S_1 \times \frac{f_{i,j}}{K_1 + f_{i,j}}$$

Figure 1.2: Term-frequency factor $\mathcal{F}(i,j)$

$$\mathcal{F}'_{i,j} = S_1 \times \frac{f_{i,j}}{\frac{K_1 \times \text{len}(d_j)}{\text{avg_doclen}} + f_{i,j}}$$

Figure 1.3: Term-frequency factor with Document length normalization $\mathcal{F}'(i,j)$

$$\mathcal{G}_{j,q} = K_2 \times \text{len}(q) \times \frac{\text{avg_doclen} - \text{len}(d_j)}{\text{avg_doclen} + \text{len}(d_j)}$$

Figure 1.4: Correction factor $G(j,q)$

$$\mathcal{F}_{i,q} = S_3 \times \frac{f_{i,q}}{K_3 + f_{i,q}}$$

Figure 1.5: Term-frequency within query factor $F(i,q)$

1.1.2 Main Formulae

$$\begin{aligned} BM1(d_j, q) &\sim \sum_{k_i[q, d_j]} \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right) \\ BM15(d_j, q) &\sim \mathcal{G}_{j,q} + \sum_{k_i[q, d_j]} \mathcal{F}_{i,j} \times \mathcal{F}_{i,q} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right) \\ BM11(d_j, q) &\sim \mathcal{G}_{j,q} + \sum_{k_i[q, d_j]} \mathcal{F}'_{i,j} \times \mathcal{F}_{i,q} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right) \end{aligned}$$

Figure 1.6: Original BM Formulae

1.1.3 Output

Original BM modelling:
Constant values:
K1: 3, K2: 1.4, K3: 100
Scores:

Document Chosen	BM1	BM11	BM15
Dilpazier Aslam.txt	6.46828462519127	8.011096520613693	5.761462503205051
coup_d'état.txt	0.0	-0.9463109934297164	-0.9463109934297164
Henry Failing.txt	0.0	1.6393576073083223	1.6393576073083223
KLF2.txt	0.0	1.9264301215953812	1.9264301215953812
Kyle Turley.txt	0.0	-1.2268322330644783	-1.2268322330644783
Creatine kinase.txt	0.0	2.8806018183281035	2.8806018183281035
Environmental certification.txt	0.0	-0.26842752323751223	-0.26842752323751223
Sustainable distribution.txt	0.0	0.34090501524299105	0.34090501524299105
Scottish National Dictionary A.txt	0.0	5.350645805496889	5.350645805496889
whatsapp.txt	6.46828462519127	2.6915723697010683	4.668639739649949
Thomas Müller.txt	6.625708843064465	2.4549536949563557	3.900287890741662
mohammad_asghar.txt	0.0	2.8017399632851783	2.8017399632851783
Moufang polygon.txt	0.0	3.4069101606479184	3.4069101606479184
Paul Ritter.txt	0.0	-0.6931542225776397	-0.6931542225776397
Andrew Sentence.txt	0.0	2.283090180663693	2.283090180663693
Protagoras.txt	0.0	1.101957531832276	1.101957531832276
maruti_suzuki.txt	0.0	-1.207381252400558	-1.207381252400558
Technological Educational Inst.txt	0.0	-0.1504266830274975	-0.1504266830274975
Grand duke.txt	0.0	0.9520642350351052	0.9520642350351052
The Age of Extremes.txt	0.0	-0.8685905437116183	-0.8685905437116183
Old Pasadena.txt	0.0	-0.22599150532296336	-0.22599150532296336
Biogenic silica.txt	0.0	0.05185270255859151	0.05185270255859151
Bay State Games.txt	0.0	0.4555041932993936	0.4555041932993936
chillum_maryland.txt	0.0	-0.6838864683364806	-0.6838864683364806
Government of the Republic of .txt	0.0	0.28491648517410556	0.28491648517410556
Donny van de Beek.txt	0.0	1.6306838147445355	1.6306838147445355
John Rudge (banker).txt	0.0	3.245664636544581	3.245664636544581
Juande Ramos.txt	0.0	1.9079303234096348	1.9079303234096348
LW9.txt	0.0	1.889507047351168	1.889507047351168
George Young (diplomat).txt	0.0	2.384430640917874	2.384430640917874
Petr Kvičala.txt	0.0	-0.2946810917061225	-0.2946810917061225
SEPnet.txt	0.0	2.7131396675818897	2.7131396675818897
New York Journal-American.txt	0.0	0.017313861701345933	0.017313861701345933
android_os.txt	6.46828462519127	-0.8352686479592233	2.533198160651369
south_china_sea.txt	6.643856189774724	11.73156465931884	16.446750383583435
Meadow's Law.txt	0.0	2.333485337068282	2.333485337068282
Kyllian Mbappé.txt	0.0	-0.8334038722265663	-0.8334038722265663
Center for Media, Religion and.txt	6.46828462519127	3.8487578308787693	3.4512481846519014
michael_jordan.txt	13.093993468255736	-0.20773144890487005	4.6440673494850095
Agroecology in West Africa.txt	0.0	0.6005234167086636	0.6005234167086636
A. E. Levett.txt	6.643856189774724	25.189070804375515	23.901451511397845

Figure 1.7: Original BM Modelling

1.1.4 Observations

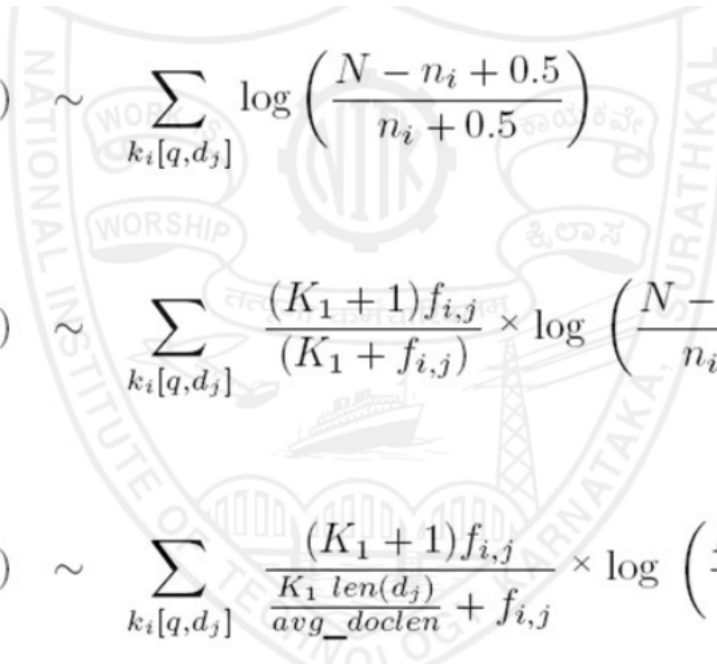
BM1 scores are quite proportional to the number of terms common to both the query and document, and thus is not really useful for ranking. BM11 and BM15 scores, considering various statistical parameters like mean and variance, are a better measure of ranking. An issue is that documents both relevant and irrelevant to the query have similar scores apart from some outliers.

On average, BM15 scores are higher than BM11, with no difference in irrelevant documents.

1.2 Simplified BM Models

Simplification eliminates the need for many factors by appropriately setting factors

1.2.1 Main Formulae



$$\begin{aligned}
 BM1(d_j, q) &\sim \sum_{k_i[q, d_j]} \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right) \\
 BM15(d_j, q) &\sim \sum_{k_i[q, d_j]} \frac{(K_1 + 1)f_{i,j}}{(K_1 + f_{i,j})} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right) \\
 BM11(d_j, q) &\sim \sum_{k_i[q, d_j]} \frac{(K_1 + 1)f_{i,j}}{\frac{K_1 \text{ len}(d_j)}{\text{avg_doclen}} + f_{i,j}} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right)
 \end{aligned}$$

Figure 1.8: Simplified BM Formulae

1.2.2 Output

```
Simplified BM modelling:
Constant values:
K1: 1.2, K2: 0, K3: inf
Scores:
```

Document Chosen	BM1	BM11	BM15
Dilpazier Aslam.txt	6.46828462519127	4.167812466114742	2.823749360308273
coup_d'état.txt	0.0	0.0	0.0
Henry Failing.txt	0.0	0.0	0.0
KLF2.txt	0.0	0.0	0.0
Kyle Turley.txt	0.0	0.0	0.0
Creatine kinase.txt	0.0	0.0	0.0
Environmental certification.txt	0.0	0.0	0.0
Sustainable distribution.txt	0.0	0.0	0.0
Scottish National Dictionary A.txt	0.0	0.0	0.0
whatsapp.txt	6.46828462519127	4.172188744421446	5.009877897321129
Thomas Müller.txt	6.625708843064465	4.164513727257927	5.300123724569014
mohammad_asghar.txt	0.0	0.0	0.0
Moufang polygon.txt	0.0	0.0	0.0
Paul Ritter.txt	0.0	0.0	0.0
Andrew Sentance.txt	0.0	0.0	0.0
Protagoras.txt	0.0	0.0	0.0
maruti_suzuki.txt	0.0	0.0	0.0
Technological Educational Inst.txt	0.0	0.0	0.0
Grand duke.txt	0.0	0.0	0.0
The Age of Extremes.txt	0.0	0.0	0.0
Old Pasadena.txt	0.0	0.0	0.0
Biogenic silica.txt	0.0	0.0	0.0
Bay State Games.txt	0.0	0.0	0.0
chillum_maryland.txt	0.0	0.0	0.0
Government of the Republic of .txt	0.0	0.0	0.0
Donny van de Beek.txt	0.0	0.0	0.0
John Rudge (banker).txt	0.0	0.0	0.0
Juande Ramos.txt	0.0	0.0	0.0
LW9.txt	0.0	0.0	0.0
George Young (diplomat).txt	0.0	0.0	0.0
Petr Kvičala.txt	0.0	0.0	0.0
SEPnet.txt	0.0	0.0	0.0
New York Journal-American.txt	0.0	0.0	0.0
android_os.txt	6.46828462519127	3.009639193771802	4.778652763598616
south_china_sea.txt	6.643856189774724	4.795190488879321	6.051662119822493
Meadow's law.txt	0.0	0.0	0.0
Kyllian Mbappé.txt	0.0	0.0	0.0
Center for Media, Religion and.txt	6.46828462519127	3.1021583755781785	2.823749360308273
michael_jordan.txt	13.093993468255736	3.908728692411693	8.123873084877287
Agroecology in West Africa.txt	0.0	0.0	0.0
A. E. Levett.txt	6.643856189774724	12.994110281112222	12.679673012961414

Figure 1.9: Simplified BM Modelling

1.2.3 Observations

Simplified BM models, particularly BM11 and BM15, are much better compared to the original counterparts, as irrelevant documents are given a value of 0. Also, the scores of relevant documents for BM11 and BM15 are somewhat more varied than BM1, offering for better ranking. The general trend of BM15 scores being higher than BM11 is present here. An issue is that the variance between scores has been reduced, which may prove to be an issue while ranking.

1.3 BM25 Modelling

1.3.1 Preliminary Calculations

$$\mathcal{B}_{i,j} = \frac{(K_1 + 1)f_{i,j}}{K_1 \left[(1 - b) + b \frac{\text{len}(d_j)}{\text{avg_doclen}} \right] + f_{i,j}}$$

Figure 1.10: Additional Factor $\mathcal{B}(i,j)$

1.3.2 Main Formulae

$$BM25(d_j, q) \sim \sum_{k_i[q,d_j]} \mathcal{B}_{i,j} \times \log \left(\frac{N - n_i + 0.5}{n_i + 0.5} \right)$$

Figure 1.11: BM25 Formula

1.3.3 Output

```
BM25 Modelling:
With b: 0.2 and K1: 1.3, scores:
Document Chosen      BM25
Dilpazier Aslam.txt  2.3992797187913877
coup_d'état.txt      0.0
Henry Failing.txt     0.0
KLF2.txt              0.0
Kyle Turley.txt       0.0
Creatine kinase.txt   0.0
Environmental certification.txt 0.0
Sustainable distribution.txt 0.0
Scottish National Dictionary A.txt 0.0
whatsapp.txt          4.841743710073616
Thomas Müller.txt     4.503402298444516
mohammad_asghar.txt  0.0
Moufang polygon.txt   0.0
Paul Ritter.txt       0.0
Andrew Sentance.txt   0.0
Protagoras.txt        0.0
maruti_suzuki.txt     0.0
Technological Educational Inst.txt 0.0
Grand duke.txt        0.0
The Age of Extremes.txt 0.0
Old Pasadena.txt      0.0
Biogenic silica.txt   0.0
Bay State Games.txt   0.0
chillum_maryland.txt 0.0
Government of the Republic of .txt 0.0
Donny van de Beek.txt 0.0
John Rudge (banker).txt 0.0
Juande Ramos.txt      0.0
LW9.txt               0.0
George Young (diplomat).txt 0.0
Petr Kvičala.txt      0.0
SEPnet.txt            0.0
New York Journal-American.txt 0.0
android_os.txt        4.5521156776545935
south_china_sea.txt   5.141968473959415
Meadow's law.txt      0.0
Kylia Mbappé.txt      0.0
Center for Media, Religion and.txt 2.3992797187913877
michael_jordan.txt    6.902682017235904
Agroecology in West Africa.txt 0.0
A. E. Levett.txt      12.994630904878367
```

Figure 1.12: BM25 with first combination of constants

```

With b: 0.8 and K1: 1.8, scores:
Document Chosen      BM25
Dilpazier Aslam.txt  1.5645184280049194
coup_d'état.txt      0.0
Henry Failing.txt    0.0
KLF2.txt             0.0
Kyle Turley.txt      0.0
Creatine kinase.txt  0.0
Environmental certification.txt 0.0
Sustainable distribution.txt 0.0
Scottish National Dictionary A.txt 0.0
whatsapp.txt         4.36647934644513
Thomas Müller.txt    2.9365712673907525
mohammad_asghar.txt  0.0
Moufang polygon.txt  0.0
Paul Ritter.txt      0.0
Andrew Sentance.txt  0.0
Protagoras.txt       0.0
maruti_suzuki.txt    0.0
Technological Educational Inst.txt 0.0
Grand duke.txt       0.0
The Age of Extremes.txt 0.0
Old Pasadena.txt     0.0
Biogenic silica.txt  0.0
Bay State Games.txt  0.0
chillum_maryland.txt 0.0
Government of the Republic of .txt 0.0
Donny van de Beek.txt 0.0
John Rudge (banker).txt 0.0
Juande Ramos.txt     0.0
LW9.txt             0.0
George Young (diplomat).txt 0.0
Petr Kvičala.txt     0.0
SEPnet.txt          0.0
New York Journal-American.txt 0.0
android_os.txt       3.9269236930609135
south_china_sea.txt  3.35296646352776
Meadow's law.txt     0.0
Kylia n Mbappé.txt   0.0
Center for Media, Religion and.txt 1.5645184280049194
michael_jordan.txt   4.501089695395672
Agroecology in West Africa.txt 0.0
A. E. Levett.txt     14.496223329780136

```

Figure 1.13: BM25 with second combination of constants

1.3.4 Observations

The first combination of b and $K1$ yielded an average higher BM25 value than the second. From the formulae, it is more apparent that the value of b plays a much larger role than

K1. The first combination has a value closer to 0, making it closer to a BM15 model, while the second combination is closer to 1, almost resembling a BM11 model. From earlier observations, BM15 has on average higher scores than BM11, hence the difference is justified.