
Name	Arnav Kumar Sinha
Roll No.	22075013
Course	B.Tech CSE

Table Of Contents

- Table Of Contents
- Examples
 - Indian Economy Technology Startup
 - IIT BHU Varanasi
 - Bollywood Movies Uttar Pradesh

Examples

NOTE: A common theme among the top matches are that they have smaller length and match some relevant terms. This may be due to using $(1 + \log(1 + tf)) \cdot \log \frac{N}{idf}$ as the weighting function. Since the documents themselves are not very large, the sublinear tf term is not helping that much.

Indian Economy Technology Startup

```
> Query as a sequence of double quoted terms
> Input text to exit
> e.g.: "short" "watch"
> e.g.: "hello" "world" "web"
> "indian" "economy" "technology" "startup"

(*) parsed query: indian economy technology startup
(*) Unnormalized "query vector": defaultdict(<class 'int'>, {'indian': 2.562952850845111, 'economy': 6.225806123571734, 'technology': 4.946082316413015, 'startup': 16.158599976013328})
[8] Got > 50 matches
-- print some results? y/n (w to save to file results.txt): w
-- written matches to results.txt :)
-- print some results? y/n (w to save to file results.txt): y

4.356 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/business/1050504_business_story_469192.utf8
4.163 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/business/1051215_business_story_5600155.utf8
3.24 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2004_utf8/business/1041103_business_story_3559002.utf8
3.201 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/business/1070820_business_story_832857.utf8
3.081 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/business/1070828_business_index.utf8
2.375 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/calcutta/1050424_calcutta_story_455131.utf8
2.264 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/foreign/1070606_foreign_story_788145.utf8
2.199 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2004_utf8/opinion/1041123_opinion_story_403915.utf8
1.994 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2004_utf8/calcutta/1041003_calcutta_story_385954.utf8
1.426 /home/aks/codin/cso_assignments/CSE561/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2004_utf8/business/1040923_business_story_579779.utf8

-- print some results? y/n (w to save to file results.txt):
```

Figure 1: indian-economy-technology-startup

matching weight	doc name
4.356	TELEGRAPH_UTF8/2005_utf8/business/1050504_business_story_469192.utf8
4.163	TELEGRAPH_UTF8/2005_utf8/business/1051215_business_story_5600155.utf8
3.24	TELEGRAPH_UTF8/2004_utf8/business/1041103_business_story_3559002.utf8
3.201	TELEGRAPH_UTF8/2007_utf8/business/1070820_business_story_832857.utf8
3.081	TELEGRAPH_UTF8/2007_utf8/business/1070828_business_index.utf8
2.375	TELEGRAPH_UTF8/2005_utf8/calcutta/1050424_calcutta_story_455131.utf8
2.264	TELEGRAPH_UTF8/2007_utf8/foreign/1070606_foreign_story_788145.utf8
2.199	TELEGRAPH_UTF8/2004_utf8/opinion/1041123_opinion_story_403915.utf8
1.994	TELEGRAPH_UTF8/2004_utf8/calcutta/1041003_calcutta_story_385954.utf8
1.426	TELEGRAPH_UTF8/2004_utf8/business/1040923_business_story_579779.utf8

As, it can be seen in the image, startup is a relatively rare term and hence has a much higher term weight. Hence the results are skewed for that.

For example, in the topmost document:

“...he said, adding that their **startup** would depend on compliance with govern...”

Other matches are also the terms. They are lower down the list primarily because they are much larger documents.

IIT BHU Varanasi

```

> Query as a sequence of double quoted terms      3.24      TELEGRAPH_UTF8/2004_utf8/business/1041103.
> Input EXIT to exit                             3.201     TELEGRAPH_UTF8/2007_utf8/business/1070820.
> e.g.: "short" "match"                          3.081     TELEGRAPH_UTF8/2007_utf8/business/1070828.
> e.g.: "hello" "world" "web"                    2.375     TELEGRAPH_UTF8/2005_utf8/calcutta/1050424.
> "lit" "bhu" "varanasi"                         2.375     TELEGRAPH_UTF8/2004_utf8/foreign/1070606.
[*] parsed query: iit bhu varanasi
[*] Unnormalized "query vector": defaultdict(cclass 'int', {'iit': 8.740255452708505, 'bhu': 15.0817738877537, 'varanasi': 7.155725982337911})
[*] Got >= 50 matches
-- print some results? y/n/cw to save to file results.txt): y
5.82 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/nation/1061118_nation_story_7018129.utf8
5.12 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/nation/1061125_nation_story_7048473.utf8
4.463 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/nation/1050914_nation_story_5235787.utf8
4.054 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/nation/1070701_nation_story_7999825.utf8
3.312 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/nation/1060403_nation_story_6049919.utf8
3.307 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/nation/1051027_nation_story_5404782.utf8
3.222 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2004_utf8/jobs/1041221_jobs_story_4124852.utf8
3.22 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/nation/1070802_nation_story_8137493.utf8
3.193 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/nation/1060328_nation_story_6023320.utf8
3.074 /home/aks/codin/cso_assignments/CSI365/assignment/3/.../IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/nation/1050914_nation_story_5235726.utf8
which highest term weight. Hence the results are skewed for that.
-- print some results? y/n/cw to save to file results.txt): n
-- written matches to results.txt :)
-- print some results? y/n/cw to save to file results.txt): n

```

Figure 2: iit-bhu-varanasi

5.82	TELEGRAPH_UTF8/2006_utf8/nation/1061118_nation_story_7018129.utf8
5.12	TELEGRAPH_UTF8/2006_utf8/nation/1061125_nation_story_7048473.utf8
4.463	TELEGRAPH_UTF8/2005_utf8/nation/1050914_nation_story_5235787.utf8
4.054	TELEGRAPH_UTF8/2007_utf8/nation/1070701_nation_story_7999825.utf8
3.312	TELEGRAPH_UTF8/2006_utf8/nation/1060403_nation_story_6049919.utf8
3.307	TELEGRAPH_UTF8/2005_utf8/nation/1051027_nation_story_5404782.utf8
3.222	TELEGRAPH_UTF8/2004_utf8/jobs/1041221_jobs_story_4124852.utf8
3.22	TELEGRAPH_UTF8/2007_utf8/nation/1070802_nation_story_8137493.utf8
3.193	TELEGRAPH_UTF8/2006_utf8/nation/1060328_nation_story_6023320.utf8
3.074	TELEGRAPH_UTF8/2005_utf8/nation/1050914_nation_story_5235726.utf8

The matches are relevant to BHU, IITs etc. More finegrained results can be obtained by further adding relevant terms like *technology*, or maybe about specific departments *chemical*

Bollywood Movies Uttar Pradesh

```

> Query as a sequence of double quoted terms
> input EXIT to exit
> e.g.: "short" "watch"
> e.g.: "hello" "world" "web"
> "hollywood" "movies" "uttar" "pradesh"

[5] parsed query: hollywood movies uttar pradesh
[6] Unnormalized "query vector": defaultdict(<class 'int'>, {'hollywood': 6.094592028990871, 'movi': 6.396691276158592, 'uttar': 5.7968862792841372, 'pradesh': 4.967745417522595})
[8] Got 350 matches
[9] print some results? y/n(↵ to save to file results.txt): y

2.782 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/sports/1060326_sports_story_6016586_utf8
2.574 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/sports/1051222_sports_story_5635956_utf8
2.215 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/sports/1051222_sports_story_56351571_utf8
2.020 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/frontpage/1070118_frontpage_index_utf8
1.887 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/nation/1070118_nation_story_7047458_utf8
1.884 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/sports/1060082_sports_story_6652116_utf8
1.860 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/nation/1070619_nation_story_7666971_utf8
1.864 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/frontpage/1070222_frontpage_index_utf8
1.826 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2004_utf8/nation/1060911_nation_story_5745927_utf8
1.814 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/frontpage/1061212_frontpage_index_utf8

[9] print some results? y/n(↵ to save to file results.txt): w
[9] written matches to results.txt
[9] print some results? y/n(↵ to save to file results.txt): y

1.812 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/nation/1050610_nation_story_4874520_utf8
1.812 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2004_utf8/nation/1060911_nation_story_5751522_utf8
1.77 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2007_utf8/nation/1070516_nation_story_7512473_utf8
1.757 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/nation/1061112_nation_story_6990892_utf8
1.701 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/nation/1060812_nation_story_5661880_utf8
1.659 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/sports/1061012_sports_story_6866465_utf8
1.687 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/sports/1060751_sports_story_6348491_utf8
1.684 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/sports/1060900_sports_story_5720169_utf8
1.678 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2005_utf8/sports/1050119_sports_story_4270881_utf8
1.678 /home/aks/codin/cso_assignments/CSI365/assignment/3/~/IR_Assignment_Dataset/english/TELEGRAPH_UTF8/2006_utf8/sports/1060825_sports_story_6655242_utf8

[9] print some results? y/n(↵ to save to file results.txt):

```

Figure 3: bollywood movies uttar pradesh

5.82	TELEGRAPH_UTF8/2006_utf8/nation/1061118_nation_story_7018129.utf8
5.12	TELEGRAPH_UTF8/2006_utf8/nation/1061125_nation_story_7048473.utf8
4.463	TELEGRAPH_UTF8/2005_utf8/nation/1050914_nation_story_5235787.utf8
4.054	TELEGRAPH_UTF8/2007_utf8/nation/1070701_nation_story_7999825.utf8
3.312	TELEGRAPH_UTF8/2006_utf8/nation/1060403_nation_story_6049919.utf8
3.307	TELEGRAPH_UTF8/2005_utf8/nation/1051027_nation_story_5404782.utf8
3.222	TELEGRAPH_UTF8/2004_utf8/jobs/1041221_jobs_story_4124852.utf8
3.22	TELEGRAPH_UTF8/2007_utf8/nation/1070802_nation_story_8137493.utf8
3.193	TELEGRAPH_UTF8/2006_utf8/nation/1060328_nation_story_6023320.utf8
3.074	TELEGRAPH_UTF8/2005_utf8/nation/1050914_nation_story_5235726.utf8

```
> Query as a sequence of double quoted terms
> input EXIT to exit
> e.g.: "short" "watch"
> e.g.: "hello" "world" "web"
> EXIT
bye :)
```

Figure 4: bye :)