Information Retrieval
Assignment - 2
Group-7

Submission id:
Roshan S
roshan20039@iiitd.ac.in

Submitted To:
Dr. Rajiv Ratn shah

1. **Sample positional index for a word**

```
term : shareware
No of Docs:  5
Doc ID: 1 Term Freq: 1 Positions: [0] File Name: 100west.txt
Doc ID: 43 Term Freq: 1 Positions: [0] File Name: arctic.txt
Doc ID: 77 Term Freq: 1 Positions: [6] File Name: breaks1.asc
Doc ID: 79 Term Freq: 1 Positions: [9383] File Name: breaks3.asc
Doc ID: 127 Term Freq: 2 Positions: [8, 12] File Name: cybersla.txt
```

Phrase query : “good day”

```
Enter number of queries 1
Enter input query good day
['good', 'day']
No of Docs:  21
13chil.txt
aesop11.txt
aesopa10.txt
brain.damage
breaks2.asc
bruce-p.txt
enchdup.hum
fantasy.hum
fantasy.txt
fic5
forgotte
history5.txt
horswolf.txt
hound-b.txt
mazarin.txt
melissa.txt
outcast.dos
sick-kid.txt
srex.txt
```

startrek.txt
superg1

2. **Jaccard Coefficient:**
   Query: 100 west 53 by north
   top 5 documents are:

   peace.fun
   snowmaid.txt
   prince.art
   campfire.txt
   glimpse1.txt

   TF-IDF Matrix:

   Query: I will endeavour, in my statement, to avoid such terms as would serve to limit the events to any particular place, or give a clue as to the people concerned

Weighting Scheme	TF Weight	TF-IDF Score	Resulting Documents
Binary	0,1	29.3879313210617 25.0205954017978 24.4466725841418 24.0327595524623 22.4336231322488	3student.txt darkness.txt history5.txt hound-b.txt radar_ra.txt
Raw count	f(t,d)	780.890531905074 549.442025462667 358.612397045353 317.378841195747 289.114632756608	gulliver.txt vgilante.txt hound-b.txt hitch3.txt hitch2.txt
Term Frequency	f(t,d)/∑f(t`,d)	0.09465521245131 0.05419773264508 0.04135114332628 0.03955580705502 0.03661454747972	jim.asc dwar quarter.c11 sre02.txt wanderer.fun
Log Normalization	log(1+f(t,d))	52.9313373813274 46.3161730104630 41.1804739267762 39.1829189318143 37.2654888098417	gulliver.txt hound-b.txt vgilante.txt radar_ra.txt hitch2.txt

Double Normalization	0.5+0.5*(f(t,d)/ max(f(t`,d))	15.3828869071525 13.3797511171881 13.2584404198979 12.5286832005816 11.5317405051663	3student.txt history5.txt darkness.txt hound-b.txt radar_ra.txt

Cosine Similarity:

Query: I will endeavour, in my statement, to avoid such terms as would serve to limit the events to any particular place, or give a clue as to the people concerned

Weighting Scheme	TF Weight	Cosine Similarity Score	Resulting Documents
Binary	0,1	0.07880057984292 0.04432548906444 0.03580292278805 0.03562733971790 0.03407998569123	3student.txt goldenp.txt monkking.txt szechuan lament.txt
Raw count	f(t,d)	0.08680444411447 0.04287494734591 0.04192362088749 0.03453633542115 0.03432647043376	jim.asc 3student.txt gulliver.txt dwar sretrade.txt
Term Frequency	f(t,d)/∑f(t`,d)	0.08680444411447 0.04287494734591 0.04192362088749 0.03453633542115 0.03432647043376	jim.asc 3student.txt gulliver.txt dwar sretrade.txt
Log Normalization	log(1+f(t,d))	0.07765792771423 0.05570947072558 0.04089764888712 0.03923409699016 0.03818585874918	3student.txt jim.asc goldenp.txt sretrade.txt wisteria.txt
Double Normalization	0.5+0.5*(f(t,d)/ max(f(t`,d))	0.07882780358023 0.04449950240156 0.04335840492979 0.03595747520057 0.03554480138058	3student.txt goldenp.txt jim.asc monkking.txt szechuan

Question3:

a.

Total Qid:4 Data 103

Data Snippet

```
['0 qid:4 1:3 2:0 3:2 4:0 5:3 6:1 7:0 8:0.666667 9:0 10:1 11:999 12:0 13:110 14:5 15:1114
16:14.976692 17:28.949002 18:25.594644 19:28.531344 20:14.972391 21:20 22:0 23:5 24:0
25:25 26:1 27:0 28:0 29:0 30:1 31:12 32:0 33:4 34:0 35:16 36:6.666667 37:0 38:1.666667 39:0
40:8.333333 41:20.222222 42:0 43:2.888889 44:0 45:37.555556 46:0.02002 47:0 48:0.045455
49:0 50:0.022442 51:0.001001 52:0 53:0 54:0 55:0.000898 56:0.012012 57:0 58:0.036364 59:0
60:0.014363 61:0.006673 62:0 63:0.015152 64:0 65:0.007481 66:0.00002 67:0 68:0.000239
69:0 70:0.00003 71:77.577533 72:0 73:30.667985 74:0 75:90.53171 76:5.52713 77:0 78:0 79:0
80:5.526745 81:57.882066 82:0 83:18.750101 84:0 85:66.125373 86:25.859178 87:0
88:10.222662 89:0 90:30.177237 91:525.177766 92:0 93:60.031269 94:0 95:675.850674 96:1
97:0 98:0 99:0 100:1 101:0.875901 102:0 103:0.66135 104:0 105:0.864571 106:28.756809
107:0 108:3.274639 109:0 110:28.985515 111:-17.640291 112:-29.251906 113:-20.596041
114:-31.107208 115:-17.519629 116:-19.440921 117:-31.580405 118:-24.146168
119:-33.960286 120:-19.161514 121:-16.596977 122:-31.750477 123:-21.267965
124:-33.908554 125:-16.503638 126:2 127:27 128:2 129:9 130:124 131:4678 132:54 133:74
134:0 135:0 136:0 \n', …………. '0 qid:4 1:3 2:0 3:2 4:0 5:3 6:1 7:0 8:0.666667 9:0 10:1 11:399
12:5 13:13 14:9 15:426 16:14.976692 17:28.949002 18:25.594644 19:28.531344 20:14.972391
21:23 22:0 23:3 24:0 25:26 26:1 27:0 28:0 29:0 30:1 31:17 32:0 33:2 34:0 35:18 36:7.666667
37:0 38:1 39:0 40:8.666667 41:46.222222 42:0 43:0.666667 44:0 45:49.555556 46:0.057644
47:0 48:0.230769 49:0 50:0.061033 51:0.002506 52:0 53:0 54:0 55:0.002347 56:0.042607 57:0
58:0.153846 59:0 60:0.042254 61:0.019215 62:0 63:0.076923 64:0 65:0.020344 66:0.00029
67:0 68:0.003945 69:0 70:0.000273 71:66.943274 72:0 73:28.523293 74:0 75:84.625987
76:5.52713 77:0 78:0 79:0 80:5.526745 81:41.344333 82:0 83:23.835768 84:0 85:57.859701
86:22.314425 87:0 88:9.507764 89:0 90:28.208662 91:216.32666 92:0 93:106.307993 94:0
95:480.740714 96:1 97:0 98:0 99:0 100:1 101:0.838129 102:0 103:0.805181 104:0
105:0.853774 106:34.08340 107:0 108:17.222283 109:0 110:35.37843 111:-14.910365
112:-29.251906 113:-17.399841 114:-31.107208 115:-14.665252 116:-18.794143
117:-31.580405 118:-23.863117 119:-33.960286 120:-18.43832 121:-13.837747 122:-31.750477
123:-17.692544 124:-33.908554 125:-13.640485 126:4 127:59 128:1415 129:14 130:5334
131:6434 132:4 133:17 134:0 135:0 136:0 \n']
```

b.

Output:

Data Stored

Q3_a.txt: Some snippet of Data

```
3 qid:4 1:3 2:0 3:2 4:1 5:3 6:1 7:0 8:0.666667 9:0.333333 10:1 11:344 12:0 13:19 14:6 15:369
16:14.976692 17:28.949002 18:25.594644 19:28.531344 20:14.972391 21:99 22:0 23:6 24:1
```

```
25:106 26:6 27:0 28:0 29:0 30:6 31:51 32:0 33:4 34:1 35:56 36:33 37:0 38:2 39:0.333333
40:35.333333 41:378 42:0 43:2.666667 44:0.222222 45:454.222222 46:0.287791 47:0
48:0.315789 49:0.166667 50:0.287263 51:0.017442 52:0 53:0 54:0 55:0.01626 56:0.148256
57:0 58:0.210526 59:0.166667 60:0.151762 61:0.09593 62:0 63:0.105263 64:0.055556
65:0.095754 66:0.003194 67:0 68:0.007387 69:0.006173 70:0.003336 71:381.086021 72:0
73:45.331988 74:9.586712 75:411.010633 76:49.589179 77:0 78:0 79:0 80:49.59403
81:281.883642 82:0 83:35.956938 84:9.586712 85:309.497726 86:127.028674 87:0
88:15.110663 89:3.195571 90:137.003544 91:11990.030799 92:0 93:231.932187 94:20.423345
95:14878.022216 96:1 97:0 98:0 99:0 100:1 101:0.699343 102:0 103:0.631671 104:0.555497
105:0.688618 106:46.563656 107:0 108:17.431926 109:11.57800 110:46.454896 111:-8.213598
112:-29.251906 113:-18.918429 114:-26.85940 115:-8.28204 116:-13.780081 117:-31.580405
118:-25.885492 119:-30.743095 120:-13.673395 121:-8.376022 122:-31.750477 123:-20.000721
124:-29.652723 125:-8.446421 126:2 127:32 128:349 129:8 130:123 131:281 132:22 133:6
134:0 135:0 136:0
.
.
.
.
0 qid:4 1:1 2:0 3:0 4:0 5:1 6:0.333333 7:0 8:0 9:0 10:0.333333 11:286 12:0 13:8 14:4 15:298
16:14.976692 17:28.949002 18:25.594644 19:28.531344 20:14.972391 21:3 22:0 23:0 24:0 25:3
26:0 27:0 28:0 29:0 30:0 31:3 32:0 33:0 34:0 35:3 36:1 37:0 38:0 39:0 40:1 41:2 42:0 43:0 44:0
45:2 46:0.01049 47:0 48:0 49:0 50:0.010067 51:0 52:0 53:0 54:0 55:0 56:0.01049 57:0 58:0 59:0
60:0.010067 61:0.003497 62:0 63:0 64:0 65:0.003356 66:0.000024 67:0 68:0 69:0 70:0.000023
71:3.542084 72:0 73:0 74:0 75:3.539923 76:0 77:0 78:0 79:0 80:0 81:3.542084 82:0 83:0 84:0
85:3.539923 86:1.180695 87:0 88:0 89:0 90:1.179974 91:2.78808 92:0 93:0 94:0 95:2.78468
96:0 97:0 98:0 99:0 100:0 101:0.14408 102:0 103:0 104:0 105:0.144067 106:2.961128 107:0
108:0 109:0 110:2.950603 111:-32.853281 112:-29.251906 113:-30.299778 114:-31.107208
115:-32.90238 116:-32.401703 117:-31.580405 118:-30.664309 119:-33.960286 120:-32.399129
121:-35.180812 122:-31.750477 123:-32.595124 124:-33.908554 125:-35.204738 126:1 127:20
128:0 129:2 130:18637 131:11377 132:12 133:110 134:0 135:122 136:29.542582010582
```

Obtained max DCG 20.989750804831445
Number of files that can be made
198934973759383705998260476149053298969368401705665705882051803127048579926951
93482412686565431050240000000000000000000000
Count of relevance score {'0': 59, '1': 26, '2': 17, '3': 1}

Analysis: The max DCG will always be obtained when the relevance score of the queries are sorted in descending order. The max DCG will be the ideal DCG in which the highest relevance scores url query will be at the top.

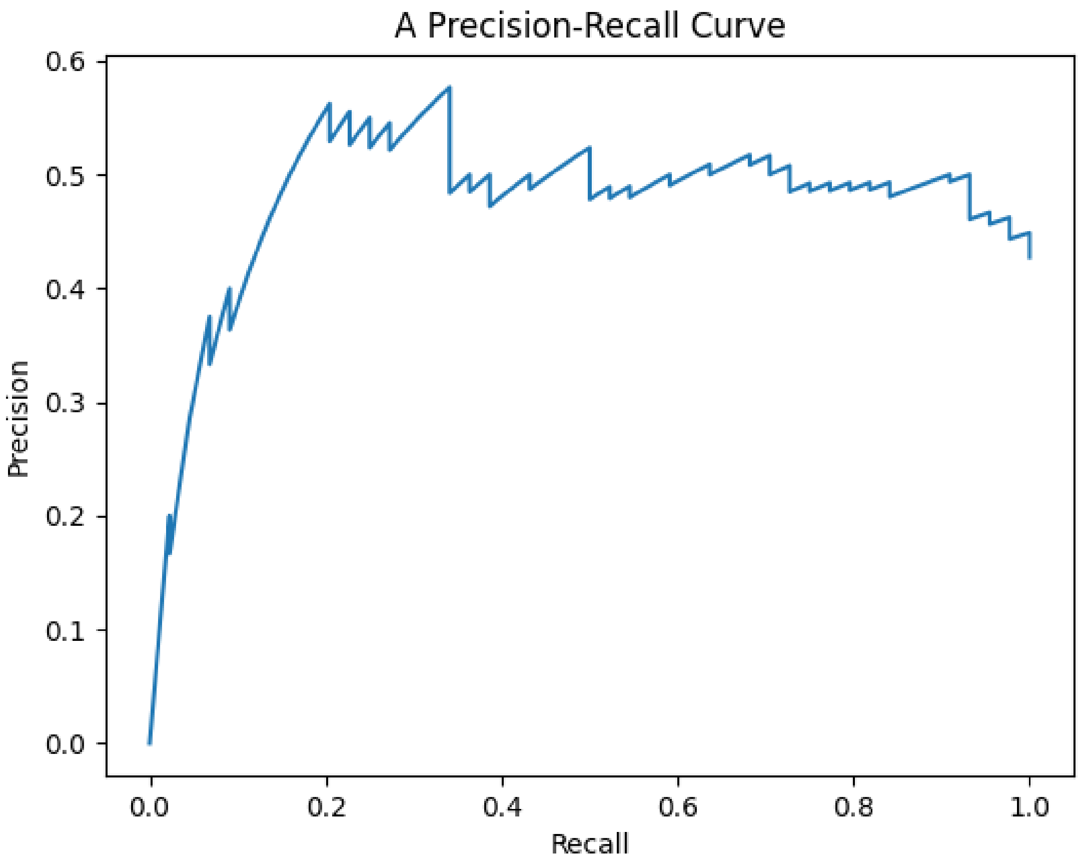
c.

Position at 50
DCG: 7.390580969258021
IDCG: 14.067092644997018
NDCG: 0.5253808413557646

Whole Dataset
DCG: 12.550247459532576
IDCG: 20.989750804831445
NDCG: 0.5979226516897831

d.

Precision Recall Curve


A Precision-Recall Curve
Precision
Recall