Experiments

For each query we have considered a total of 10 web pages in which 5 are taken from google and 5 are taken from Bing. The file named “queries.txt” has 50 URLs where 1-10 are relevant documents for Query1, 11-20 are relevant to Query2 and so on.

The precision at kth position is calculated as follows:

Precision@k = (Number of Relevant Doc. till k)/(Total number of Doc. till k)

Average Precision is calculated as follows:

$$\frac{1}{|Rel|} \cdot \sum_{i=1}^{k} (\text{relevant}(i) \times \text{P@i})$$

Here relevant(i) is 1 if the ith result is relevant and 0 otherwise.
Rel = the set of actual relevant documents.
k = total number of documents.

Query#1: What time is it?

O/P:

```
(8,0.021460443242957) -> Precision at 1 = 1
(5,0.015777896727363) -> Precision at 2 = 1
(3,0.015462381223619) -> Precision at 3 = 1
(2,0.0092496986551338) -> Precision at 11 = 0.36363636363636
(10,0.0082403849298738) -> Precision at 15 = 0.33333333333333
(1,0.0081158452475989) -> Precision at 16 = 0.375
(6,0.0081158452475989) -> Precision at 17 = 0.41176470588235
(9,0.0071813502701465) -> Precision at 19 = 0.42105263157895
(4,0.0063812471971615) -> Precision at 24 = 0.375
(7,0.0063812471971615) -> Precision at 25 = 0.4
The Average Precision = 0.5679787034431
```

Here the first part represents the document number and the cosine similarity, and the second part represents the precision of that particular document. According to my_seeds.txt we have provided URLS first 10 URLS 5 from google and 5 from bing which are related to the query “What time is it?”,

so the most relevant documents should be among the first 10 documents which is why you can see only the documents from 1-10 represented in the O/P

Total Relevant Documents = 10

Average Precision1 = (Sum of all Precision@k values)/(Total number of Relevant Documents)

= 0.5679787034431

Query#2: How to register to vote?

O/P:

```
(17,0.038734489457471) -> Precision at 1 = 0
(11,0.027833605334399) -> Precision at 2 = 1
(19,0.027833605334399) -> Precision at 3 = 1
(14,0.024867173667433) -> Precision at 4 = 1
(15,0.020831701325575) -> Precision at 5 = 1
(12,0.020219524466585) -> Precision at 6 = 1
(16,0.020219524466585) -> Precision at 7 = 1
(13,0.017043612039518) -> Precision at 8 = 1
(20,0.017043612039518) -> Precision at 9 = 1
(18,0.014261820145446) -> Precision at 10 = 1
The Average Precision = 1
```

Here the first part represents the document number and the cosine similarity, and the second part represents the precision of that particular document. According to my_seeds.txt we have provided URLS between 11 and 20 URLS, 5 from google and 5 from bing which are related to the query “How to register to vote?”, so the most relevant documents should be between 11 and 20 documents which is why you can see only the documents from 11-20 represented in the O/P

Total Relevant Documents = 10

Average Precision2 = (Sum of all Precision@k values)/(Total number of Relevant Documents)

= 1

Query#3: How to tie a tie?

O/P:

```
(21,0.226364068059560) -> Precision at 1 = 1
(23,0.029723884772171) -> Precision at 2 = 1
(28,0.029723884772171) -> Precision at 3 = 1
(24,0.009918914816437 2) -> Precision at 4 = 1
(27,0.009918914816437 2) -> Precision at 5 = 1
(29,0.009391733447523 5) -> Precision at 6 = 1
(25,0.009179378095971 2) -> Precision at 7 = 1
(26,0.005793449830827 7) -> Precision at 8 = 1
(30,0.004300108690501 3) -> Precision at 9 = 1
(22,0.004124624557016 1) -> Precision at 10 = 1
The Average Precision = 1
```

Here the first part represents the document number and the cosine similarity, and the second part represents the precision of that particular document.

Total Relevant Documents = 10

Average Precision3 = (Sum of all Precision@k values)/(Total number of Relevant Documents)

= 1

Query#4: Can you run it?

O/P:

```
(32,0.025061090348037) -> Precision at 1 = 1
(37,0.025061090348037) -> Precision at 2 = 1
(40,0.023772381774872) -> Precision at 3 = 1
(31,0.021444463151186) -> Precision at 4 = 1
(36,0.021444463151186) -> Precision at 5 = 1
(33,0.018998141861556) -> Precision at 9 = 0.66666666666667
(38,0.018998141861556) -> Precision at 10 = 0.7
(35,0.010509596193832) -> Precision at 14 = 0.57142857142857
(39,0.009448217257852) -> Precision at 15 = 0.6
(34,0.0091874241187901) -> Precision at 16 = 0.625
The Average Precision = 0.81630952380952
```

Here the first part represents the document number and the cosine similarity, and the second part represents the precision of that particular document.

Total Relevant Documents = 10

Average Precision4 = (Sum of all Precision@k values)/(Total number of Relevant Documents)

= 0.81630952380952

Query#5: What song is this?

O/P:

```
(42,0.036977836616282) -> Precision at 1 = 1
(48,0.036977836616282) -> Precision at 2 = 1
(50,0.030224488407132) -> Precision at 3 = 1
(47,0.027894969615602) -> Precision at 4 = 1
(45,0.023852681081951) -> Precision at 5 = 1
(43,0.019023611557121) -> Precision at 6 = 1
(49,0.017433638851078) -> Precision at 7 = 1
(44,0.013115014294578) -> Precision at 9 = 0.88888888888889
(41,0.0060901929601637) -> Precision at 11 = 0.81818181818182
(46,0.0049649056537619) -> Precision at 12 = 0.83333333333333
The Average Precision = 0.9540404040404
```

Here the first part represents the document number and the cosine similarity, and the second part represents the precision of that particular document.

Total Relevant Documents = 10

Average Precision5 = (Sum of all Precision@k values)/(Total number of Relevant Documents)

= 0.954040404040404

Mean Average Precision (MAP) Score:

(AP1 + AP2 + AP3 + AP4 + AP5)/5

(0.5679787034431 + 1 + 1 + 0.81630952380952 + 0.954040404040404)/5

= 0.8676657262586048

Conclusion:

lookup_index.php outputs all the documents and their respective cosine similarity with the query in a decreasing order. The important point to note is that, cosine similarity is greater than 0 if there is atleast 1 word that is common in both document and query.

Since tf-idf doesn't take document normalization into account (unlike Okapi BM25), a large document, although not relevant to the query, might appear in the top search results for a query.