

CPS 842 Assignment 2 report

Xiangyu Lu 500679658

Marc McCombe 500786634

1. For each term, in the posting list, is ordered in document ID;
2. For our program we integrated a top-K retrieval method by only returning the top-50 results based off of descending cosine similarity score (first result = highest score) as any results returned after this point (50) generally have too low a similarity score to be worthwhile.
3. For the tf-idf weighting scheme
 - a. $Tf_i = 1 + \log(fi)$
 - b. $Idf_i = \log(1+N/df_i)$
 - c. $Weight = tf * idf$
 - d. Cosine similarity score = $\frac{\bar{d} \cdot \bar{q}}{|d| * |q|}$

How to run the program

1. Open the terminal window under linux system, then type the following;
2. `% javac Invert.java`
3. `% java Invert`
4. Following the instruction
5. Then posting.txt and dictionary.txt will be generated
6. `%javac Search.java`
7. `%java Search`
8. User enters desired query term, then hit return; ranked documents with their similarity score will be printed out;
9. `%javac Eval.java`
10. `%java Eval`
11. The program will give AP values of each query found in query.txt as well as the final MAP value
 - a. One side note is that Eval.java always assumes stemming and stop word removal is on

Not working

1. R-precision is not implemented in the program.

Sample runs

1. Generating posting.txt and dictionary.txt among document collection.

```
~/Desktop/842 Assignment2(marc*) » javac Invert.java xiangyulu@Seans-Mac
-----
~/Desktop/842 Assignment2(marc*) » java Invert xiangyulu@Seans-Mac
Enter y to use stopwords removal and stemming algorithm
Enter n to continue without stopwords removal and stemming algorithm
y
-----
~/Desktop/842 Assignment2(marc*) »
```

2. User enter query term to perform search.

```
~/Desktop/842 Assignment2(marc*) » javac Search.java xiangyulu@Seans-Mac
-----
~/Desktop/842 Assignment2(marc*) » java Search xiangyulu@Seans-Mac
java.lang.ArrayIndexOutOfBoundsException: 5296
Stop word removal/stemming? (y/n):
y
y
Search:
computer monitor
IDocID= 2796 |Title= Monitors: An Operating System Structuring Concept (Corrige
dum) |Similarity= 0.5427093381920287
IDocID= 754 |Title= Ye Indiscreet Monitor |Similarity= 0.36201548100857456
IDocID= 1340 |Title= Multiplexing of Slow Peripherals |Similarity= 0.2784891404
448597
IDocID= 2866 |Title= Proving Monitors |Similarity= 0.2329532630983162
IDocID= 1363 |Title= A General Method of Systematic Interval Computation |Simil
rity= 0.2328173716050817
IDocID= 2571 |Title= An Analytic Model of the Hasp Execution Task Monitor |Simi
arity= 0.22530337105864381
IDocID= 1829 |Title= An Interactive Graphical Display Monitor in |Similarity= 0
2170877568862665
IDocID= 1642 |Title= Time Sharing on a Computer with a Small Memory |Similarity
0.20883031870914118
```

3. Evaluation for query.txt and qrels.txt

```
~/Desktop/842 Assignment2(marc*) » javac Eval.java                                xiangyulu@Seans-Mac
-----
~/Desktop/842 Assignment2(marc*) » java Eval                                     xiangyulu@Seans-Mac
End of file reached
AP value for query: What articles exist which deal with TSS (Time Sharing System), an operating system for IBM computers? is: 0.2
AP value for query: I am interested in articles written either by Prieve or Udo Pooch. APrieve, B.Pooch, U. is: 0.0
AP value for query: Intermediate languages used in construction of multi-targeted compilers; TCOLL is: 0.041666666666666664
AP value for query: I'm interested in mechanisms for communicating between disjoint processes, possibly, but not exclusively, in a distributed environment. I would rather see descriptions of complete mechanisms, with or without implementations, as opposed to theoretical work on the abstract problem. Remote procedural calls and message-passing are examples of my interests. is: 0.007575757575757576
AP value for query: I'd like papers on design and implementation of editing interfaces, window-managers, command interpreters, etc. The essential issues are human interface design, with views on improvements to user efficiency, effectiveness and satisfaction. is: 0.19691928475935827
AP value for query: Interested in articles on robotics, motion planning particularly the geometric and combinatorial aspects. We are not interested in the dynamics of arm motion. is: 0.016666666666666666
AP value for query: I am interested in distributed algorithms - concurrent programs in which processes communicate and synchronize by using message passing. Area
```

```
Ons between parallel and sequential algorithms. is: 0.17020800552551002
AP value for query: List all articles on EL1 and ECL (EL1 may be given as EL/1;
I don't remember how they did it. is: 0.0
Final MAP value is: 0.1484602173009952
-----
~/Desktop/842 Assignment2(marc*) »
```