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. Tfi = 1 + log(fi)
    b. Idfi = log(1+N/dfi)
    c. Weight = tf * idf
    d. Cosine similarity score = $\frac{\overline{d} \cdot \overline{q}}{|d| * |q|}$

How to run the program

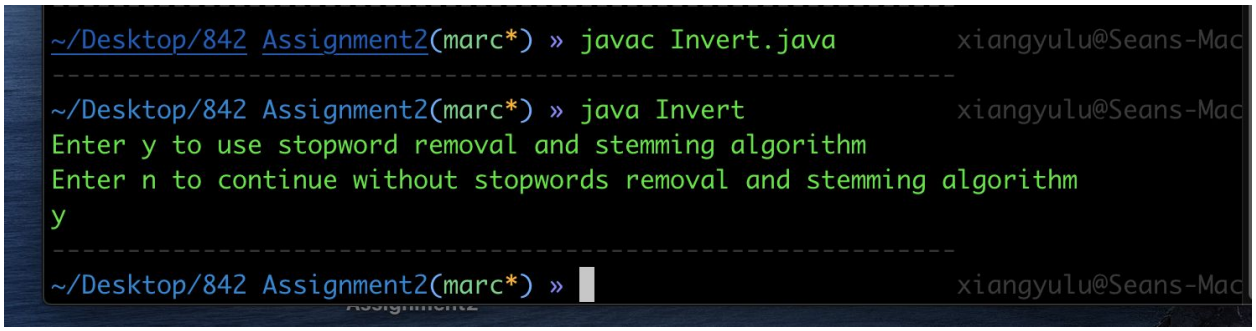
1. Open the terminal window under linux system, then type the following;
2. % javac Invert.java
3. % java Invert
4. Following the instruction
5. Then posting.txt and dictionary.txt will be generated
6. %javac Search.java
7. %java Search
8. User enters desired query term, then hit return; ranked documents 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

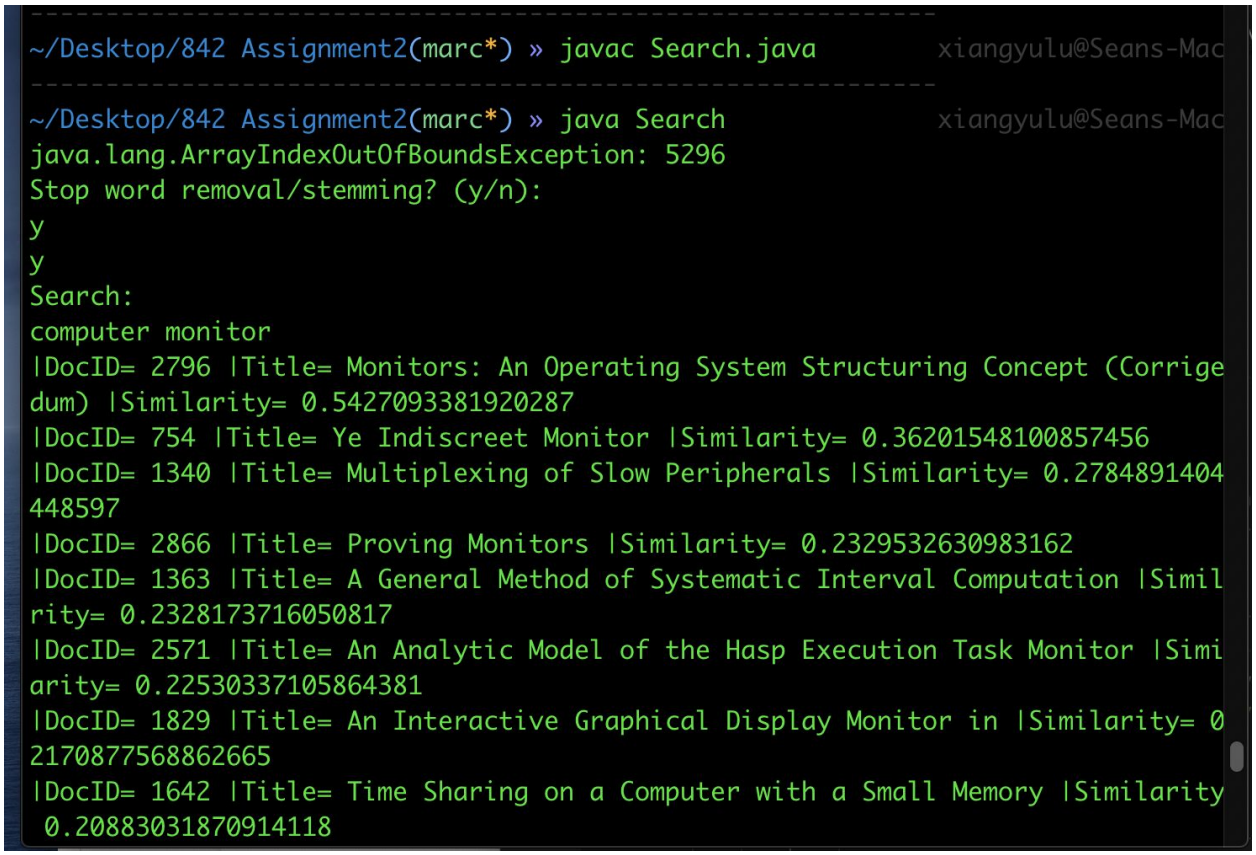
1. R-precision is not implemented in the program.

Sample runs

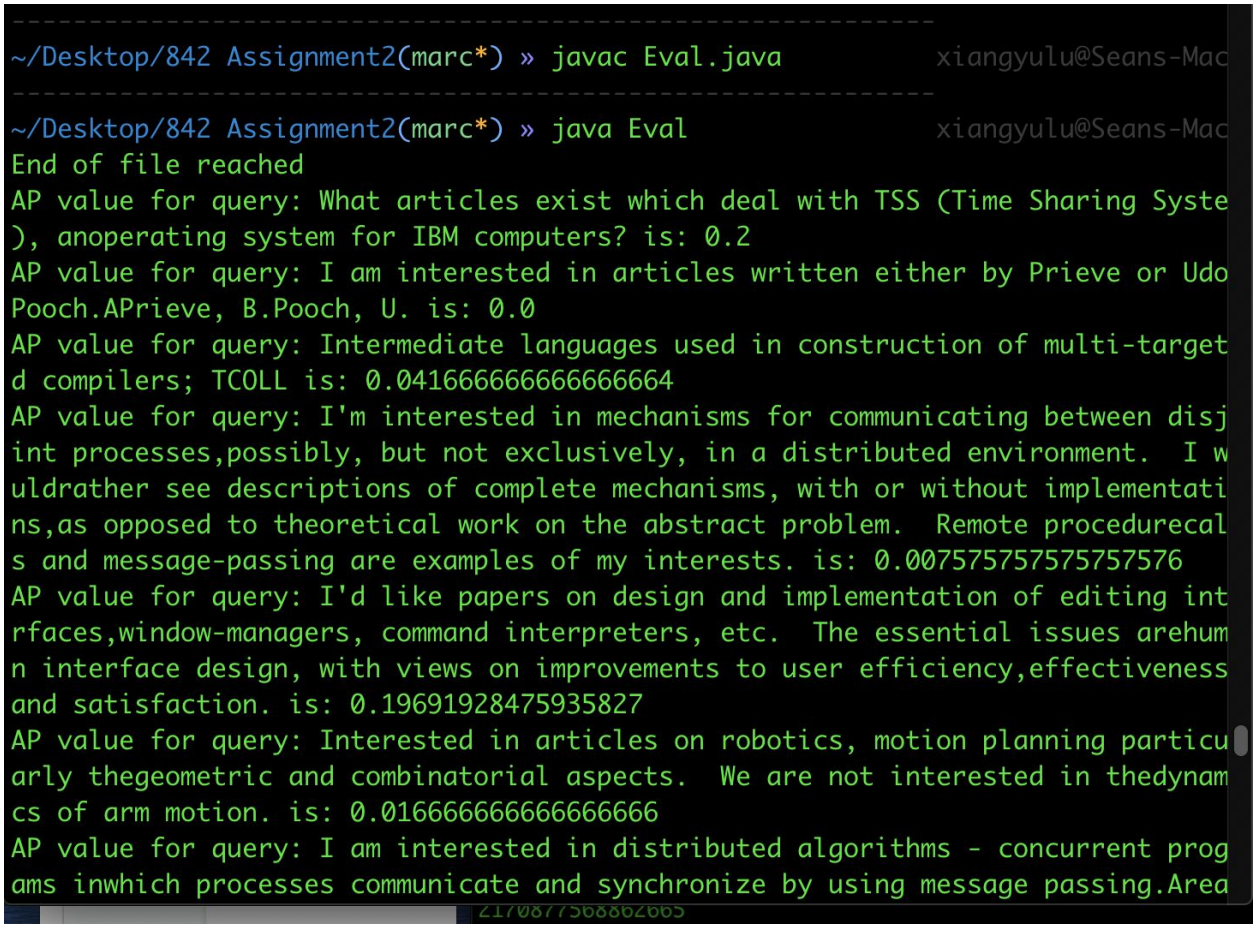
1. Generating posting.txt and dictionary.txt among document collection.



2. User enter query term to perform search.



3. Evaluation for query.txt and qrels.txt



….