

CPS842 Assignment 2

Daniyil Kotov, 500877422

Vladislav Brovikov, 500864655

Implementation information

Postings lists are ordered in document id.

We did not implement a predefined K top-K retrieval method, however we did implement a Champion List-based collection of relevant documents. In our implementation of the champion list for each term t , is the set of r documents with weights for t that are higher than the median weight for t . The set of relevant to the query documents is the union of all the champion lists for each term.

For the tf-idf weighting scheme, we used:

documents:

$$TF = 1 + \log F,$$

$$IDF = \log(N / DF),$$

$$W = tf * idf$$

queries:

$$TF = F,$$

$$IDF = \log(N / DF),$$

$$W = tf * idf$$

where F is the number of occurrences of the term in the document. N is the total number of documents and DF is in how many documents the term occurs.

Screenshots

```
Dan@DESKTOP-N9B0CL3 ~/Documents/Coding/CPS842/a2 (master)
$ python invert.py
Do you want to enable the stopwords removal? (y/n) y
Stopwords removal enabled
Do you want to enable stemming? (y/n) y
Stemming enabled
```

Invert program

```
Dan@DESKTOP-N9B0CL3 ~/Documents/Coding/CPS842/a2 (master)
$ python user_interface.py
Please input a query to search for: cold
1.
Document Title: Cold-Start vs. Warm-Start Miss Ratios
Document Author(s): Easton, M.C. Fagin, R.
Please input a query to search for: ZZEND
```

```
PS C:\Users\test\Desktop\cps 842\assignment 2\cps842> python .\user_interface.py
Please input a query to search for: Parallel languages; languages for parallel computation
1.
Document Title: Glypnir-A Programming Language for Illiac IV
Document Author(s): Lawrie, D. H. Layman, T. Baer, D. Randal, J. M.
2.
Document Title: Procedure-Oriented Language Statements to Facilitate Parallel Processing
Document Author(s): Opler, A.
3.
Document Title: Three Criteria for Designing Computing Systems to Facilitate Debugging
Document Author(s): Van Horn, E. C.
4.
Document Title: A Language for Formal Problem Specification
Document Author(s): Greif, I.
5.
Document Title: Programming Semantics for Multiprogrammed computations
Document Author(s): Dennis, J. B. VanHorn, E. C.
6.
Document Title: Control Structures in Illiac IV Fortran
Document Author(s): Millstein, R. E.
7.
Document Title: Computing Connected Components on Parallel Computers
Document Author(s): Hirschberg, D.S. Chandra, A.K. Sarwate, D.V.
8.
Document Title: Parallel Methods for Integrating Ordinary Differential Equations
Document Author(s): Nievergelt, J.
9.
Document Title: Cellular Arrays for the Solution of Graph Problems
Document Author(s): Levitt, K. N. Kautz, W. H.
10.
Document Title: On Simulating Networks of Parallel Processes in which Simultaneous Events May occur
Document Author(s): Parnas, D. L.
11.
Document Title: The Parallel Execution of DO Loops
Document Author(s): Lamport, L.
12.
Document Title: A Highly Parallel Algorithm for Approximating All Zeros of a Polynomial with only Real Zeros
Document Author(s): Patrick, M. L.
13.
Document Title: On the Optimal Detection of Curves in Noisy Pictures
Document Author(s): Montanari, U.
14.
Document Title: Fast Parallel Sorting Algorithms
Document Author(s): Hirschberg, D.S.
15.
Document Title: Multiprogramming STRETCH: Feasibility Considerations
Document Author(s): Codd, E. F. Lowry, E. S. McDonough, E. Scalzi, C. A.
16.
Document Title: Scheduling Meetings with a Computer
Document Author(s): Schultz, C. K. Brooks, A. Schwartz, P.
```

Search program

```
Dan@DESKTOP-N9B0CL3 ~/Documents/Coding/CPS842/a2 (master)
$ python eval.py
MAP: 0.18184736248251063
R-Precision: 0.20170888433431408
```

Eval program