باسمه تعالی

تکلیف دوم درس مبانی بازیابی اطلاعات

A small search engine

The main steps are:

* Read in the already stemmed document collection provided in the file tccorpus.txt inside tccorpus.zip. This is an early standard collection of abstracts from the *Communications of the ACM*.
* Build a simple inverted indexer that reads the corpus and writes the index. You should invoke it with
* indexer tccorpus.txt index.out

* Implement a new version of TFIDF ranking algorithm, and write a program to provide a ranked list of documents for a file with one or more queries. You should pass parameters for the index file, the query file, and the maximum number of document results, and return documents on the standard output, like so:
* NewTFIDF index.out queries.txt 100 > results.eval
* Submit the output from this run, with the top 100 document IDs and their NewTFIDF scores for each test query according to the following format:
* query\_id Q0 doc\_id rank NewTFIDF\_score system\_name

The string Q0 is a literal used by the standard TREC evaluation script. You can use any space-free token for your system\_name.

* Also, submit your code, instructions for compiling it, and a short report describing your implementation and your NewTFIDF ranking algorithm.

Tokenized Document Collection

* The provided tccorpus.txt file is in the format:
  + A # followed by a document ID
  + Lines below the document ID line contain stemmed words from the document.
* For example:
* # 1
* this is a tokenzied line for document 1
* this is also a line of document 1
* # 2
* from here lines for document 2 begin
* ...
* ...
* # 3
* ...
* For tokenization, simply break the character sequence at any run of whitespace (space, newline, etc.).

Test Queries: Use the following stemmed test queries, also provided in the file queries.txt:

|  |  |
| --- | --- |
| Query ID | Query Text |
| 1 | portabl oper system |
| 2 | code optim for space effici |
| 3 | parallel algorithm |
| 4 | distribut comput structur and algorithm |
| 5 | appli stochast process |
| 6 | perform evalu and model of comput system |
| 7 | parallel processor in inform retriev |