IR Assignment #4 [Vector Space Model]

Implement the Vector Space Model for Scoring and displaying top k documents for a query. Refer section 6.3.3 and implement algorithm 6.14. Use tf-idf weightings and logarithmic functions for weights as explained below.

The sample text file (“cranfield.all.1400”) is a single file that contains 1400 documents. You have to preprocess and filter the file to read the body of each document (preceded by .W). You can use code from previous assignments to preprocess the documents including normalization, stopword removal and stemming.

To implement the VSM, you may choose to implement (you can can do it differently) your dictionary and postings lists in the format shown in the table. The tuple in each posting represents (doc ID, term freq).

|  |  |  |  |
| --- | --- | --- | --- |
| Term | Doc freq(df) | -> | Postings lists |
| Ambitious | 5 | -> | (1,5)->(7,2)->(21,7)….. |
|  |  |  |  |

In the searching step, you will need to rank documents by cosine similarity based on tf×idf. In terms of SMART notation of ddd.qqq, you will need to implement the ltc.ltn ranking scheme. That means, log tf and idf for both documents and queries. Only the document component needs to be cosine normalized. Compute the similarity between the query and each document, with the weights following the tf×idf calculation, where

term freq = 1 + log(tf) and inverse document frequency idf = log(N/df). That is,

tf-idf = (1 + log(tf)) \* log(N/df).

It's suggested that you use log base 10 for your logarithm calculations (i.e., math.log(x, 10), but be careful of cases like math.log(0, 10)). Your searcher should output a list of up to 10 most relevant (less if there are fewer than ten documents that have matching stems to the query) docIDs in response to the query.

You can use Java or Python to implement this assignment.

Dataset :

1. cranfield collection

The queries and results are shown on the next page:

Test Case #1 ##Concept Search

Query : experimental results on hypersonic viscous interaction

Output : [305, 26, 540, 570, 525, 1253, 573, 323, 63, 1395]

Documents Titles:

305 : hypersonic strong viscous interaction on a flat plate with surface mass transfer .

26 : inviscid leading-edge effect in hypersonic flow .

540 : use of local similarity concepts in hypersonic viscous interaction problems .

570 : on the boundary layer equations in hypersonic flow and their approximate solutions .

525 : on hypersonic viscous flow over an insulated flat plate with surface mass transfer .

1253 : hypersonic viscous flow near the stagnation point in the presence of magnetic field .

--------------------------------------------------------------------------------------------------------------------------

Test Case #2 ##Title Search

Query : properties of impact pressure probes in free molecule flow .

Output : [183, 906, 10, 1139, 1151, 1227, 1257, 355, 1107, 1391]

(Doc 183 has the exact title as the query)

-----------------------------------------------------------------------------------------------------------------------

Test Case #3 ##Search for text from documents

Query : manufacturing and maintainance of ideally sharp leading edges and noses is practically impossible

Output : [211, 900, 332, 167, 918, 465, 212, 1196, 544, 337]

(Doc 211 has the exact matching test)

----------------------------------------------------------------------------------------------------------------

Test Case #4 ##Question Search 1

Query : why does the compressibility transformation fail to correlate the high speed data for helium and air .

Output : [502, 68, 421, 343, 686, 1176, 389, 460, 1026, 1007]

(Doc 502 seems to have the answer (?))

--------------------------------------------------------------------------------------------------------------------------------

Test Case #5 ##Question Search 2

Query : can increasing the edge loading of a plate beyond the critical value for buckling change the buckling mode .

Output : [862, 15, 1177, 1015, 1023, 245, 642, 1069, 1398, 412]

(Doc 15 seems to have the answer)

-------------------------------------------------------------------------------------------------------------------------------