Web Systems and Algorithms

M. Sc. CSIT 2nd Semester

Central Department of Computer Science and Information Technology

Lab Assignment 1

Issued Date: - 2019 Nov 11 **Due Date: - 2019 Dec 01**

The Cranfield collection is a standard IR text collection, consisting of 1400 documents from the aerodynamics field. It is available from the class web page. (Check the "Links and resources" section) or goto the link *http://web.eecs.utk.edu/research/lsi/corpa.html* or you can consult your instructor for the materials or simply find the attached file.

1. Write a program that preprocesses the collection. This preprocessing stage should specifically include:
    a. Function that eliminates SGML tags
    b. Function that tokenizes the text. In doing this, pay particular attention to characters that need special handling, as discussed in class (. , - etc.). For this task, please use _your own_ implementation of a tokenizer.

2. Determine the frequency of occurrence for all the words in this collection and Answer the following questions:
    a. What is the vocabulary size? (i.e. number of unique terms)
    b. What are the top 10 words in the ranking? (i.e. the words with the highest frequencies)
    c. From these top 10 words, which are "meaningful" (i.e. they are not stopwords), and which ones you would eliminate as "stopwords".
    d. What is the minimum number of unique words accounting for half of the total number of words in the collection?

Example: if the total number of words in the collection is 100, and we have the following wordfrequency pairs: “the” – 30, “of” – 10, “a” – 10, “clear” – 8, “cut: - 7 etc. the answer to this question will be 3 (3 unique words account for half of the total 100 words)

3. Integrate the Porter stemmer and a stopword eliminator into your code. Answer again questions a-d from the previous point. (Check the "Links and resources" section for a link to various implementations of the Porter stemmer and to lists of stopwords).

4. Implement an indexing scheme based on the vectorial model, as discussed in class. The steps pointed out in class can be used as guidelines for the implementation. For weighting, use the TF/IDF weighting scheme.

5. For each of following given the ten queries provided on the class webpage, determine a ranked list of documents, in descending order of their similarity with the query. The output of your retrieval should be a list of (query_id, document_id) pairs.

Determine the average precision and recall for the ten queries, when you use:
- top 10 documents in the ranking
- top 50 documents in the ranking
- top 100 documents in the ranking
- top 500 documents in the ranking

6. A list of relevant documents for each query is provided on the following page, such that you can determine precision and recall.

Submission instructions:

- Write a README file including:
  - a detailed note about the functionality of the above programs,
  - complete instructions on how to run them
- Make sure you include your name in each program and in the README file.
- Make sure all your programs run correctly
- Submit your assignment, including programs and README file by the due date using the 'project' program.
- Class code is 2019CollegeNameRollNo, project Assignment 1

List of Queries

1. Is it possible to find an analytical, similar solution of the strong blast wave problem in the Newtonian approximation .
2. How can the aerodynamic performance of channel flow ground effect machines be calculated.
3. What is the basic mechanism of the transonic aileron buzz .
4. Papers on shock-sound wave interaction .
5. Material properties of photoelastic materials .
6. Can the transverse potential flow about a body of revolution be calculated efficiently by an electronic computer .
7. Can the three-dimensional problem of a transverse potential flow about a body of revolution be reduced to a two-dimensional problem .
8. Are experimental pressure distributions on bodies of revolution at angle of attack available .
9. Does there exist a good basic treatment of the dynamics of re-entry combining consideration of realistic effects with relative simplicity of results .
10. Has anyone formally determined the influence of joule heating, produced by the induced current, in magneto hydrodynamic free convection flows under general conditions . **List of Relevance Documents**

Query Number	Relevance Doc ID
1	27, 28, 262, 160, 20, 263, 654, 495
2	86, 194, 650, 649, 652, 624
3	64, 265, 65, 311, 496
4	64, 65, 496
5	463, 462, 497
6	266, 106, 196, 498
7	106, 196, 498
8	196, 197, 198, 498
9	32, 67, 164, 639, 715, 716, 719, 1379, 717, 499
10	87, 88, 104, 267, 268, 269, 270, 407, 408, 500