Group Assignment 2

CS 5180 Information Retrieval

Grade: 40 points.

* This is group work. You will work with your team members and make one submission.

**Due Date: 10/12/2022 11:59PM on Canvas**

**PROJECT DESCRIPTION**

In this assignment, you will improve on the indexes from group assignment 1, and implement the vector space model with cosine similarity for retrieving the top K documents (ranked retrieval) from the collection of document provided as an attachment. You will then compare exact Top K search with methods of inexact top K search. The specifications are:

**Part A: Exact Top K Retrieval [10]**

* In part A, you will construct a retrieval system for exact top K retrieval (henceforth denoted as Method 1).
* Use the TF-IDF weighting for the terms to construct the vectors for the documents and queries.
* Your index will be of the form
  + ***Term ID: [,(ID1, [pos1,pos2,..]), (ID2, [pos1,pos2,…]),….]***
* As in assignment 1, your program will accept the collection, process the documents and construct the index. Additionally, you will use the stop list (provided as an attachment) to weed out words from your dictionary. Note you should not index the document for each query. Simply do this at the beginning.
* Your program will then accept a free text query, generate the vectors for the documents and the query and compute the cosine similarity score for the documents. You can assume that you will not get single term queries.
* The program will retrieve and display the names of the top K documents for each query in decreasing order of their score.
* For each query, you will consider the results for your top K results as the *baseline results*. You will use this in part B
* You will also note the time taken to retrieve the results for each query.

**Part B: Comparison of Inexact retrieval methods with inexact retrieval [15]**

* You will also compare the performance of the exact top K with three inexact methods given below.
* Method 2 Champion List: Implement a champion list method which will produce, for each term, a list of r documents that is based on the weighted term frequency . Come up with a formula for *r*. Discuss this in your report. Remember the champion list is created only once for a collection.
* Method 3 Index Elimination: Implement index elimination by using only half the queries terms sorted in decreasing order of their IDF values.
* Method 4 Simple Cluster pruning: You will randomly pick leaders (where N is the number of documents in the collection) and then use them to implement the cluster pruning. Note that you need to select the leaders only once. For each query, you will select a leader closest to the query and then retrieve the top K results. Remember, if you do not get the top K results with the first leader, you will also look at the next best leader and so on.

**Part C: Comparison of Inexact retrieval methods with inexact retrieval [15]**

Experimentally compare the performance of your implementations of the exact retrieval with that of inexact retrieval methods. What are your performance measures? How are they measured and compared? In your report, you will provide justifications for these measures. Use graphs, as appropriate, to compare performance and describe insights and conclusion drawn from your results. Also include details for your implementation in the report that may have an impact on performance.

**Other instructions:**

* Implement in Python
* Comment your code appropriately
* You may reuse the code from earlier assignment

**Attachments:**

* 1. Collection of documents
* 2. List of 25 stop words
* 3. Skeleton code – implement the functions in the code. Use additional functions as needed.

Submission

**Deadline for submission: 10/12/2022 11:59 PM**

* Submit the following files on canvas as a .zip file.
  + A PowerPoint file that provides a description of your implementation, including pseudocode, experimental results. You will also provide the tasks performed by each team member.
  + The source code files including the following
    - index.py
    - README file --- briefly describing your code and how to execute the code.
    - Output.txt: containing 5 queries queries and the output generated by your code. This is for testing your code.
* Each team will submit a single copy on canvas.

**Each team will also schedule a presentation with the instructor (Week of Oct 17).** Your presentation will last about 10 minutes. The instructor will email you to set up the presentations.

**Anyone who misses the final presentation will not receive a grade for the assignment.**

**Late/re-submission**

* Team will have the option to resubmit or make a late submission with a penalty of 25%. Team who want to make use of this option will make their presentations during the last two weeks of the semester.