Group Assignment 4

CS 5180 Information Retrieval

Grade: 70 points.

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

**Due Date: 12/13/2022 11:59PM on Canvas**

**PROJECT DESCRIPTION**

**Page rank algorithm**

You will implement the page rank algorithm as discussed in class (see relevant lecture notes)

* Your program should read the directed graph of the webpages from a text file with the following format. You can also implement code that will automatically generate input files for various numbers of pages, so as to test your code effectively. A few example input files are provided with the assignment. Note the pageids range from 0 to N-1 where N is the total no. of pages.

(# of pages)

(# of links)

(src page 1) (dest page2) // an outgoing link from webpage page1 to page2

(src page1) (dest page1)

...

Figure 1 Format of the input file

* You will implement the iterative algorithm that uses the power method to generate the page rank. The teleportation rate =0.15.
* Your program should print the pagerank for the top 10 pages (both page id and pagerank value). It should also store the pageranks for all the pages (both page id and pagerank value) in the output file “out.txt”.

**Other instructions:**

* Implement in Python 3.0
* Comment your code appropriately
* You may reuse the code from earlier assignment

**Attachments:**

* Skeleton code – implement the functions in the code. Use additional functions as needed.
* Sample input files.(test1.txt and test2.txt)

**Files:**

Submit the following files:

1. pagerank.py

2. Output.txt: containing the output generated by your code for an input file. Also include the input file.

**Bonus Assignment: Distributed Indexing using Map reduce (100 pts)**

This bonus assignment will build on the lucene indexes that you build in the previous assignment. You will now scale this up for potentially large datasets by implementing distributed processing to index the data. You will need to implement the following

1. Utilize map reduce distributed processing paradigm to index the documents. You may use the Apache Hadoop framework or any other map reduce based library that works along with Lucene to generate the indexes.
2. You will use the twitter API[[1]](#footnote-1) to collect tweets that will form your data sets. Choose your own keyword to collect tweets. You will be indexing these tweets using lucene + map reduce. Collect a large enough data set that will be helpful for your performance study.
3. For your performance study, you will compare the run time to index the dataset with distributed processing (i.e. map-reduce) and with traditional processing. Provide analysis of the speedup and use other metrics as appropriate.

**Files:**

Submit the source code and the performance study results.

Submission

**Deadline for submission: 12/13/2022 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 files as requested/
* Each team will submit a single copy on canvas.

**Each team will also schedule a presentation with the instructor (Dec 14th).** Your presentation will last about 15 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**

* Late or re-submission options is not available for this assignment.

1. <https://developer.twitter.com/en/docs/twitter-api> [↑](#footnote-ref-1)