

## Project #3 : Implement PageRank

**Due Date : 11:59pm on Wednesday, October 29<sup>th</sup> , 2009**

The goal of Project #3 is to complete your search engine. You have to build PageRank of the given Wikipedia corpus using Map/Reduce. After that, complete the search engine using your web interface (Project #1), inverted index (Project #2) and PageRank (Project #3). Below is a step-by-step description of Project #3.

### 1. Build PageRank of Wikipedia

- A. Build a link graph of the given corpus (node = Wikipedia page; directed link = Wiki link tag in a page).
- B. Calculate PageRank; iteratively update the PageRank values of the articles.
- C. Output the PageRanks and article titles.

Details of the implementation are up to you.

- Link graph format, intermediate file format
- Choices of algorithm, parameters, number of iteration, convergence criteria

Write down details and reasons on the final report.

### 2. Build a simple search engine

Today's search engines use a mixture of complicate algorithms for information retrieval. Here we only use the PageRank algorithm to build a prototype of a search engine.

The search engine should run as follows:

1. Take a keyword (only one word)
2. Find all articles that contain the keyword by using the inverted index.
3. Sort those articles according to PageRank scores.
4. Display titles of the top 10 articles according to the PageRank scores.

Run your search engine on your own VM (assigned in Project #1).  
Do not launch on vc221.kaist.ac.kr

You have to demonstrate your search engine to TAs. Keywords will be picked randomly. Your search engine should have reasonable response time.

### 3. Cluster Information

Thanks to NexR, we will use CCI:U cluster.

Here is the procedure to use CCI:U cluster:

1. Make your account in CCI:U Open Course Labs located at <http://cciu.or.kr> (email validation required).
2. Join the KAIST CS492 course lab. (The direct URL is <http://cciu.or.kr/lab/kaist-p4sj4x/>)  
When TAs confirm and approve your joining, you will receive a notification email.
3. After the sysop or TAs put you in the "Project3" team, you will be able to see instance information on the team console page. The instance marked with a star is the master

that you can log in with the same account of the cluster used in Project #2.

4. As you can see, the job tracker status is on <http://143.248.160.245:50030> and the HDFS status is on <http://143.248.160.245:50070>.

You can still use the cluster you have used in Project #2. But we have plans for other tasks and cannot guarantee computing resources as before.

If you find any problems with CCI:U cluster, report them to [daybreaker@nexr.co.kr](mailto:daybreaker@nexr.co.kr).

NOTE: Due to the network topology of CCI:U cluster, you can no longer see each task's standard output and error directly because each data node has private IP only. But on the master node, you can use w3m (a text-mode web browser) to open the links starting with "host#" or "10.8.x.y" from the job tracker.

## 4. Deliverables

Send e-mail to <[hosung.golbaengi@an.kaist.ac.kr](mailto:hosung.golbaengi@an.kaist.ac.kr)>

Subject : [CS492] Proj3 studentID name

ex) [CS492] Proj3 20098765 GildongHong

Attached file name : studentID\_name

ex) 20098765\_GildongHong.zip

The attachment should include:

- Source code for PageRank
- Top 10 pages for each selected word according to the PageRank score. Selected words are the same as those of Project #2. (27 words)
- List of top 100 pages with their PageRank scores.
- URL of the search engine
- Final Report in English or Korean : limited to 4 pages (single column, 10pt)
  1. Describe your implementation of **Project #2** and **Project #3**. (How & Why)
  2. Write how to execute your code.
  3. Give us your feedback on:
    - What you've done & what you haven't done.
    - How long you expected this project would take to finish and how much time you actually spent on this project?
    - Acknowledge any assistance you received from anyone and any online resource

## 5. Tips

- Use of database is not mandatory. You can change your implementation of Project #1.
- It may take 10 to 20 minutes to generate a link graph and about 10 minutes for a single iteration of the PageRank algorithm.
- You can make your search engine more realistic (for fun).
  - Show contents of the article
  - Add a URL link to the title of the article so that users can navigate
  - Multi-word keywords
  - Fancy design