HOMEWORK 3:

LINK ANALYSIS

KEY - OVERVIEW

- Site-level PageRank Calculation on a given real Web sites sample collection
- Collection:
  - ~500,000 sites, 72,584,579 links (~1G original / ~250M zip file)
  - A small sample from 200,000,000 sites, ~5 Billion links
- Link-relationships have been extracted and given
  - URLs Have been transformed into SiteIDs
  - Out-links file “linkgraph.txt”: sourceID\tdestID\n
- An id-url mapping file “id-url.txt”
  - For your manual validation

RESULT REQUIREMENTS

- Using external libraries, frameworks are NOT allowed. Please develop your OWN solution.

+Page-rank result file

+ Source code, with detailed instruction about how to run it. Code must be successfully compiled & run (C/C++, Java, JavaScript, Perl, Python, PHP, Scala, ……)

+ Report (Efficiency of your program, score distribution……see details below)

Compress the result files in one ZIP archive

DETAILS OUTPUT

- File name convention:

  PR_“StudentId”.txt

  e.g. PR_2017999999.txt

  StudentId - please use your own student id

- Each line of the file should be in the format “SiteID\tScore\n”

  Example:

  15 0.00003

  3434 0.0000002

REPORT

- Efficiency of your program
  - Time for one loop, and Total running time
  - With your computer hardware configuration
- Distribution of PageRank scores (figure)
  - X: rank order from top one (with highest PR) to ~500,000
  - Y: log of PageRank score
- Comparison on changes of the top 20 PageRank sites and scores when loop = 5 and loop = 15(stable) (figure)
- Analyses on the top 10 sites with highest PageRank score one bye one (site's url can be found in the id-url.txt)
  - Why it gets a high PR score, many in-links?
  - Or being linked by a/several site with high PR score(s)?
  - Or any other reasons?
- Any observation/conclusion/discussion what you think may be important.

TIPS

- Try first on a small sample for the correctness of your program (we also provide very small set for test, with PR score after convergence)
- Deal with pages with no out-links and no in-linker
- Convergence condition: 15 loops
- On the whole set, each loop may take ~30 minutes.
- You'd better to find some way to optimize your calculation

DEADLINE:

10:00am (UTC+08:00, Beijing Time) Mar. 22, Monday
Submit your homework to web learning platform of thu:

Http://learn.Tsinghua.Edu.Cn our course section “Assignment”(课程作业)

QUESTIONS ?