

Homework 2 - Grading Guidelines

Total Points 10

Crawl Report

- 1) # fetches attempted = # fetches succeeded + # fetches failed or aborted
- 2) Number of rows of fetch_*.csv statistics should be close to 20,000 (close means within a 1,000 or 2,000; if not explain why)
- 3) # unique URLs extracted = # unique URLs within news site + # unique URLs outside the news site
- 4) Status code - 200 codes should be equal to fetches succeeded
- 5) Number of files in the size statistics should be less than or equal to the number of fetches succeeded.
- 6) Number of files in the content types should be less than or equal to the number of fetches succeeded.

CSV files

- 7) Inspect fetch.csv, visit.csv: all of the data in both files will be cross validated against the crawl reports.