Homework 2 - Grading Guidelines

Total Points 10

Crawl Report

1) Number of threads used
2) # fetches attempted = # fetches succeeded + # fetches failed or aborted
3) 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)
4) # unique URLs extracted = # unique URLs within news site + # unique URLs outside the news site
5) The total number of URLs extracted should be equal to the number of outgoing links encountered on the fetched pages.
6) Status code - 200 codes should be equal to fetches succeeded
7) Number of files in the size statistics should be less than or equal to the number of fetches succeeded.
8) Number of files in the content types should be less than or equal to the number of fetches succeeded.

CSV files

9) Inspect fetch.csv, visit.csv: all of the data in both files will be cross validated against the crawl reports.
10) **Note**: column headers need to be included