Homework 2 - Grading Guidelines

Total Points 10

Crawl Report
1) # fetches attempted = # fetches succeeded + # fetches failed or aborted (2 points)
2) Number of rows of fetch.csv statistics should be close to 5000 (close means within 500; if not explain why) (1 point)
3) # unique URLs extracted = # unique URLs within School + # unique USC URLs outside School + # unique URLs outside USC (1 point)
4) Status code (1 point) 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. (1 point)
6) Number of files in the content types should be less than or equal to the number of fetches succeeded. (1 point)

CSV files
7) Inspect fetch.csv, visit.csv, urls.csv (3 points).
   a) All the urls in visit.csv and fetch.csv should be within the school.
   b) urls.csv will have urls outside the school.
   c) number of urls in visit.csv <= fetch.csv <= urls.csv.