CSCI 572 ASSIGNMENT 2 RUBRICS

Web Crawling

Total Marks: **10 marks**
The distribution of marks is shown in the table below:

Submission Requirements	**Marks Allotted**
Task 1: CrawlReport_NewSite.txt	5 mark/s
Task 2: fetch_NewsSite.csv	2 mark/s
Task 3: visit_NewsSite.csv	2 mark/s
Source Code	1 mark

Late Submission Penalty

- 10% penalty per day after the deadline

Rubrics – Web Crawling

- Task 1: CrawlReport_NewSite.txt:
  - If this file is missing: -5 mark
  - For each part missing: -1 mark
    - Fetch Statistics
    - Outgoing URLs
    - Status Codes
    - File Sizes
    - Content Types
  - Multiple threads not used : -0.5 mark **(Note: this does not hold if you are using scrapy)**
  - # fetches attempted = # fetches succeeded + # fetches failed or aborted, if not: -1 mark
  - Number of rows of fetch_*.csv statistics should be close to 20,000 (close means within 1,000 or 2,000). If not explain why; if no explanation: -0.5 mark
    - # unique URLs extracted = # unique URLs within news site + # unique URLs outside the news site. If not: -1 mark
  - The total number of URLs extracted should be equal to the sum of outgoing links encountered on the fetched pages, if not: -0.5 mark
  - Status code - 200 codes should be equal to fetches succeeded, if not: -0.5 mark
  - Number of files in the size statistics should be less than or equal to the number of fetches succeeded, if not: -0.5 mark

    - Number of files in the content types should be less than or equal to the number of fetches succeeded, if not: -0.5 mark
- Task 2: fetch_NewsSite.csv:
    - If this file is missing: -2 mark
    - If less than 18K rows present: -0.5 marks
    - Verify e visited (Search "https://news_site" and see if results match the number of rows): -0.5 marks
    - For each missing column OR partially filled columns (4.5K or more empty values): -0.5 mark
        - URL, Status
- Task 3: visit_NewsSite.csv:
    - If this file is missing: -2 mark
    - Check total number of rows in this file should be less than or equal to the "fetches succeeded " in CrawlReport: -0.5 marks
    - Verify no external URLs are visited (Search "https://news_site" and see if results closely match the number of rows): -0.5 marks
    - For each missing column OR partially filled columns (4.5K or more empty values): -0.5 mark
        - URL, Size (Bytes), # of Outlinks, Content Type
- Task 4: Source Code Files:
    - If this file(s) is missing: -1 mark