Additional FAQ from Piazza forum

1. I'm working on the HW2 and need clarification on how to calculate the "#fetches succeeded" metric.

I noticed a conflict between the assignment PDF and the grading rubric:

Assignment PDF: States that "#fetches succeeded" is based on all URLs returning HTTP status code of 2XX(all success codes, e.g., 200, 201, 204).

Grading Rubric: States that "Status code 200 codes should be equal to fetches succeeded."

For the final value of "#fetches succeeded" in our report, should we strictly count:

A) All 2XX status codes (200, 201, 204, etc.)?

B) Only the 200 (OK) status codes?

A: Please follow instructions on Bytes..(not from the grading Rubrics)

Question:

Hello,

I set my crawler configuration according to the HW2 requirements (maxPagesToFetch = 10,000, maxDepthOfCrawling = 16, politeness delay = 2000ms, 8 threads, seed = nytimes.com). The crawler runs successfully, fetches and visits pages correctly, and generates the CSV files and CrawlReport.

However, the run only reached about 6,400 fetches before finishing. It seems the queue of available URLs was exhausted (likely due to domain constraints, robots.txt, or duplicates), so it stopped before reaching 10,000.

1. Is it acceptable to submit with ~6,400 fetches, as long as I explain in the report why the number is less than 10,000?
2. Or should I add more seeds / try another news site to ensure I reach closer to 10,000?

Answer: crawled USA, where my major parameter difference from yours is my politeness delay = 1.5 sec. also, I did 10 threads. I retrieved 20,000.

Downside, it took about 45 min to run. Hope thats helpful.

Q3. The description for the submission says to submit the crawler source files. Does that include the code file also?

A Hi yes.