Assignment 4 Group 33

Aim

Learn details of Hyper-Text Transmission Protocol-HTTP, Secured HTTP (HTTPS) and develop services using socket programming

Tech stack used

* Python – need not stick to python/Java/C++ can use any language
* Using csv files for storage… better to use postgres? Or something else?
* How to compile a report? ChatGPT plugin to format facts and source links and graph links into text, https://platform.openai.com/docs/introduction/key-concepts

Use case

Education: Popularity of private tuition in Southeast Asia/ Asia.

Questions that will be answered in the report:

* What percentage of students attend private tuition in Singapore?
* What are the trends in age groups involved in, popular subjects taught in, and fees spent on private tuition?
* How does expenditure and interest in private tuition differ between countries?

Types of data we need

* Numerical data (percentage, numbers, statistics, age, fees)
* Keyword filter “private tuition”, “education”, “tuition fees”, “primary”, “secondary”, etc.
* Subject names (English, Maths, Science, Physics, Chemistry)

Data Analytics

* Create graph for trend in private tuition fee expenditure in Singapore over time?
* Model trend in tuition enrolment over time
* Bar graph for age groups enrolled in tuition.
* Bar graph for popular subjects by age group
* Plot on map geographic locations of schools used in survey.

Input to program

Seed.txt: Contains initial URLs to crawl.

Visited.csv: pages that have been found and a crawler has started to visit it (note: crawlers should not revisit URLS in this file, one atomic action on this file = read url to check if found url is unique + write url to file if it is unique)

Found.csv: store information found by crawler (note: need to make sure only one crawler can access this file at a time, one atomic action on this file = write to file)

Final output

1 page report on statistics of tuition industry in Singapore.

Time limit

Run parallel crawler for xxx minutes only and then force to stop

What we have built so far

Crawlers lacking parallelization,

Difficult to get precise geolocation – currently using external API.

Timeline for development

Work assignment

What to submit

Submit one ZIP/RAR file (change filename to your Group Number) containing:

* your source code
* short video demo showing your DB/text file/screen showing new URLs found (minimum 10 new urls)
* one PDF report
* short text file with instructions for compiling/running your source code

Marking scheme

(3 marks) Starts with the initial set of URLS in a text file/DB.

(7 marks) Coordinated access by multiple Threads to the text file/DB.

(10 marks) Adds newly found URLs to the text file/DB.

(5 marks) Response times to the severs, IP addresses and geolocation of the servers are printed (either in screen or in text file/DB).

(8 marks) One page report on useful data collection/statistics (open ended requirement)

(2 marks) Well written code with in-line comments.

Sources used so far

<https://www.geeksforgeeks.org/remove-all-style-scripts-and-html-tags-using-beautifulsoup/>