ECS 152A: HW1 Part 2

Nghi Dao (921147615), Bian Lee (920763430)

November 2024

Original implementation (collection) : `part2_collection_original.py`
Revised implementation (collection): `part2_collection_refined.py`
Original implementation (analysis) : `part2_analysis_original.py`
Revised implementation (analysis): `part2_analysis_refined.py`
ChatGPT conversation link: here
HAR files: `har_file_directory`
Analysis output: `analysis.txt`

Collecting the HAR files

We used our script to collect the HAR files from the first 1000 websites listed on the CSV.

- We first started the BrowserMob proxy server after specifying the path of the binary. Then, we set a timeout for the proxy and implemented a retry mechanism in case of errors. We kept a count of retries and specified a maximum retries limit of 3, ensuring that restarting the BrowserMob proxy occurred only up to the specified number of times.
- We configured our Selenium WebDriver to operate in headless mode to avoid having the webpage physically pop up on the screen for every site access.
- To collect the HAR files, we read the CSV placed in the same directory and parsed the domains of the first 1000 sites, saving them into a 'sites' array. The script checked whether the '.har' file for a site already exists in the designated folder (if it does, it just skips over) and then began capturing and saving into the har_file_directory folder..

Analysis

- The output result of running the analysis script can be found in analysis.txt where it lists the top 10 most commonly seen third-party domains, as well as the top 10 third-party cookies.
- We used Cookiepedia to look up each of the top 10 cookies to understand their functionality. See the tables below for the analysis output:

Rank	Domain	Request Count
1	doubleclick.net	2763
2	media-amazon.com	2497
3	google.com	1999
4	googletagmanager.com	1666
5	rubiconproject.com	1379
6	googlesyndication.com	1377
7	gstatic.com	1284
8	com.br	1241
9	cookielaw.org	1081
10	bing.com	1073

Rank	Cookie Name	Count	Intended Functionality (researched with Cookiepedia)
1	_ga	3533	Used by Google Analytics to distinguish unique users in order to store and count pageviews, and conduct analytics.
2	__cf_bm	2896	Used by Cloudflare to distinguish bot traffic and human web traffic to ensure secure browsing.
3	ar_debug	2738	Debugging cookie used by advertisers to track ad performance and troubleshoot application behaviors.
4	receive-cookie-deprecation	2330	Indicates acceptance of cookie usage or tracking policies.
5	_gid	2249	Google Analytics cookie used to distinguish users for analytics. Expires after 24 hours
6	_gcl_au	2114	Google Conversion Linker cookie; tracks advertising effectiveness and conversions from advertising services .
7	IDE	2014	DoubleClick cookie by Google for optimizing advertising and user targeting.
8	_abck	1899	Used by Akamai to detect and prevent automated attacks.
9	bm_sz	1899	Bot management session cookie by Akamai for performance and security (protect against automated threats by storing session specific information).
10	MUID	1513	Microsoft User Identifier; cookie to identify unique users across sessions for ad targeting.