Billy Ouattara (920603707)

Jose Gavidia (921477055)

Files: crawling.py, crawlingGPT.py, iterate.py, iterateGPT.py

Har files folder: harFiles

RESULTS

After implementing and running our code, which can be found in the “crawling.py” file, we were able to retrieve the one thousand files needed for our analysis. Unfortunately, some of the websites were throwing exceptions such as “TimeoutException”, and “connectionRefusedError”, but we were able to catch these exceptions and continue with our process of collecting the .har for a thousand websites. Once we successfully collected a thousand har files, each representing a website, we created the “iterate.py” file to conduct our analysis of the data.

These were our results:

```
Total requests: 60993

TOP 10 THIRD PARTIES:
 www.googletagmanager.com: 1079
 cdn.cookielaw.org: 969
 res.cdn.office.net: 834
 www.ifood.com.br: 768
 fonts.gstatic.com: 617
 pagead2.googlesyndication.com: 611
 www.google-analytics.com: 607
 content-autofill.googleapis.com: 590
 www.google.com: 589
 cm.g.doubleclick.net: 511

TOP 10 THIRD PARTY COOKIES:
 _ga: 1909
 _gcl_au: 1762
 IDE: 1721
 ar_debug: 1669
 audit: 1586
 khaos: 1435
 _fbp: 1371
 _gid: 1361
 NID: 1143
 __cf_bm: 1074
```

We found a total of 60993 requests to third party websites. These websites are notability in charge of advertising, targeting, analytics, marketing, cookie consent, among other functions.

The functionality of our top-10 third party cookies is as follows:

1) _ga: This cookie is part of Google Universal Analytics. It is used to identify users by generating and assigning a unique ID. It is notably used to calculate visits and session data, from which analytics reports are generated. It expires in two years. According to Cookiepedia, the main purpose of this cookie is performance.
2) _gcl_au: According to Cookiepedia, it is “used by Google AdSense for experimenting with advertisement efficiency across websites using their services”. The main purpose of this cookie is targeting and/or advertising.
3) IDE: Owned by Doubleclick, and the main purpose of this cookie is targeting and advertising.
4) ar_debug: Unfortunately, Cookiepedia states that “there is not yet any general information about this cookie”, although this cookie name has been found on 111 websites.
5) Audit: As for ar_debug, Cookiepedia states that there is no general information about this cookie and its main purpose is unknown. Interestingly, although unknown purpose, its name have been found on 7273 websites set by 19 host domains.
6) Khaos: This cookie carries information about how the end user interacts with the website and records information about any advertising that the end user may have seen before visiting the current website.
7) _fbp: Used by Facebook for targeting and advertising. It delivers advertising products from third party advertisers.
8) _gid: Cookiepedia doesn’t say much about this cookie, besides that it is used for performance. Surprisingly though, “cookies with this name have been found on 169,506 websites.
9) NID: This domain is owned by Google, and the main purpose of this cookie is targeting and advertising. Google gathers data from the users by using this cookie, so they can sell advertising to other institutions based on the interests of the users.
10) __cf_bm: Functional cookie used by cloudflare.com.