


# Assignment 1

 Grade weight: **3/10** of the final grade

 Due: **8 March** 2022 (23:59, CET)

## Start fresh and pick a website

- Create a new Chrome/Chromium profile for the assignment
- Pick a news website **or** an online shop from the lists below.

News websites	Online shops
<ul style="list-style-type: none"><li>• <a href="https://www.nu.nl/">https://www.nu.nl/</a></li><li>• <a href="https://www.ad.nl/">https://www.ad.nl/</a></li><li>• <a href="https://www.telegraaf.nl/">https://www.telegraaf.nl/</a></li><li>• <a href="https://nos.nl/">https://nos.nl/</a></li><li>• <a href="https://www.rtlnieuws.nl/">https://www.rtlnieuws.nl/</a></li><li>• <a href="https://www.volkskrant.nl/">https://www.volkskrant.nl/</a></li><li>• <a href="https://www.nrc.nl/">https://www.nrc.nl/</a></li><li>• <a href="https://www.metronieuws.nl/">https://www.metronieuws.nl/</a></li><li>• <a href="https://www.trouw.nl/">https://www.trouw.nl/</a></li></ul>	<ul style="list-style-type: none"><li>• <a href="https://www.coolblue.nl/">https://www.coolblue.nl/</a></li><li>• <a href="https://www.ah.nl/">https://www.ah.nl/</a></li><li>• <a href="https://www.zalando.nl/">https://www.zalando.nl/</a></li><li>• <a href="https://www.wehkamp.nl/">https://www.wehkamp.nl/</a></li><li>• <a href="https://www.amazon.nl/">https://www.amazon.nl/</a></li><li>• <a href="https://www.jumbo.com/">https://www.jumbo.com/</a></li><li>• <a href="https://www.aboutyou.nl/">https://www.aboutyou.nl/</a></li><li>• <a href="https://www.debijenkorf.nl/">https://www.debijenkorf.nl/</a></li><li>• <a href="https://www.hm.com/">https://www.hm.com/</a></li></ul>

## Capture the HTTP traffic

For the website you chose:

1. Start with a fresh profile (clear all browsing data)
2. Open the Devtools/Network panel
3. Check "Preserve log" (that'll retain all requests made during a session)
4. Load the website's homepage; accept all cookies/data processing, dismiss other potential dialogs (permission to send notifications, location access, email signup etc.)
5. Scroll down until the bottom of the page
6. Click on an article or a product page (multiple clicks are okay if you have to). Avoid external links, the inner page should be under the same first-party domain as the homepage
7. Scroll down until the bottom of the page
8. Save all HTTP request/responses as HAR to a file using the following naming convention: example.com.har. No www. or other prefixes; just domain\_name.har.

### Capture the HTTP traffic with an adblocker/tracking protection add-on

Now, install [uBlock Origin](#) or [Adblock Plus](#) on Chrome/Chromium. Repeat steps 1-8 **starting again with a fresh profile**, this time with the add-on installed. Name the second HAR file as `domain_name_adblocker.har`. Now you should have two HAR files: one with the adblocker and one without.

### Analyze the HAR Data



Write an analysis script as a Jupyter Notebook (.ipynb) or as a standalone Python script (.py) that processes the captured HAR files and outputs the following as two separate JSON files, each containing a (Python) dictionary of results. The overall processing pipeline should look like the following:

- HAR -> analysis -> results dict -> serialize to JSON

The dictionary serialized in each JSON should contain the following keys:

- `num_reqs`: Integer, number of requests (observed in the HAR file)
- `num_requests_w_cookies`: Integer, number of requests with cookies
- `num_responses_w_cookies`: Integer, number of responses that set at least one cookie
- `third_party_domains`: list of distinct third-party domains (eTLD+1)
- `domains_w_cookies`: list of distinct domains that set at least one cookie
- `server_countries`: list of distinct server countries
- `xorigin_cookie_domains`: list of domains that set at least one cookie with `SameSite=None`, and `lifespan >=3 months`
- `requests`: a list of dict containing the following request details:
  - `request_domain`: String; e.g. `example.com`
  - `server_country`: String; e.g. `Germany`; “unknown” if server IP is unavailable
  - `server_in_eu`: Boolean; whether the server is located in the EU or not; “unknown” if server IP is unavailable
  - `num_request_cookies`: Integer
  - `num_response_cookies`: Integer
  - `is_tracker`: Boolean; whether the domain is listed in [EasyList](#) or [EasyPrivacy](#) “*just domains*” blocklists
  - `url_first_128_char`: String; e.g. `https://example.com/pixel.gif`

### Tips:

- Don't use "Save all as HAR with content" when saving the HAR files. Rather use "Copy -> Copy all as HAR", and paste to an empty file (v0.1)
- The requests list will contain a dict for each request-response pair
- Unless specified, "domain" means eTLD+1
- Make sure you open the Devtools/Network panel and check "Preserve log" before loading any page
- Comment your code when what you are doing is not very obvious
- DRY: Don't Repeat Yourself. Break your code into reusable small functions
- Avoid deep indentations
- Use meaningful variable and function names
  - a.  good: request\_domain, response\_headers, get\_country\_by\_ip\_address
  - b.  not good: foo, bar, tmp, a, do\_stuff, get\_data

### Practicalities

- Upload a zip file containing the files listed below (a-f). Name the zip file after your student number; e.g. s012345.zip. File names should look like this:
  - a. example.com.har
  - b. example.com.json
  - c. example.com\_adblocker.har
  - d. example.com\_adblocker.json
  - e. s012345.ipynb OR s012345.py (analysis script)
  - f. requirements.txt: Python packages required to run your script, if any
- You can assume the following files will be available in the same folder as your code. You do **not** need to upload them in your zip file. When testing your code, we will extract your zip file and copy the below files (a-d) to your folder (v0.3):
  - a. easylist-justdomains.txt ([link](#))
  - b. easyprivacy-justdomains.txt ([link](#))
  - c. dbip-country-lite.mmdb ([link](#))
  - d. GeoLite2-Country.mmdb ([link](#)): Alternative to (c), requires a MaxMind account
    - Note: using either c or d is acceptable
- Your code should **not** make any calls to online APIs (v0.3). It should be able to work offline
- You are free to use publicly available Python packages (v0.2)
- You can print log messages from your code (you don't have to) (v0.2)
- Your code should work with Python 3
- Your code should be able to run without any command line parameters
  - a. Hard-code the HAR filenames in your code, assume they are in the same folder as the analysis script/notebook
  - b. Python script: Running "python s012345.py" once should re-generate the exact JSON outputs
  - c. Jupyter Notebook: Should run without any intervention and re-generate the exact JSON outputs

### Help / Office hours

- Wednesdays between 11h-13h, **starting from Feb 23rd**
- Zoom link:  
<https://radbouduniversity.zoom.us/j/84643591429?pwd=T2MwYU9mdDZjT0JhV0tJeTRWTStrZz09>

🍀 Good luck! 🍀