Assignment 1

ℹ️ 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
● https://www.nu.nl/ ● https://www.ad.nl/ ● https://www.telegraaf.nl/ ● https://nos.nl/ ● https://www.rtlnieuws.nl/ ● https://www.volkskrant.nl/ ● https://www.nrc.nl/ ● https://www.metronieuws.nl/ ● https://www.trouw.nl/	● https://www.coolblue.nl/ ● https://www.ah.nl/ ● https://www.zalando.nl/ ● https://www.wehkamp.nl/ ● https://www.amazon.nl/ ● https://www.jumbo.com/ ● https://www.aboutyou.nl/ ● https://www.debijenkorf.nl/ ● https://www.hm.com/

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
    - server_in_eu: Boolean; whether the server is located in the EU or not
    - 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:

- 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 before loading any page
- Make sure you check "Preserve log" on the Devtools/Network panel
- File names should look like this:
    - example.com.har, example.com.json
    - example.com_adblocker.har, example.com_adblocker.json
    - s012345.ipynb *OR* s012345.py (analysis script)
    - requirements.txt: Python packages required to run your script, if any

- Upload a zip file containing the above files. Name the zip file after your student number; e.g. s012345.zip

Coding style and practicalities

- 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
  - ❌ not good: foo, bar, tmp, a, do_stufff, get_data
  - ✅ good: request_domain, response_headers, get_country_by_ip_address
- Your code should work with Python 3
- Your code should be able to run without any command line parameters
  - Hard-code the HAR filenames in your code, assume they are in the same folder as the analysis script/notebook
  - Python script: Running "python s012345.py" once should re-generate the exact JSON outputs
  - 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!** 🍀