Task 1

1. Create a script crawler that visits www.cnn.com browses 10 random links,(e.g. cnn.com/sports, cnn.com/money/article1.html). For this first session store all cookies to a cookie.out file,store all http traffic to a json output (hint : browsermob proxy, mitmproxy), and all the visited pages to html form.
   Parse the json output and save to file.
2. Visit www.ebay.com and trigger the next queries : macbook pro, dell xps, nike shoes, meller glasses.
   For each query result go to the 3 first pages, store the same as task1 and also a screenshot of the page.

3. Visit www.gmail.com and login to your email.
   Do the same with facebook.

4. Visit www.netflix.com find the sign in button and login :
   a. With gmail
   b. With facebook

5. Visit https://www.news.gr/, 20 links, 4 times a day(every 6 hours e.g. or on standard timeslots) for three days (hint : crontab) and store the same as with task1 and task2.
   Report the following :
   a. All the third party trackers.
      domain name -IP (hint: dns resolver).
   b. On average how many trackers (unique) can you discover in every visit?
   c. Plot the number of trackers found on every crawl of the day, for the total of three days.(hint : **gnuplot**).
   d. Use the 1st crawl as baseline, and report the number and the domain names of the “new” trackers that you observed in every crawl.

e.g:
1st crawl : 40 trackers
2nd crawl : 5 new trackers: a.com,b.com,d.com…
Do you see any change in the volume of trackers?Do you observe new domains?Why do they change?

Final task

Use all the above implementations to create a real crawler :)
Your crawler will run for 5 days three times a day, and will visit **15** domains(choose whatever you want-popular sites) and 10 sublinks for each domain.You will have to store everything reported above and also:

- Login to every site that has a login page on it's homepage (facebook/gmail) and then visit the subdomains.
- Your script has to be more "human" than a bot, or else you might get a ban on some domains.Think how could you do that.How does a domain understands that a bot/crawler does the requests?
- Compare the third party trackers that you find on every domain.Plot :
    - The 5 "top" domains with the largest number of trackers.
    - The number of trackers found on every domain each day of the 5 experimental days.
    - The most popular trackers found across domains.

Do you notice any correlation between the popularity of the site and the number of trackers?

Hint:

→ Use disconnect/easylist- adblock/ghostery/ google for ad-domains list :) , to identify the trackers.