Ahmed Irtija (920742338)
Python: webCrawler.py

Identify the top-10 most commonly seen third-parties across all sites.

- 

,
- Here are the top 10 third party domains that we got from running our script. We used the tldextract library to be able extract the info. We had to use the urllib.parse library to parse the url from the har file. This script works once the har folder has the contents it's supposed to have. We can see the result of each domain as fonts.gstatic.com being the most popular one.

Domain Explanation

- **fonts.gstatic.com** - Google uses this subdomain to store font files used on websites.
- **www.google-analytics.com** - A Google web analytics service that tracks and reports website traffic and also the mobile app traffic & events.
- **www.googletagmanager.com** - A tag management system that has the same functionality as the Google tag and lets you configure and instantly deploy tags on your website or mobile app from an easy-to-use web-based interface.
- **www.youtube.com** - A video sharing platform and social media platform
- **www.google.com** - Google search engine
- **fonts.googleapis.com** - An interactive directory of free hosted application programming interfaces for web fonts.
- **pagead2.googlesyndication.com** - It is Google's tracker. It provides advertising or advertising-related services such as data collection, behavioral analysis or retargeting.
- **static.parastorage.com** - It allows users to create websites and mobile sites through the use of online drag and drop tools.
- **googleads.g.doubleclick.net** - It is Google's advertising company. They use this referral scheme in order to track your browsing habits (to learn what words you searched for that led you to click on their ad), and for Google to collect a referral bonus for displaying such a good advertisement that you clicked on it.
- **www.controlant.com** - It is a global leader in the digital transformation of pharma supply chains

```
Top 10 Third-Party Cookies:
_ga: 3495
_gid: 2142
_gcl_au: 1338
1P_JAR: 1272
IDE: 1161
PHPSESSID: 1147
VISITOR_INFO1_LIVE: 937
YSC: 937
__cf_bm: 890
_gat: 627
```

- Here are the top 10 most common used cookies from all 1000 websites which we got using the similar method. Now using the Cookiepedia, here are the functionality for each

Cookie Functionality

- **_ga** - This cookie is used to distinguish unique users by assigning a randomly generated number as a client identifier. Link to Cookiepedia
- **_gid** - The main purpose of this cookie is: Performance Link to Cookiepedia
- **_gcl_au** - Used by Google AdSense for experimenting with advertisement efficiency across websites using their services. The main purpose of this cookie is: Targeting/Advertising. Link to Cookiepedia
- **1p_JAR** - This cookie carries out information about how the end user uses the website and any advertising that the end user may have seen before visiting the said website. The main purpose of this cookie is: Targeting/Advertising. Link to Cookiepedia
- **IDE** - This domain is owned by Doubleclick (Google). The main business activity is: Doubleclick is Google's real time bidding advertising exchange. The main purpose of this cookie is: Targeting/Advertising. Link to Cookiepedia
- **PHPSESSID** - PHP session cookie associated with embedded content from this domain. The main purpose of this cookie is: Strictly Necessary. Link to Cookiepedia
- **VISITOR_INFO1_LIVE** - This cookie is set by Youtube to keep track of user preferences for Youtube videos embedded in sites;it can also determine whether the website visitor is using the new or old version of the Youtube interface. The main purpose of this cookie is: Targeting/Advertising. Link to Cookiepedia
- **YSC** - This cookie is set by YouTube to track views of embedded videos. The main purpose of this cookie is: Performance. Link to Cookiepedia
- **__cf_bm** - is a cookie necessary to support Cloudflare Bot Management. (Had to use different source to find info)
- **_gat** - This cookie name is associated with Google Universal Analytics, according to documentation it is used to throttle the request rate - limiting the collection of data on high traffic sites. It expires after 10 minutes. The main purpose of this cookie is: Performance. Link to Cookiepedia