**ATG Task #1 Report**

-Devansh Shukla

13 Nov’19

Objective:

Scrape the following websites :

- eventshigh.com

- insider.in

- naadyogacouncil.com

for all the details about any 10 interesting events and store them in a database.

Procedure:

- Visited the website and imported the html using BeautifulSoup library

- Extracted all URLs from the website and performed checks on them

- Found 10 event URLs and stored them in a list

- Stored all other URLs that underwent the check in a non-interesting URL list

- Requested each interesting URL one at a time

- Found structured data inside the soup by searching for JSON script tags

- Loaded the JSON into a python dictionary

- Used the JSON dictionary to scrape useful data and stored it in a list of dictionaries

- Used pandas to convert the list into a DataFrame

- Exported Data to CSV

Details:

All the websites supported structured data for their individual event page. The home pages that didn’t support structured data were tackled by scraping all the URLs from them and visiting the particular event URLS.

The scraped data that was recovered from the JSON was stored in a list of dictionary in key, value format to make finding data easier.

Since, the data wasn’t dynamic and operations weren’t needed to be performed on the database, I refrained from using sqlite3 or psycopg2 and simply exported the DataFrame to a CSV file.