Internship Project – Web Scraping and Python App Development

Project Title: "From Digital Governance to Tourist Delight: Unravelling the Tourism Impact of India's Smart Cities"

Deliverables:

- Phase 1: Clean Scraped Data (in excel or other formats) to be submitted by the end of **February 2024**
- Phase 2: A python app for dynamically scarping data from specific platforms (based on keyword or other relevant parameters) by **May 2024**

Phase 1 - Data Collection

- In this phase, the team will work on scraping data from the following platforms – Tripadvisor (mandatory) and Google Reviews (mandatory)
- Identify and learn the process for scraping data from these platforms
- Using a programming language (preferably, Python), write the code to scrape the data based on few specific parameters (like city, location, landmark, hotel, etc…)
- The data has to be structured and exported to excel.
- The cities chosen for the first phase – ***List of cities will be communicated in the email***
- Collect data on top 30 tourist attractions in each of these cities
- Deliverable: Clean scraped data by the **end of February, 2024**.

Phase 2 – Web Application Development:

- Build a user-friendly application for data scraping, automating the process developed during Phase 1.
- You may choose to develop separate applications for each portal (TripAdvisor and Google Reviews) or a single application for both platforms.
- Ensure the application is scalable and adaptable for future use.
- The application should be user friendly and should be easy to use without coding knowledge.
- Document the application development process, including code documentation and user manuals.

Guidelines for Trip Advisor Reviews

This involves creating 2 web scrapers – One to get the details of all attractions in city and another to extract review information for each attraction in the city

Trip Advisor Web Scraper 1

1. Go to Tripadvisor.com (please ensure it is .com and not .in)
2. Search for the city – eg. Varanasi
3. You will get a page like – https://www.tripadvisor.com/Attractions-g297685-Activities-Varanasi_Varanasi_District_Uttar_Pradesh.html
4. This page lists all attractions in Varanasi.
5. The scraper has to get all details of all attractions from this page.
6. In the app, the input should be the link similar to point 3. The scraper should run and extract data in csv file (TripAdvisor - Web Scraper 1 Output.csv).
7. In the app there should be an option to provide multiple links (cities) and the data should be systematically downloaded for each city (refer example below).

Trip Advisor Web Scraper 2

1. Go to Tripadvisor.com (please ensure it is .com and not .in)
2. For each attraction there will be a page – eg. https://www.tripadvisor.com/Attraction_Review-g297685-d319858-Reviews-Ganges_River-Varanasi_Varanasi_District_Uttar_Pradesh.html
3. This page has all reviews for this specific attraction.
4. The scraper has to get all details of all reviews of this attraction from this page.
5. In the app, the input should be the link similar to point 3. The scraper should run and extract data in csv file (TripAdvisor - Web Scraper 2 Output.csv).
6. In the app there should be an option to provide multiple links (attractions) and the data should be systematically downloaded for each attraction.

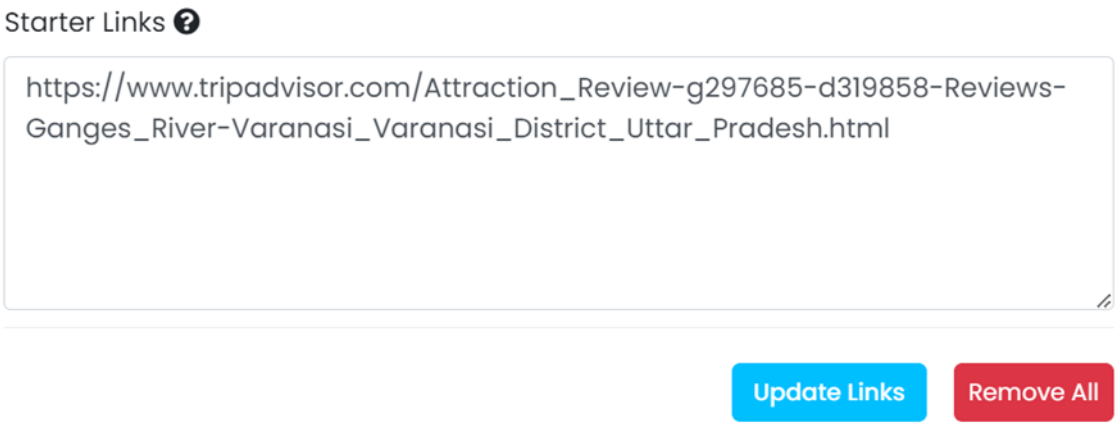
Guidelines for Google Reviews

Google Reviews Web Scraper 1

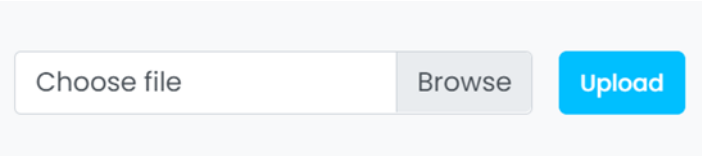
1. Google reviews for a place can be accessed using a link such as: https://www.google.com/search?q=saheliyon+ki+bari#lrd=0x3967e5d77db7d277:0x9d61a020249ce441,1,,,,
2. In this link, the place is 'Saheliyon ki bari' and the link directly takes you to the reviews tab – This will be the input link.
3. This page has all reviews for this specific attraction.
4. The scraper has to get all details of all reviews of this attraction from this page.
5. In the app, the input should be the link similar to point 1. The scraper should run and extract data in csv file (Google Reviews - Web Scraper 1 Output.csv).
6. In the app there should be an option to provide multiple links (attractions) and the data should be systematically downloaded for each attraction.

Example User Interface:

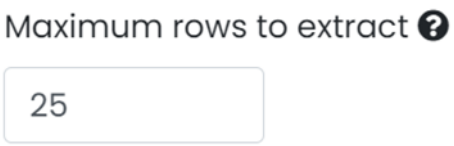
- Please note that this is just a sample. You may design the user interface that is easy to use and has several options.



- In the text box, it should be possible to enter multiple links
- Also, add an upload file option to upload a .txt file with links in each line



- In the app, add another input, where one can specify the number of rows to be extracted (0 indicates all possible rows to be extracted)



Remember:

- By end of Feb 2024, your team is only expected to send the data in csv file (after extracting the csv files for each of the 5 cities for top 30 attractions).
- You can then work on converting this into a generic app, and complete the app development after that.
- Note that you may have to use multiple ips to get the data (explore using free ip services, etc… or you may come up with you own logic for scraping the data in an effective manner).