Academic Year: 2023-24 **Semester: VIII**
Class / Branch: BE Computer **Subject: Social Media Analytics Lab**

Name: Shivam Pandey
Date of performance: 14/02/2024
Date of Submission: 24/02/2024

Experiment No. 02

Aim: To perform web crawling, scraping, and parsing using Instant Data Scraper, Octoparse, and Netlytics.

Objective: To apply web crawling, scraping, and parsing techniques to extract data from Google reviews using Instant Data Scraper, extract data from YouTube comments using Netlytic, and set up and run web scraping tasks to extract data using Octoparse

Software used: Instant Data Scraper (Chrome Extension), Octoparse, Netlytics.

Theory:

Web Crawling: Web crawling involves systematically browsing through web pages and following hyperlinks to discover and index content. It starts with a seed URL and then traverses through links to other pages, recursively fetching and indexing data. This process is essential for gathering a comprehensive dataset from the web.

Web Scraping: Web scraping refers to the extraction of specific data elements from web pages. It involves parsing HTML or XML documents to locate and extract desired information, such as text, images, or structured data. Web scraping tools like Instant Data Scraper automate this process by allowing users to specify the data to be extracted and the rules for extraction.

Parsing: Parsing is the process of analyzing a string of symbols or data to determine its structure and meaning. In the context of web scraping, parsing involves interpreting the HTML or XML markup of web pages to extract relevant data elements. This may include using techniques such as DOM parsing or regular expressions to locate and extract specific data patterns.

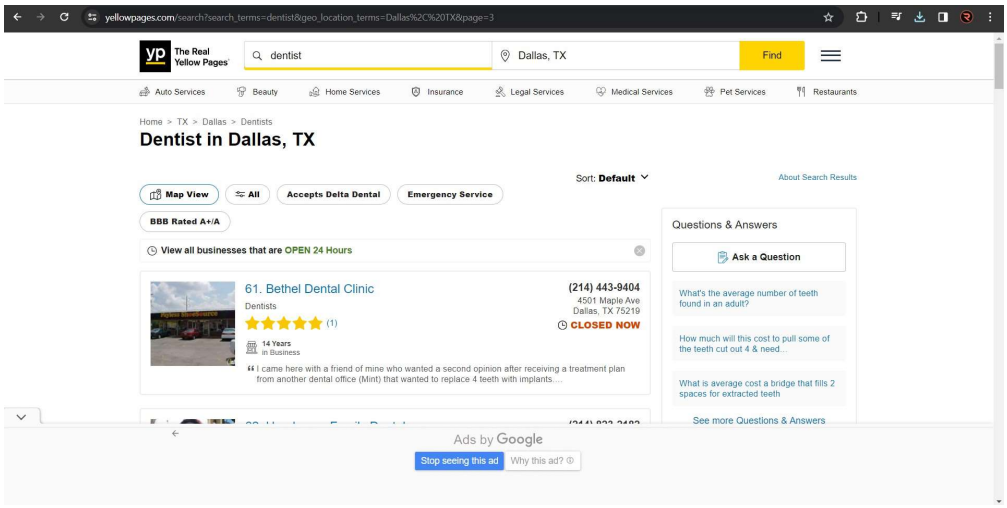
Instant Data Scraper: Instant Data Scraper is a Chrome extension that allows scraping data from websites directly in your browser. It provides a simple interface for selecting and extracting data elements, and it can export the data in various formats like CSV or Excel. Instant Data Scraper is useful for quick and easy web scraping tasks, but it may have limitations compared to more advanced scraping tools.

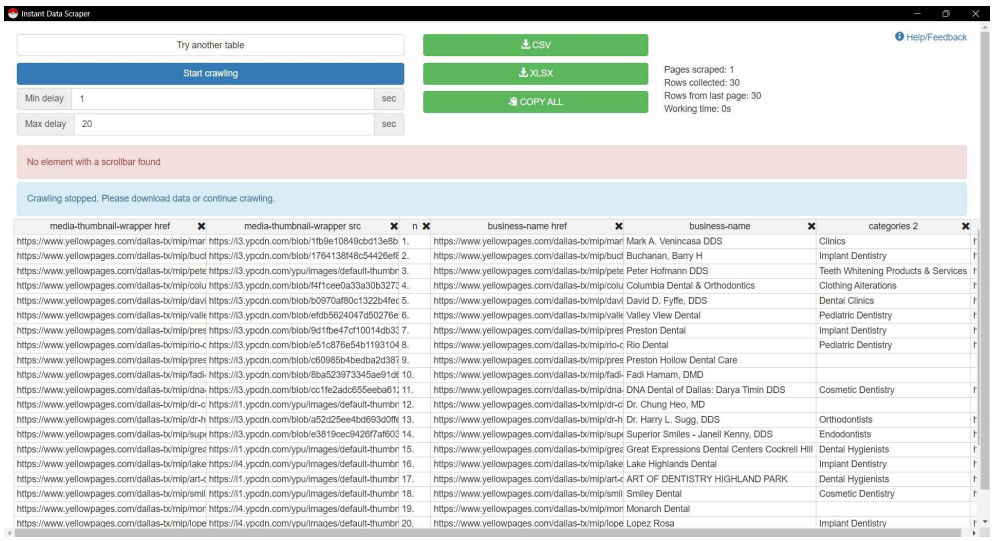
Netlytic: Netlytic is a cloud-based text and social network analyzer that allows users to collect, analyze, and visualize social media data. It can be used to study online communities, track social media trends, and analyze text data from various sources, including Twitter, Facebook, YouTube, and web forums. Netlytic offers features for data collection, text analysis, and network analysis, making it a versatile tool for social media research and analysis.

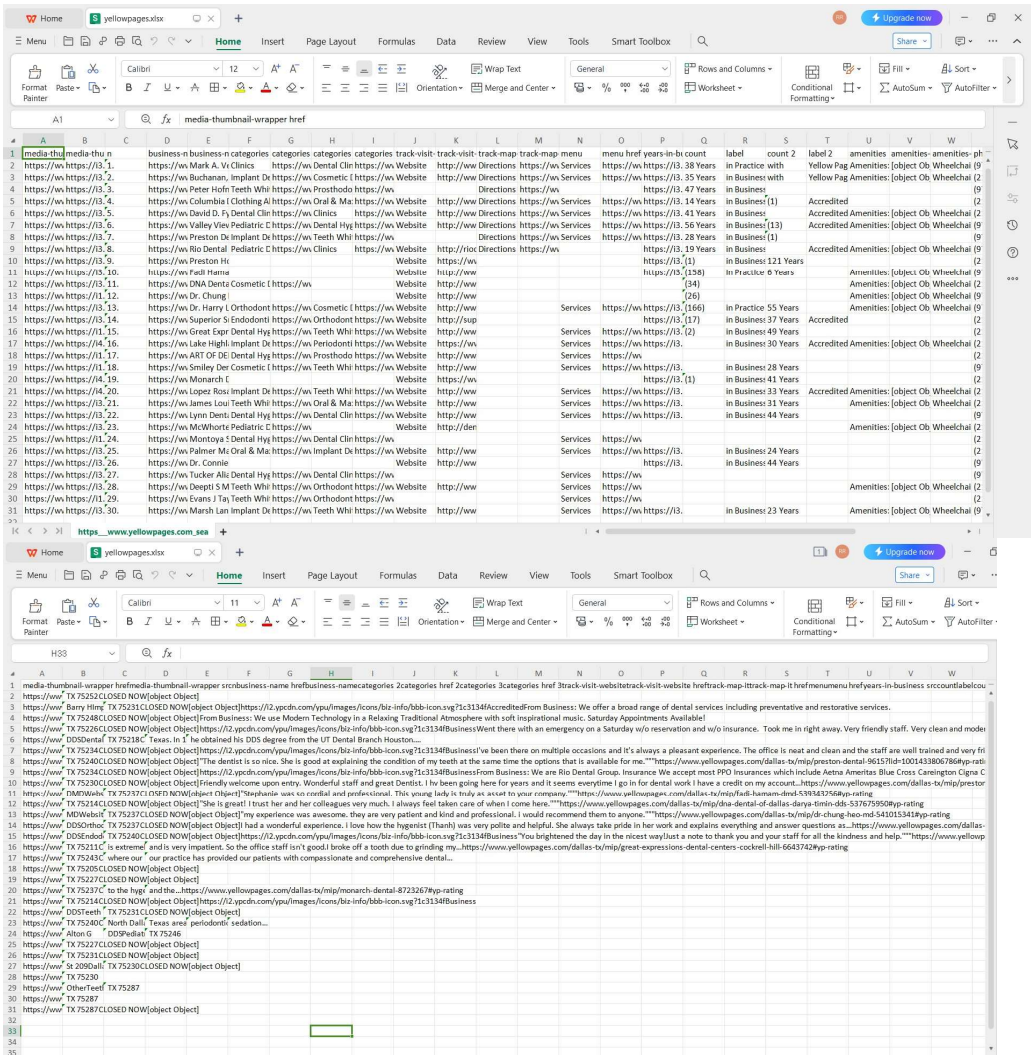
Octoparse: Octoparse is a web scraping tool that allows you to extract data from websites without the need for programming. It provides a visual interface for selecting the data to scrape and offers features like scheduled scraping, cloud extraction, and data export options. It's commonly used for tasks such as web data collection, price monitoring, and market research.

Implementation and Output:

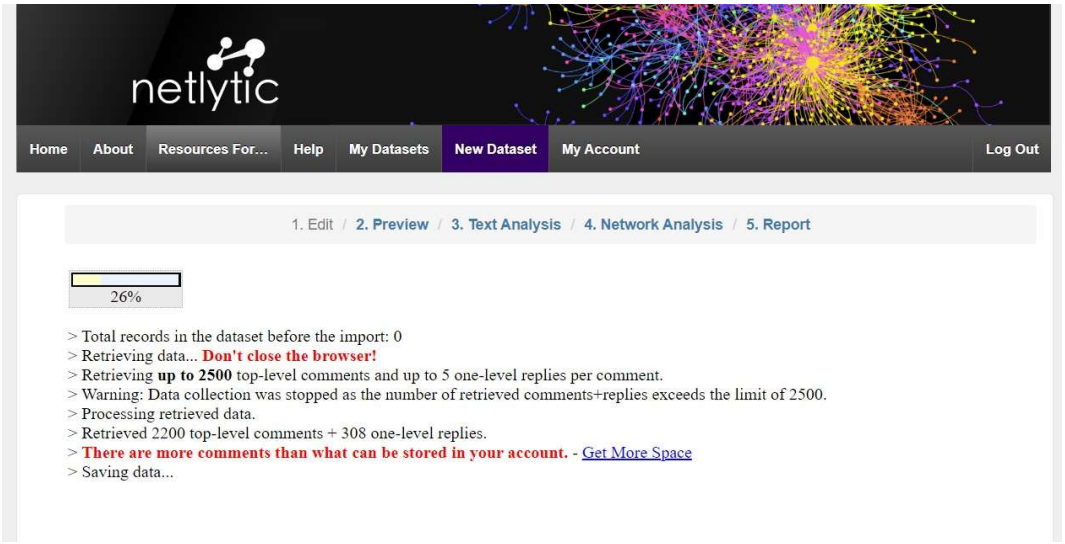
Instant Data Scrapper: Scraping the data from the real yellow web page using the IDS Chrome extension.

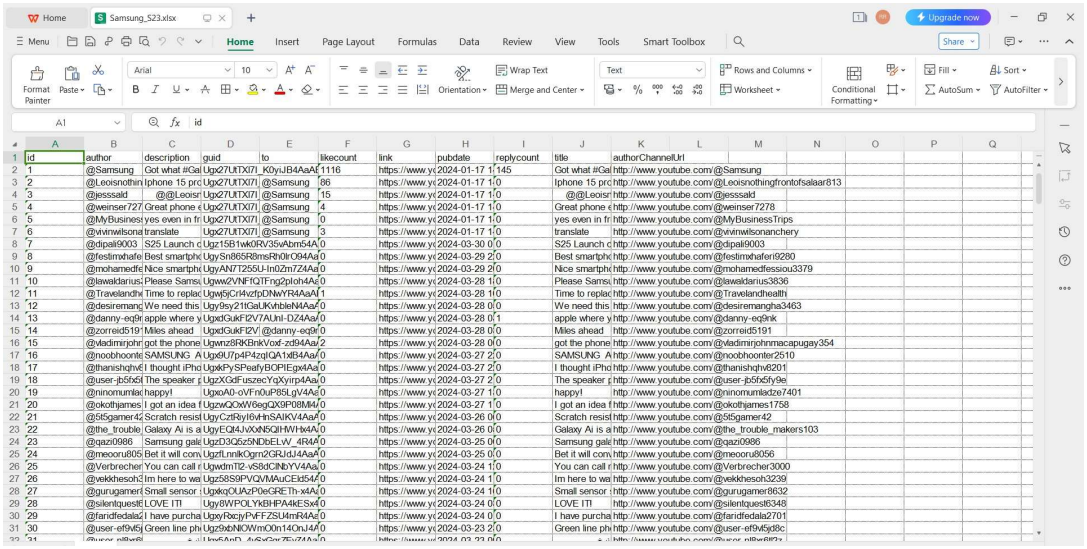




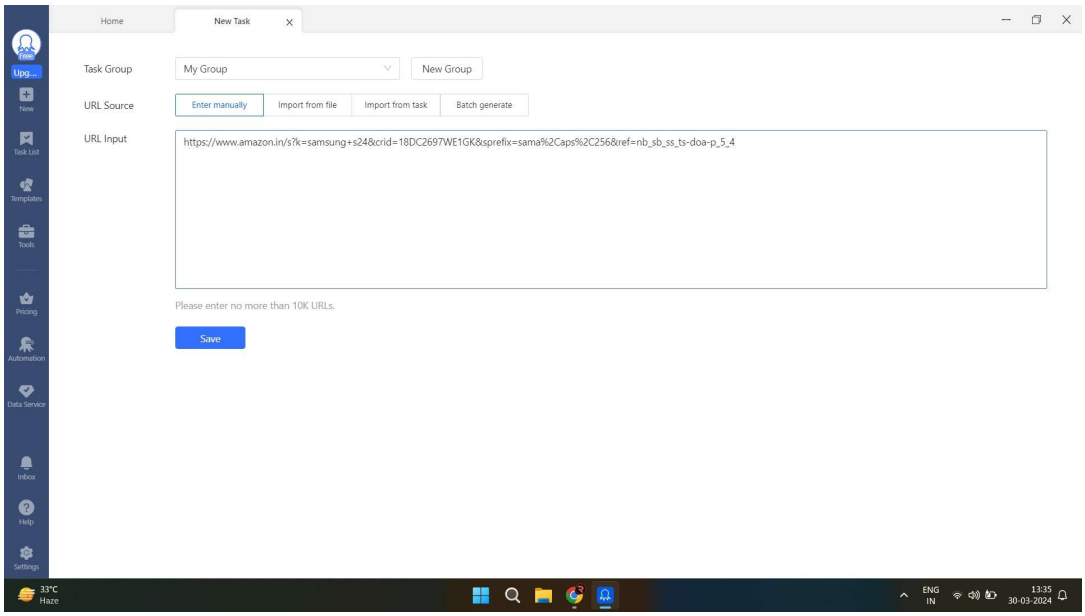


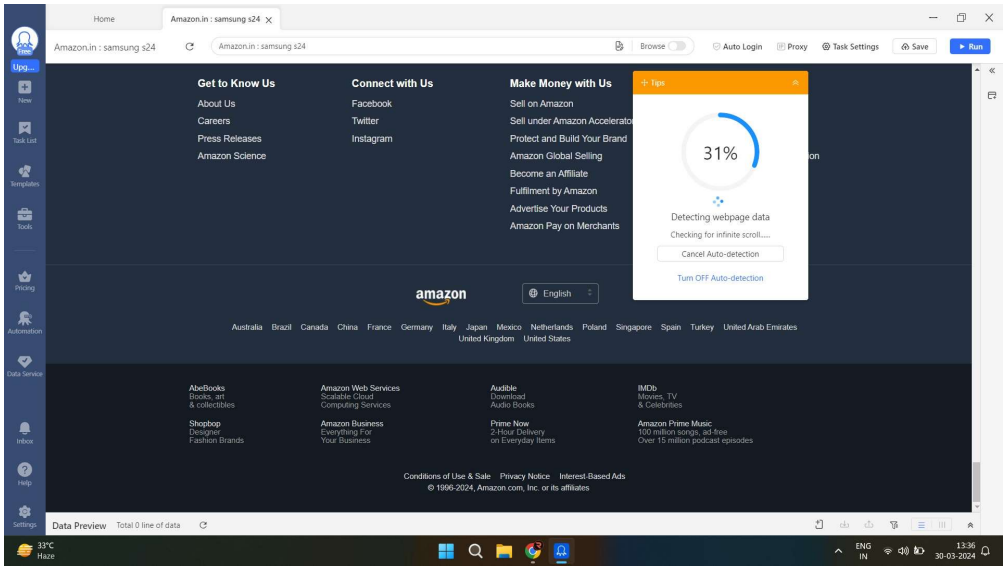
Netlytics: Using Netlytics to get the Data from a YouTube video of the Samsung S23 launch.

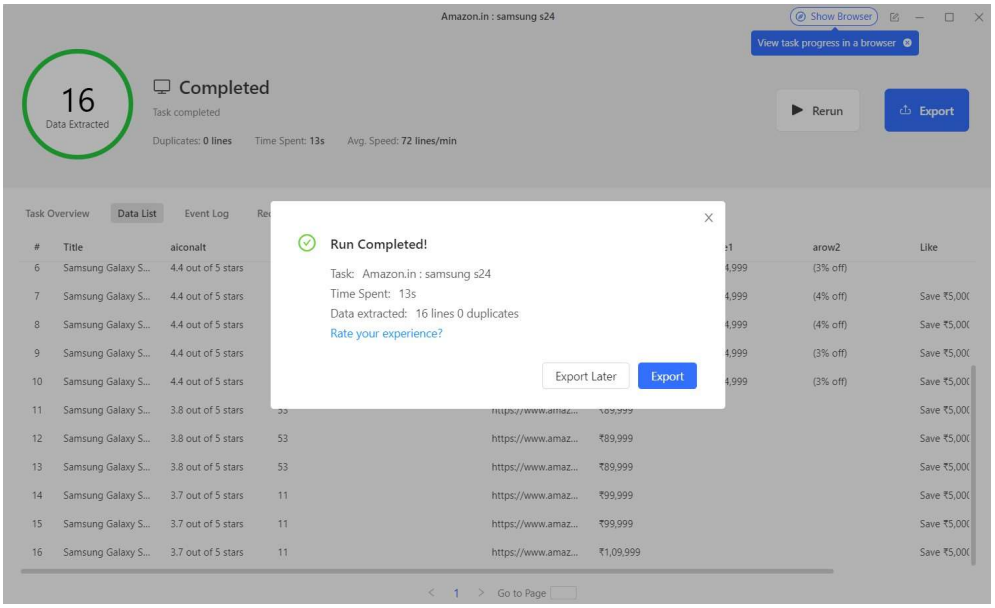


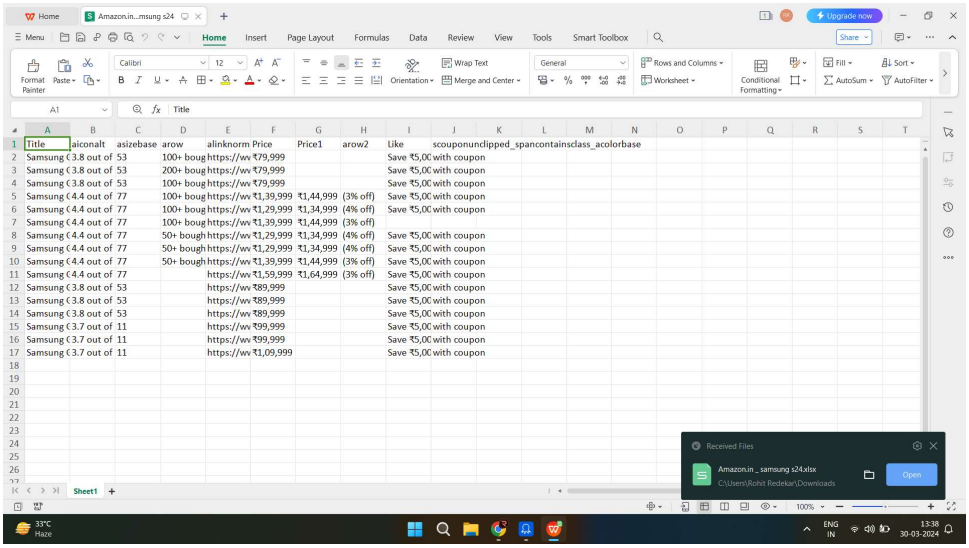


Octoparse: Using Octoparse to get the data from the product (Samsung S24) page of Amazon.




Amazon.in : samsung s24
Get to Know Us
About Us
Careers
Press Releases
Amazon Science
Connect with Us
Facebook
Twitter
Instagram
Make Money with Us
Sell on Amazon
Protect and Build Your Brand
Amazon Global Selling
Become an Affiliate
Fulfilment by Amazon
Advertise Your Products
Amazon Pay on Merchants
31%
Detecting webpage data
Checking for infinite scroll.....
Cancel Auto-detection
Turn OFF Auto-detection
amazon
English
Australia Brazil Canada China France Germany Italy Japan Mexico Netherlands Poland Singapore Spain Turkey United Arab Emirates
United Kingdom United States
AbeBooks
Amazon Web Services
Audible
IMDb
Shopbop
Amazon Business
Prime Now
Amazon Prime Music
Conditions of Use & Sale Privacy Notice Interest-Based Ads
© 1996-2024, Amazon.com, Inc. or its affiliates
Data Preview


Amazon.in : samsung s24
Show Browser
View task progress in a browser
16
Data Extracted
Completed
Task completed
Duplicates: 0 lines
Time Spent: 13s
Avg. Speed: 72 lines/min
Rerun
Export
Task Overview
Data List
Event Log
Run Completed!
Task: Amazon.in : samsung s24
Time Spent: 13s
Data extracted: 16 lines 0 duplicates
Rate your experience?
Export Later
Export
Samsung Galaxy S...
https://www.amaz...
Go to Page


Amazon.in...msung s24
Home
Insert
Page Layout
Formulas
Data
Review
View
Tools
Smart Toolbox
Title
aiconalt
asizebase
arow
alinknorm
Price
Price1
arow2
Like
scouponundlipped_spancontainsclass_acolorbase
Received Files
Amazon.in _ samsung s24.xlsx
Open
Sheet1

Conclusion:

In summary, this experiment highlighted the practical utility of web crawling, scraping, and parsing techniques facilitated by tools like Instant Data Scraper, Netlytic, and Octoparse. Each tool demonstrated unique strengths: Instant Data Scraper's simplicity for Google review extraction, Netlytic's effectiveness in analyzing social media comments from YouTube, and Octoparse's versatility in handling complex scraping tasks across multiple pages. Collectively, these tools offer a comprehensive suite of capabilities for web data extraction, catering to diverse user needs and skill levels. Their integration into research, analysis, and decision-making processes empowers users to derive valuable insights and make informed decisions based on web-based data sources.