Vidyavardhini's College of Engineering & Technology
Department of Computer Engineering
Academic Year : 2024-25

Experiment No. 2
To perform web crawling, scraping and parsing using Instant data scraper.
Date of Performance:18/01/2024
Date of Submission:19/04/2024

Aim: To perform web crawling, scraping and parsing using Instant data scraper, Netlytic and Octoparse.

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.

Theory:

Web crawling: Web crawling is the process of automatically browsing the internet and indexing web pages. It is typically done by search engines to discover new content and update their indexes. Web crawlers, also known as spiders or bots, follow links from one page to another and download the content of each page for indexing. While web crawling is not the same as web scraping, web scraping often involves web crawling to navigate through a website and extract data from multiple pages.

Web scraping: This is the process of extracting specific information from websites. It involves using software or programming scripts to access the HTML of web pages and extract the desired data, such as text, images, or links. Web scraping can be done manually or automatically, and it is used for various purposes, including data collection, market research, and price monitoring.

Parsing: Parsing is the process of analyzing the structure of a document or data file to extract meaningful information. In the context of web scraping, parsing is used to extract specific data elements from the HTML or other markup languages used to create web pages. This process involves identifying the patterns and structures of the data and using techniques like regular expressions or HTML parsers to extract the desired information.

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.

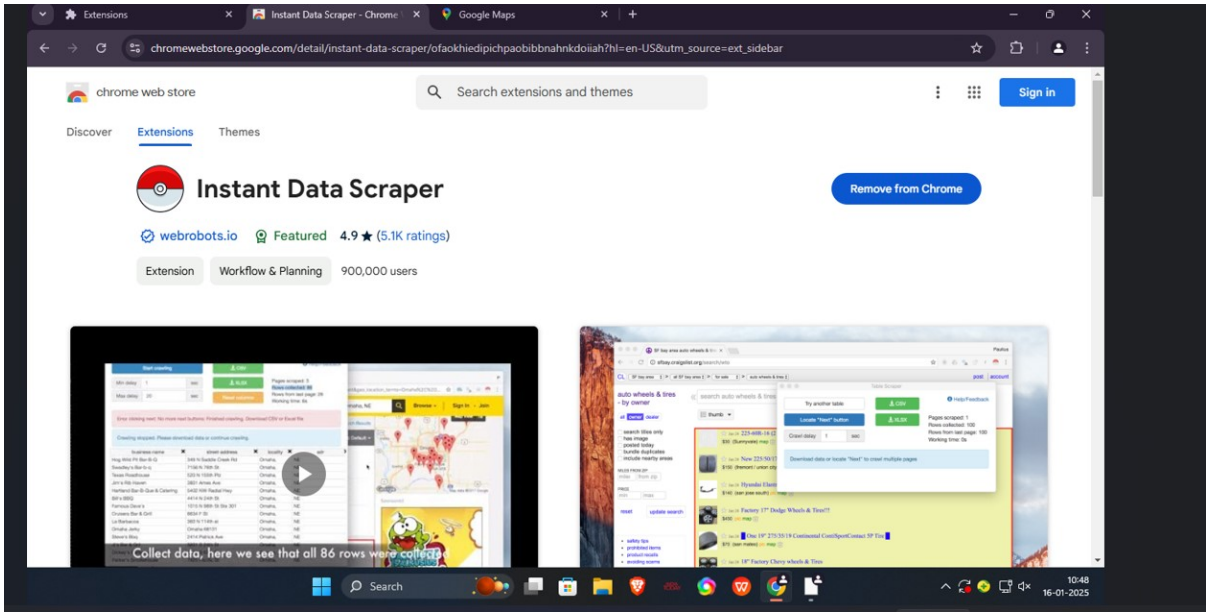
Implementation and Output:

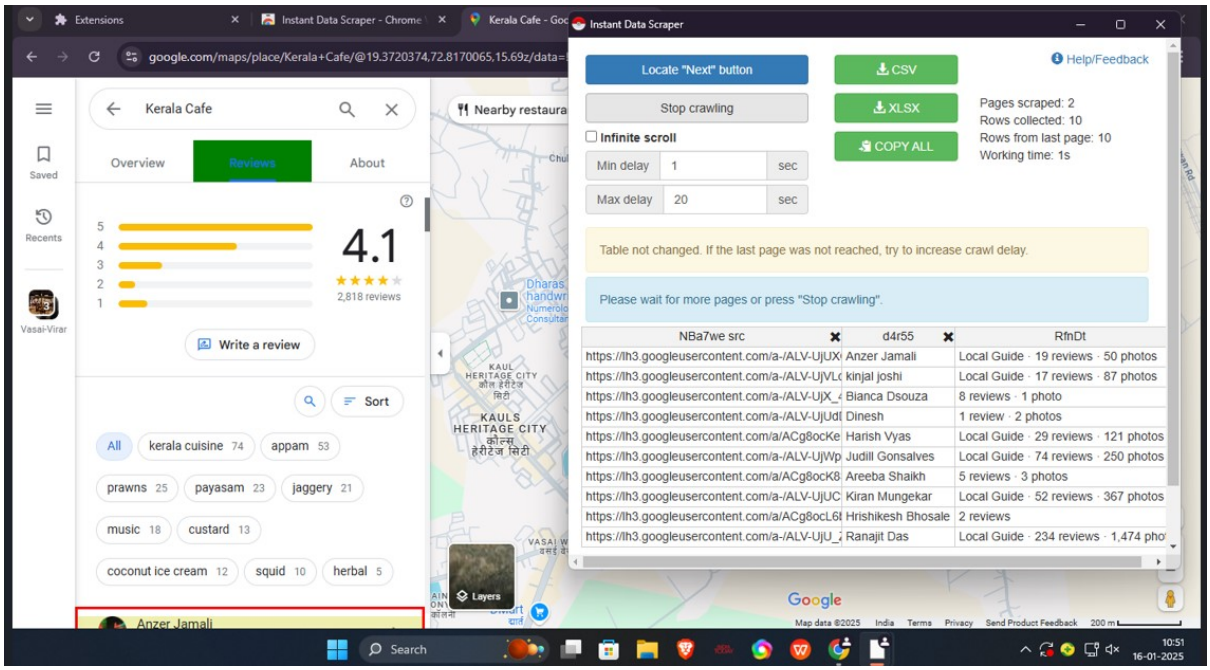
Scrape Google Reviews

Step 1 : Install the Google Chrome extension Instant Data Scraper to scrape Google reviews for any local business

Step 2 : Go to Google Maps and look for a business that interests you

Step 3 : Choose the reviews and launch Instant Data Scraper to crawl Google reviews. Wait until all reviews have been scraped



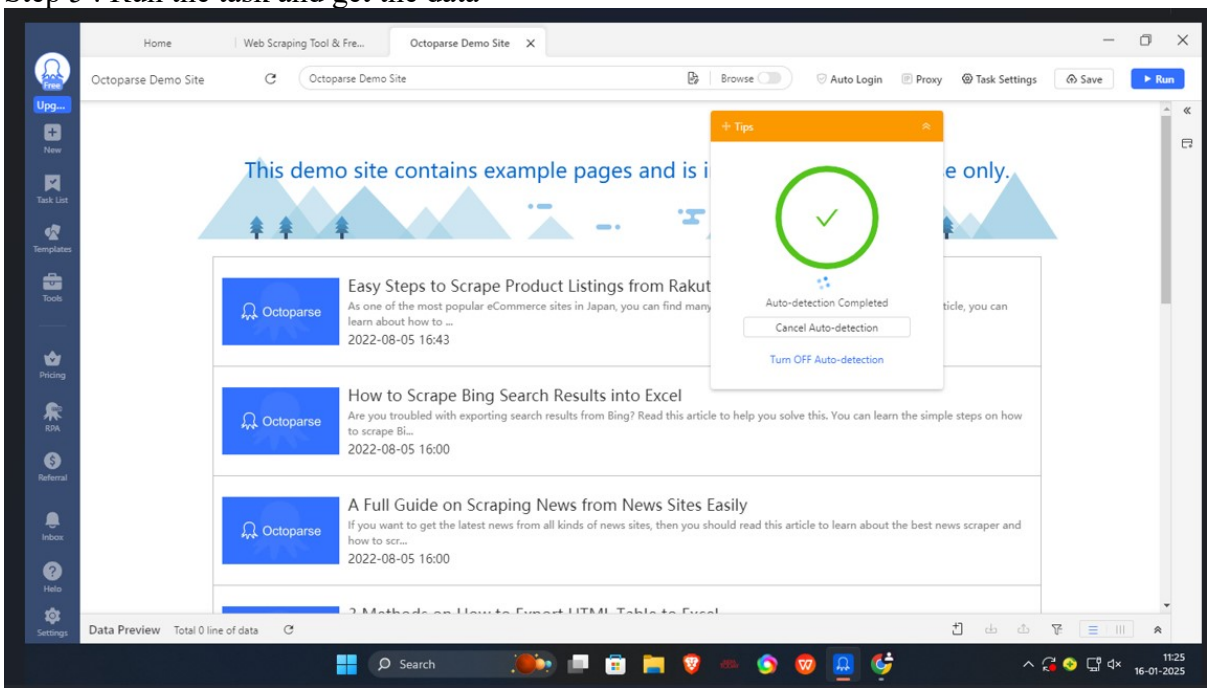


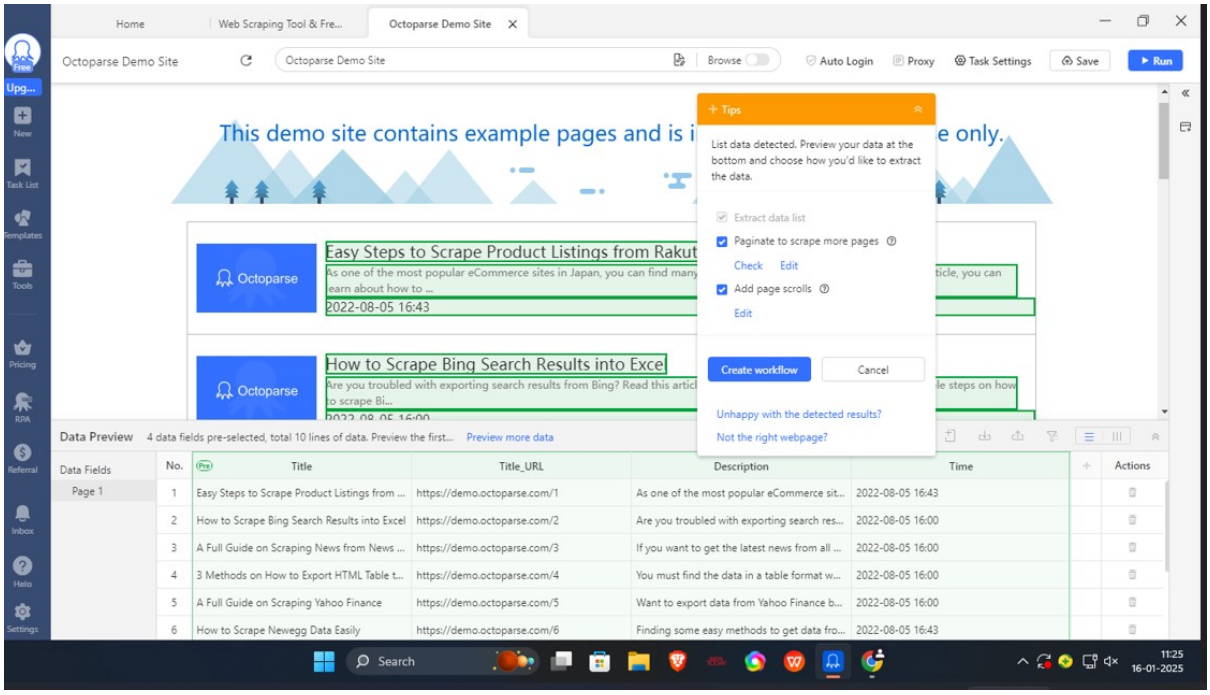
Scrape YouTube Comments using Netlytic

Step 1 : Sign up for Netlytic
Step 2 : Click "New Dataset"
Step 3 : Select "YouTube" as the data source
Step 4 : Copy the YouTube video ID you want to scrape comments from and paste it into Netlytic, also enter Dataset Name and click import
Step 5 : Go to "My Datasets" tab where you can find your dataset

Web Scraping using Octoparse

Step 1 : Go to web page
Step 2 : Create pagination
Step 3 : Build a loop item
Step 4 : Extract the data
Step 5 : Run the task and get the data





Conclusion:

In this experiment, we successfully performed web crawling, scraping, and parsing using Instant Data Scraper, Netlytic, and Octoparse. We extracted Google reviews, YouTube comments, and structured data from web pages using these tools. The process involved installing necessary tools, setting up data extraction tasks, and obtaining useful datasets. This experiment demonstrated the effectiveness of automated web data collection techniques. The extracted data can be further analyzed for insights in various domains like market research and sentiment analysis.