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: 06-01-2025
Date of Submission: 28-01-2025

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.

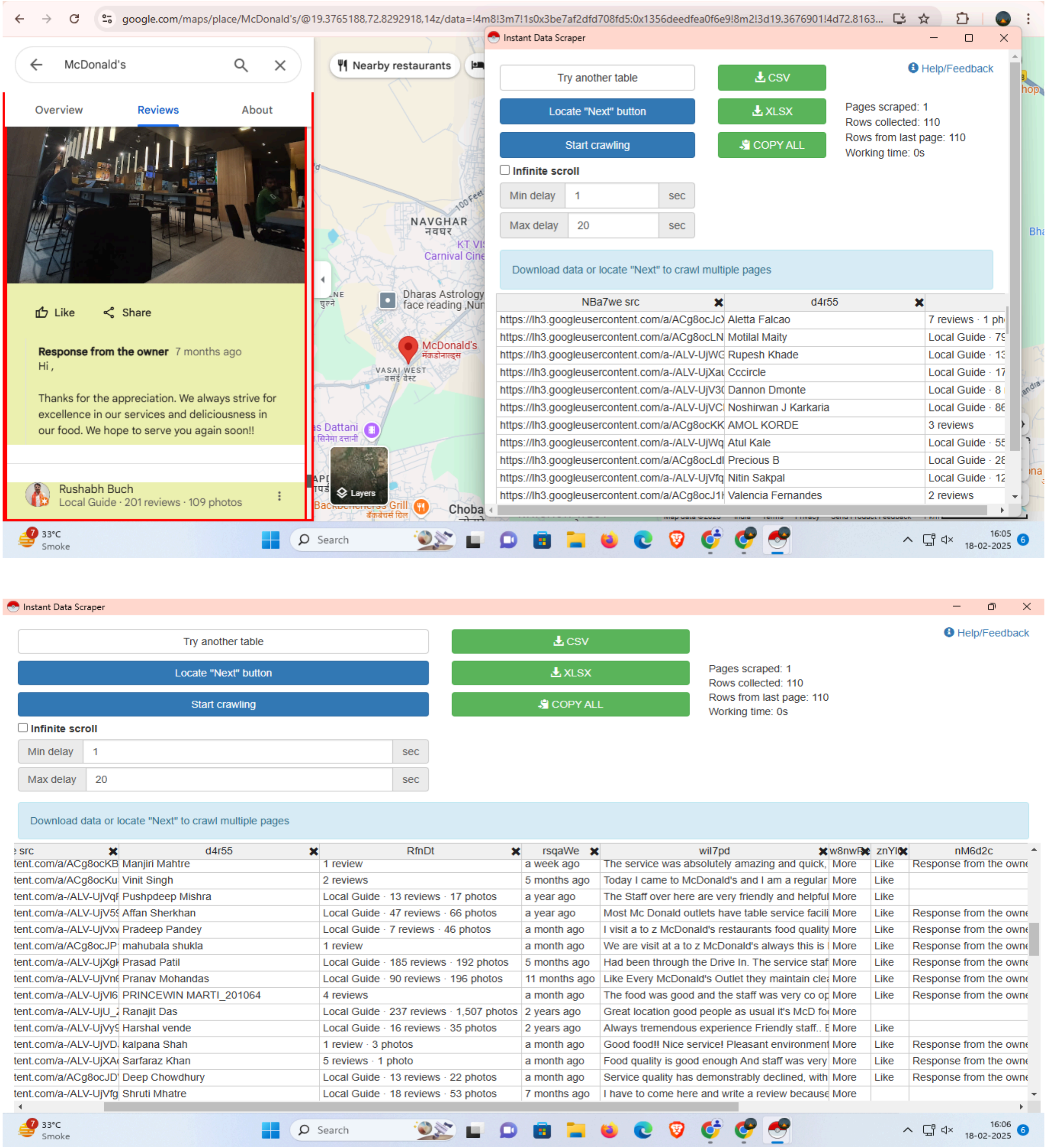
Implementation and Output:

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



google.com/maps/place/McDonald's/@19.3765188,72.8292918,14z/data=!4m8!3m7!1s0x3be7af2dfd708fd5:0x1356deedfea0f6e9!8m2!3d19.3676901!4d72.8163...

McDonald's

Overview Reviews About

Like Share

Response from the owner 7 months ago

Hi ,

Thanks for the appreciation. We always strive for excellence in our services and deliciousness in our food. We hope to serve you again soon!!

Rushabh Buch
Local Guide · 201 reviews · 109 photos

Instant Data Scraper

Try another table | CSV | XLSX | COPY ALL

Locate "Next" button

Start crawling

Help/Feedback

Pages scraped: 1
Rows collected: 110
Rows from last page: 110
Working time: 0s

□ Infinite scroll

Min delay 1 sec

Max delay 20 sec

Download data or locate "Next" to crawl multiple pages

NBa7we src	d4r55	
https://lh3.googleusercontent.com/a/ACg8ocJcX	Aletta Falcao	7 reviews · 1 ph
https://lh3.googleusercontent.com/a/ACg8ocLN	Motilal Maity	Local Guide · 79
https://lh3.googleusercontent.com/a-/ALV-UjWG	Rupesh Khade	Local Guide · 13
https://lh3.googleusercontent.com/a-/ALV-UjXau	Cccircle	Local Guide · 17
https://lh3.googleusercontent.com/a-/ALV-UjV3(	Dannon Dmonte	Local Guide · 8
https://lh3.googleusercontent.com/a-/ALV-UjVCI	Noshirwan J Karkaria	Local Guide · 86
https://lh3.googleusercontent.com/a/ACg8ocKK	AMOL KORDE	3 reviews
https://lh3.googleusercontent.com/a-/ALV-UjWq	Atul Kale	Local Guide · 55
https://lh3.googleusercontent.com/a/ACg8ocLdl	Precious B	Local Guide · 28
https://lh3.googleusercontent.com/a-/ALV-UjVfq	Nitin Sakpal	Local Guide · 12
https://lh3.googleusercontent.com/a/ACg8ocJ1I	Valencia Fernandes	2 reviews

16:05
18-02-2025

Instant Data Scraper

Try another table | CSV | XLSX | COPY ALL

Locate "Next" button

Start crawling

Help/Feedback

Pages scraped: 1
Rows collected: 110
Rows from last page: 110
Working time: 0s

□ Infinite scroll

Min delay 1 sec

Max delay 20 sec

Download data or locate "Next" to crawl multiple pages

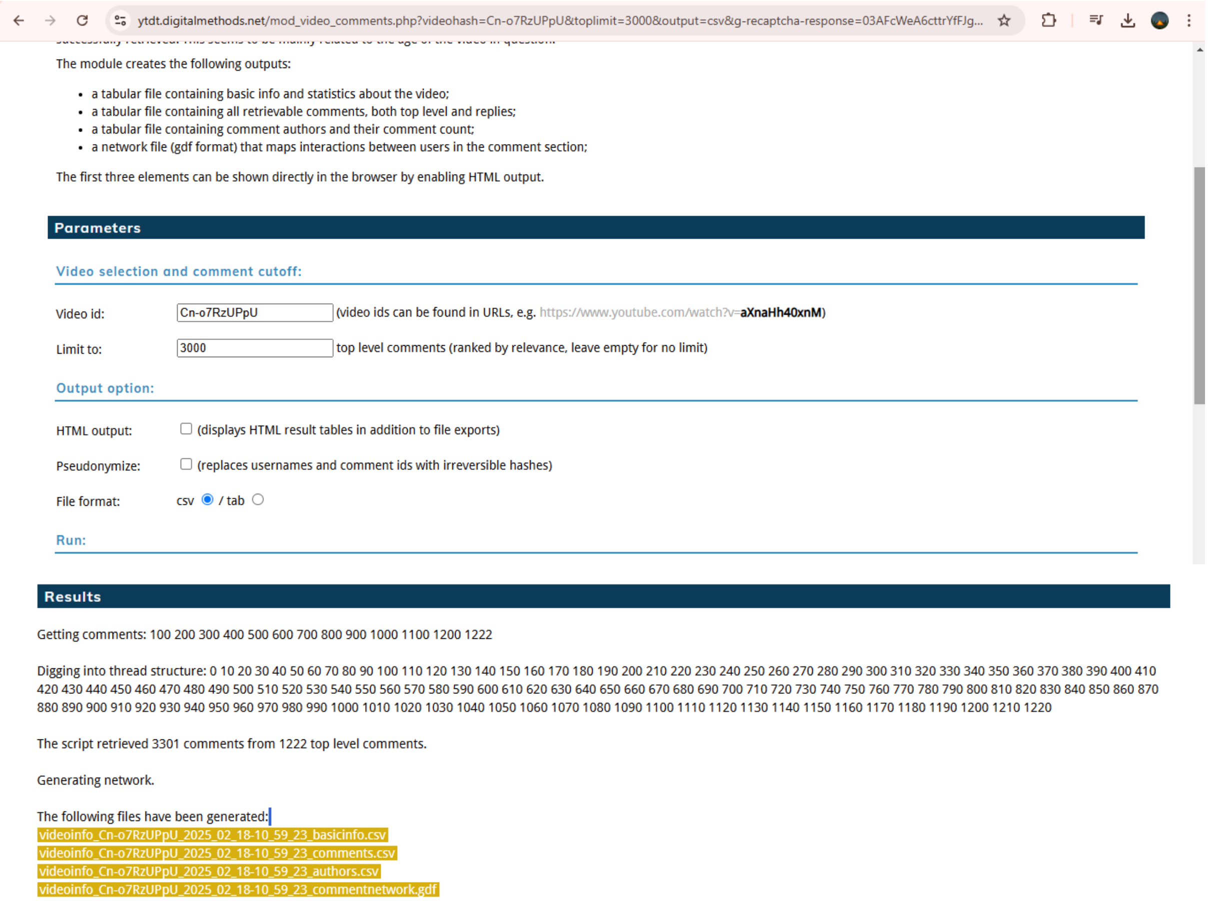
e src	d4r55	RfnDt	rsqaWe	wil7pd	w8nwRe	znYl	nM6d2c
tent.com/a/ACg8ocKB	Manjiri Mahtre	1 review	a week ago	The service was absolutely amazing and quick,	More	Like	Response from the owne
tent.com/a/ACg8ocKu	Vinit Singh	2 reviews	5 months ago	Today I came to McDonald's and I am a regular	More	Like	
tent.com/a-/ALV-UjVqF	Pushpdeep Mishra	Local Guide · 13 reviews · 17 photos	a year ago	The Staff over here are very friendly and helpful	More	Like	
tent.com/a-/ALV-UjV59	Affan Sherkhan	Local Guide · 47 reviews · 66 photos	a year ago	Most Mc Donald outlets have table service facili	More	Like	Response from the owne
tent.com/a-/ALV-UjVxw	Pradeep Pandey	Local Guide · 7 reviews · 46 photos	a month ago	I visit a to z McDonald's restaurants food quality	More	Like	Response from the owne
tent.com/a/ACg8ocJP	mahubala shukla	1 review	a month ago	We are visit at a to z McDonald's always this is	More	Like	Response from the owne
tent.com/a-/ALV-UjXgk	Prasad Patil	Local Guide · 185 reviews · 192 photos	5 months ago	Had been through the Drive In. The service staf	More	Like	Response from the owne
tent.com/a-/ALV-UjVn6	Pranav Mohandas	Local Guide · 90 reviews · 196 photos	11 months ago	Like Every McDonald's Outlet they maintain clea	More	Like	Response from the owne
tent.com/a-/ALV-UjVl6	PRINCEWIN MARTI_201064	4 reviews	a month ago	The food was good and the staff was very co op	More	Like	Response from the owne
tent.com/a-/ALV-UjU_z	Ranajit Das	Local Guide · 237 reviews · 1,507 photos	2 years ago	Great location good people as usual it's McD fo	More		
tent.com/a-/ALV-UjVy9	Harshal vende	Local Guide · 16 reviews · 35 photos	2 years ago	Always tremendous experience Friendly staff.. E	More	Like	
tent.com/a-/ALV-UjVD.	kalpana Shah	1 review · 3 photos	a month ago	Good food!! Nice service! Pleasant environment	More	Like	Response from the owne
tent.com/a-/ALV-UjXAc	Sarfaraz Khan	5 reviews · 1 photo	a month ago	Food quality is good enough And staff was very	More	Like	Response from the owne
tent.com/a/ACg8ocJD	Deep Chowdhury	Local Guide · 13 reviews · 22 photos	a month ago	Service quality has demonstrably declined, with	More	Like	Response from the owne
tent.com/a-/ALV-UjVfg	Shruti Mhatre	Local Guide · 18 reviews · 53 photos	7 months ago	I have to come here and write a review because	More		

16:06
18-02-2025

Fig: Scrapping of McDonald's(Vasai) Reviews

Scrape YouTube Comments using YouTube Data Tools (YTDT)

Step 1 : Go to tools.digitalmethods.net/netvizz/youtube/
Step 2 : Select "Video Info and Comments"
Step 3 : Enter the video ID
Step 4 : Set your desired parameters (number of comments, etc.)
Step 5 : Click "Run Query"
Step 6 : Download the results in CSV format



ytdt.digitalmethods.net/mod_video_comments.php?videohash=Cn-o7RzUPpU&toplimit=3000&output=csv&g-recaptcha-response=03AFcWeA6cttrYfFJg...

The module creates the following outputs:

- a tabular file containing basic info and statistics about the video;
- a tabular file containing all retrievable comments, both top level and replies;
- a tabular file containing comment authors and their comment count;
- a network file (gdf format) that maps interactions between users in the comment section;

The first three elements can be shown directly in the browser by enabling HTML output.

Parameters

Video selection and comment cutoff:

Video id: Cn-o7RzUPpU (video ids can be found in URLs, e.g. https://www.youtube.com/watch?v=**aXnaHh40xnM**)

Limit to: 3000 top level comments (ranked by relevance, leave empty for no limit)

Output option:

HTML output: □ (displays HTML result tables in addition to file exports)

Pseudonymize: □ (replaces usernames and comment ids with irreversible hashes)

File format: csv ● / tab ○

Run:

Results

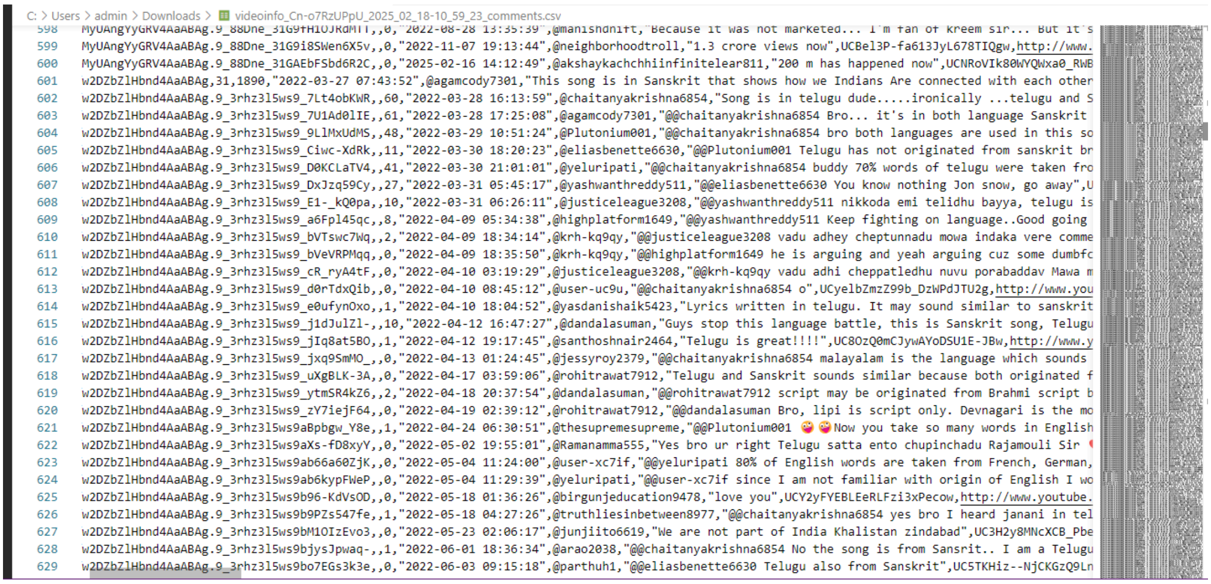
Getting comments: 100 200 300 400 500 600 700 800 900 1000 1100 1200 1222

Digging into thread structure: 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370 380 390 400 410 420 430 440 450 460 470 480 490 500 510 520 530 540 550 560 570 580 590 600 610 620 630 640 650 660 670 680 690 700 710 720 730 740 750 760 770 780 790 800 810 820 830 840 850 860 870 880 890 900 910 920 930 940 950 960 970 980 990 1000 1010 1020 1030 1040 1050 1060 1070 1080 1090 1100 1110 1120 1130 1140 1150 1160 1170 1180 1190 1200 1210 1220

The script retrieved 3301 comments from 1222 top level comments.

Generating network.

The following files have been generated:
videoinfo_Cn-o7RzUPpU_2025_02_18-10_59_23_basicinfo.csv
videoinfo_Cn-o7RzUPpU_2025_02_18-10_59_23_comments.csv
videoinfo_Cn-o7RzUPpU_2025_02_18-10_59_23_authors.csv
videoinfo_Cn-o7RzUPpU_2025_02_18-10_59_23_commentnetwork.gdf



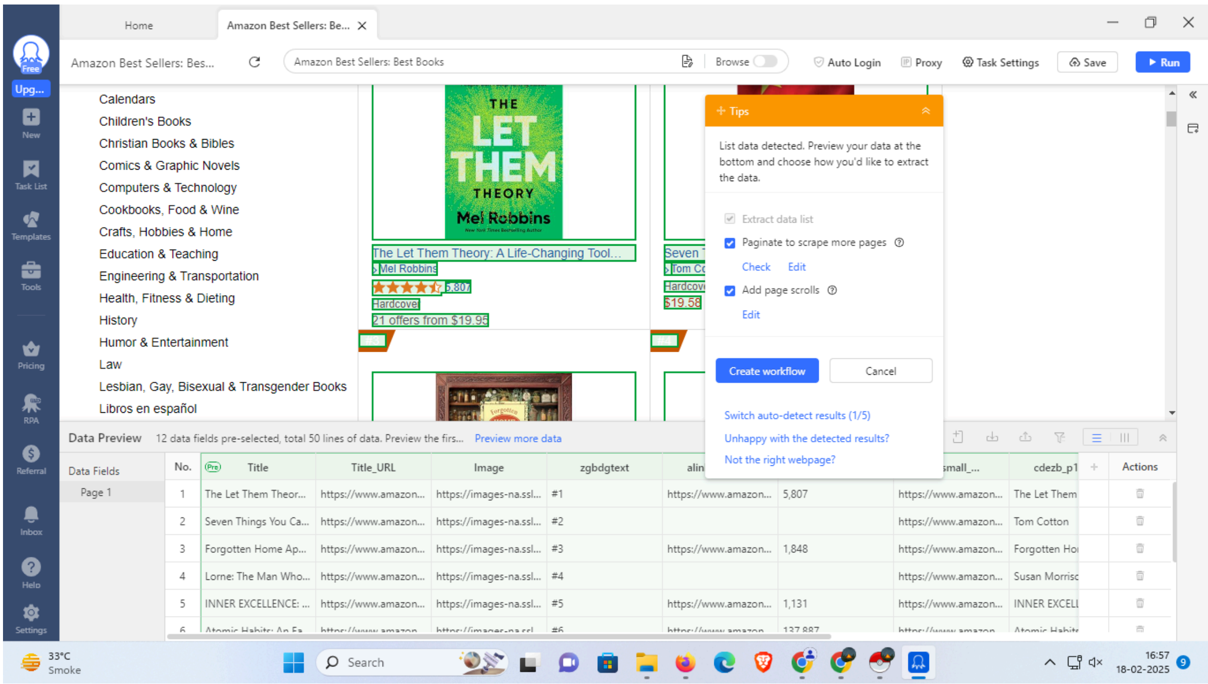
C: > Users > admin > Downloads > videoinfo_Cn-o7RzUPpU_2025_02_18-10_59_23_comments.csv

```
MyUAngYyGRV4AaABAg.9_88Dne_31G9fH1OJRdMTT,,0,"2022-08-28 13:35:39",@manishdnift,"Because it was not marketed... I'm fan of kreem sir... But it's
MyUAngYyGRV4AaABAg.9_88Dne_31G9i8SWen6X5v,,0,"2022-11-07 19:13:44",@neighborhoodtroll,"1.3 crore views now",UCBel3P-fa613JyL678TIQgw,http://www.
MyUAngYyGRV4AaABAg.9_88Dne_31GAEbFSbd6R2C,,0,"2025-02-16 14:12:49",@akshaykachchhiinfinitelear811,"200 m has happened now",UCNRoVIk80WYQWxa0_RWB
w2DZbZlHbnd4AaABAg,31,1890,"2022-03-27 07:43:52",@agamcody7301,"This song is in Sanskrit that shows how we Indians Are connected with each other
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_7Lt4obKWR,,60,"2022-03-28 16:13:59",@chaitanyakrishna6854,"Song is in telugu dude.....ironically ...telugu and S
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_7U1Ad0lIE,,61,"2022-03-28 17:25:08",@agamcody7301,"@@chaitanyakrishna6854 Bro... it's in both language Sanskrit
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_9LlMxUdMS,,48,"2022-03-29 10:51:24",@Plutonium001,"@@chaitanyakrishna6854 bro both languages are used in this so
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_Ciwc-XdRk,,11,"2022-03-30 18:20:23",@eliasbenette6630,"@@Plutonium001 Telugu has not originated from sanskrit br
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_D0KCLaTV4,,41,"2022-03-30 21:01:01",@yeluripati,"@@chaitanyakrishna6854 buddy 70% words of telugu were taken fro
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_DxJzq59Cy,,27,"2022-03-31 05:45:17",@yashwanthreddy511,"@@eliasbenette6630 You know nothing Jon snow, go away",U
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_E1-_kQ0pa,,10,"2022-03-31 06:26:11",@justiceleague3208,"@@yashwanthreddy511 nikkoda emi telidhu bayya, telugu is
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_a6Fpl45qc,,8,"2022-04-09 05:34:38",@highplatform1649,"@@yashwanthreddy511 Keep fighting on language..Good going
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_bVTswc7Wq,,2,"2022-04-09 18:34:14",@krh-kq9qy,"@@justiceleague3208 vadu adhey cheptunnadu mowa indaka vere comme
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_bVeVRPMqq,,0,"2022-04-09 18:35:50",@krh-kq9qy,"@@highplatform1649 he is arguing and yeah arguing cuz some dumbfc
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_cR_ryA4tF,,0,"2022-04-10 03:19:29",@justiceleague3208,"@@krh-kq9qy vadu adhi cheppatledhu nuvu porabaddav Mawa m
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_d0rTdxQib,,0,"2022-04-10 08:45:12",@user-uc9u,"@@chaitanyakrishna6854 o",UCyelbZmzZ99b_DzWPdJTU2g,http://www.you
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_e0ufynOxo,,1,"2022-04-10 18:04:52",@yasdanishaik5423,"Lyrics written in telugu. It may sound similar to sanskrit
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_j1dJulZl-,,10,"2022-04-12 16:47:27",@dandalasuman,"Guys stop this language battle, this is Sanskrit song, Telugu
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_jIq8at5BO,,1,"2022-04-12 19:17:45",@santhoshnair2464,"Telugu is great!!!!",UC8OzQ0mCJywAYoDSU1E-JBw,http://www.y
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_jxq9SmMO_,,0,"2022-04-13 01:24:45",@jessyroy2379,"@@chaitanyakrishna6854 malayalam is the language which sounds
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_uXgBLK-3A,,0,"2022-04-17 03:59:06",@rohitrawat7912,"Telugu and Sanskrit sounds similar because both originated f
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_ytmSR4kZ6,,2,"2022-04-18 20:37:54",@dandalasuman,"@@rohitrawat7912 script may be originated from Brahmi script b
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9_zY7iejF64,,0,"2022-04-19 02:39:12",@rohitrawat7912,"@@dandalasuman Bro, lipi is script only. Devnagari is the mo
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9aBpbgw_Y8e,,1,"2022-04-24 06:30:51",@thesupremesupreme,"@@Plutonium001 😜 😜Now you take so many words in English
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9aXs-fD8xyY,,0,"2022-05-02 19:55:01",@Ramanamma555,"Yes bro ur right Telugu satta ento chupinchadu Rajamouli Sir
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9ab66a60ZjK,,0,"2022-05-04 11:24:00",@user-xc7if,"@@yeluripati 80% of English words are taken from French, German,
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9ab6kypFWeP,,0,"2022-05-04 11:29:39",@yeluripati,"@@user-xc7if since I am not familiar with origin of English I wo
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9b96-KdVsOD,,0,"2022-05-18 01:36:26",@birgunjeducation9478,"love you",UCY2yFYEBLEeRLFzi3xPecow,http://www.youtube.
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9b9PZs547fe,,1,"2022-05-18 04:27:26",@truthliesinbetween8977,"@@chaitanyakrishna6854 yes bro I heard janani in tel
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9bM1OIzEvo3,,0,"2022-05-23 02:06:17",@junjiito6619,"We are not part of India Khalistan zindabad",UC3H2y8MNcXCB_Pbe
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9bjysJpwaq-,,1,"2022-06-01 18:36:34",@arao2038,"@@chaitanyakrishna6854 No the song is from Sansrit.. I am a Telugu
w2DZbZlHbnd4AaABAg.9_3rhz3l5ws9bo7EGs3k3e,,0,"2022-06-03 09:15:18",@parthuh1,"@@eliasbenette6630 Telugu also from Sanskrit",UC5TKHiz--NjCKGzQ9Ln
```

Fig : Youtube Comments Scrapping

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



List data detected. Preview your data at the bottom and choose how you'd like to extract the data.

☑ Extract data list

☑ Paginate to scrape more pages

Check Edit

☑ Add page scrolls

Edit

Create workflow Cancel

Switch auto-detect results (1/5)

Unhappy with the detected results?

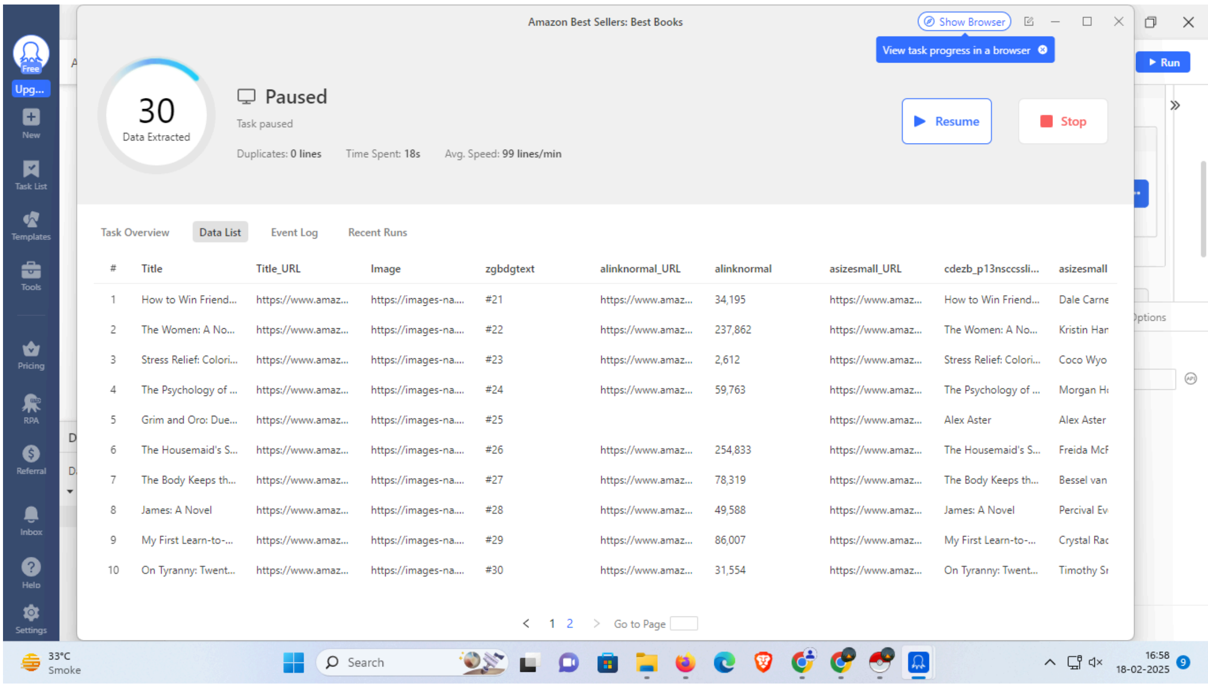
Not the right webpage?

Data Preview 12 data fields pre-selected, total 50 lines of data. Preview the firs... Preview more data

Data Fields

Page 1

No.	Title	Title_URL	Image	zgbdgtext	alin		small_...	cdezb_p1	Actions
1	The Let Them Theor...	https://www.amazon...	https://images-na.ssl...	#1	https://www.amazon...	5,807	https://www.amazon...	The Let Them	
2	Seven Things You Ca...	https://www.amazon...	https://images-na.ssl...	#2			https://www.amazon...	Tom Cotton	
3	Forgotten Home Ap...	https://www.amazon...	https://images-na.ssl...	#3	https://www.amazon...	1,848	https://www.amazon...	Forgotten Ho	
4	Lorne: The Man Who...	https://www.amazon...	https://images-na.ssl...	#4			https://www.amazon...	Susan Morris	
5	INNER EXCELLENCE: ...	https://www.amazon...	https://images-na.ssl...	#5	https://www.amazon...	1,131	https://www.amazon...	INNER EXCELL	
6	Atomic Habits: An Ea	https://www.amazon	https://images-na.ssl	#6	https://www.amazon	137,887	https://www.amazon	Atomic Habits	



30
Data Extracted

Paused

Task paused

Duplicates: 0 lines Time Spent: 18s Avg. Speed: 99 lines/min

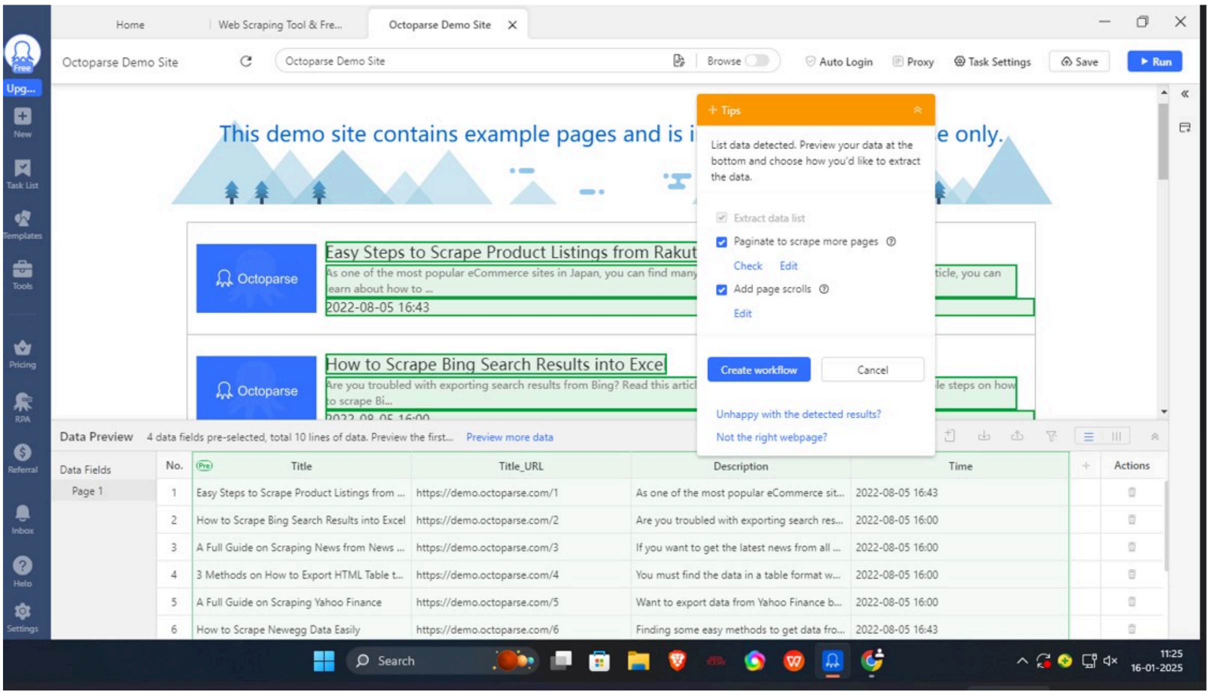
Resume Stop

Task Overview Data List Event Log Recent Runs

#	Title	Title_URL	Image	zgbdgtext	alinknormal_URL	alinknormal	asizesmall_URL	cdezb_p13nsccssli...	asizesmall
1	How to Win Friend...	https://www.amaz...	https://images-na....	#21	https://www.amaz...	34,195	https://www.amaz...	How to Win Friend...	Dale Carne
2	The Women: A No...	https://www.amaz...	https://images-na....	#22	https://www.amaz...	237,862	https://www.amaz...	The Women: A No...	Kristin Har
3	Stress Relief: Colori...	https://www.amaz...	https://images-na....	#23	https://www.amaz...	2,612	https://www.amaz...	Stress Relief: Colori...	Coco Wyo
4	The Psychology of ...	https://www.amaz...	https://images-na....	#24	https://www.amaz...	59,763	https://www.amaz...	The Psychology of ...	Morgan H
5	Grim and Oro: Due...	https://www.amaz...	https://images-na....	#25			https://www.amaz...	Alex Aster	Alex Aster
6	The Housemaid's S...	https://www.amaz...	https://images-na....	#26	https://www.amaz...	254,833	https://www.amaz...	The Housemaid's S...	Freida McF
7	The Body Keeps th...	https://www.amaz...	https://images-na....	#27	https://www.amaz...	78,319	https://www.amaz...	The Body Keeps th...	Bessel van
8	James: A Novel	https://www.amaz...	https://images-na....	#28	https://www.amaz...	49,588	https://www.amaz...	James: A Novel	Percival Ev
9	My First Learn-to-...	https://www.amaz...	https://images-na....	#29	https://www.amaz...	86,007	https://www.amaz...	My First Learn-to-...	Crystal Rac
10	On Tyranny: Twent...	https://www.amaz...	https://images-na....	#30	https://www.amaz...	31,554	https://www.amaz...	On Tyranny: Twent...	Timothy Si

< 1 2 > Go to Page

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



This demo site contains example pages and is i... e only.

Easy Steps to Scrape Product Listings from Rakut
As one of the most popular eCommerce sites in Japan, you can find many ... ticle, you can learn about how to ...
2022-08-05 16:43

How to Scrape Bing Search Results into Exce
Are you troubled with exporting search results from Bing? Read this articl ... le steps on how to scrape Bi...

Tips

List data detected. Preview your data at the bottom and choose how you'd like to extract the data.

☑ Extract data list
☑ Paginate to scrape more pages
Check Edit
☑ Add page scrolls
Edit

Create workflow Cancel

Unhappy with the detected results?
Not the right webpage?

Data Preview 4 data fields pre-selected, total 10 lines of data. Preview the first... Preview more data

Data Fields
Page 1

No.	Title	Title_URL	Description	Time
1	Easy Steps to Scrape Product Listings from ...	https://demo.octoparse.com/1	As one of the most popular eCommerce sit...	2022-08-05 16:43
2	How to Scrape Bing Search Results into Excel	https://demo.octoparse.com/2	Are you troubled with exporting search res...	2022-08-05 16:00
3	A Full Guide on Scraping News from News ...	https://demo.octoparse.com/3	If you want to get the latest news from all ...	2022-08-05 16:00
4	3 Methods on How to Export HTML Table t...	https://demo.octoparse.com/4	You must find the data in a table format w...	2022-08-05 16:00
5	A Full Guide on Scraping Yahoo Finance	https://demo.octoparse.com/5	Want to export data from Yahoo Finance b...	2022-08-05 16:00
6	How to Scrape Newegg Data Easily	https://demo.octoparse.com/6	Finding some easy methods to get data fro...	2022-08-05 16:43

Conclusion:

Web crawling, scraping, and parsing tools like Instant Data Scraper and Octoparse provide powerful means to extract and analyze online data systematically. Through this experiment, we learned how to collect Google reviews, YouTube comments, and structured web data using different tools and techniques. These tools demonstrate the practical application of web scraping for data collection and analysis, enabling researchers and businesses to gather valuable insights from online sources. The combination of these tools showcases how different scraping approaches can be used depending on the source and type of data needed, making data collection more efficient and accessible even without programming knowledge.