



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

Name: Divya Patil
Roll No.: 08
Experiment No. 2
To perform web crawling, scraping and parsing using Instant data scraper, Netlytic and Octoparse
Date of Performance: 08/02/2024
Date of Submission: 15/02/2024



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

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



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

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



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

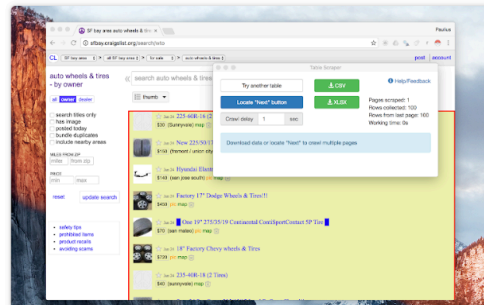
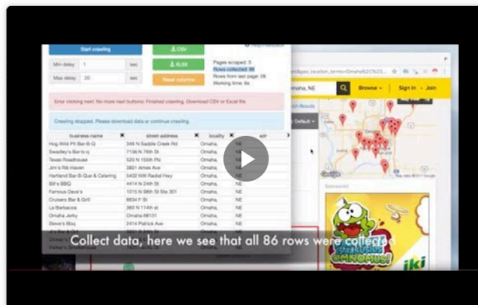


Instant Data Scraper

webrobots.io Featured 4.8 ★ (3.8K ratings)

Extension Workflow & Planning 600,000 users

Add to Chrome



google.com/maps/place/McDonald's/@19.3919376,72.8397201,15z/data=!4m8!3m7!1s0x3be7a91652b3a48d0x36133ed90b6ab68d!8m2!3d19.4039412!4d72.8376004!9m1!1b...

McDonald's

Overview Reviews About

Chanchal Singh 1 review

★★★★★ a year ago

Hii good afternoon this place is osm and me and my friend very enjoy the moment thanks Mcdonald

Serv... More

Like Share

Krishna Singh 1 review

★★★★★ 3 months ago

Food quality is good

Food: 5

Service... More

Like Share

Vaibhav chougule Local Guide · 34 reviews · 16 photos

Nearby restaurants Hotels Things to do Bars Coffee Takeout Groceries

McDonald's

Vasant Nagari Ground

Vishnu enterprises Perfect Tyres for Innova

Bhoidapada

Target Hydrautech

Tugar phata

OM NAGAR

BAXAY PRIME Multi-specialty ICU & Baxay Prime Hospital

Wash N Dry Laundry-Mart

Asus Exclusive Store - Kings Group

Manikpur

Instant Data Scraper

Stop crawling

Min delay 1 sec

Max delay 20 sec

CSV

XLSX

COPY ALL

Pages scraped: 43

Rows collected: 445

Rows from last page: 445

Working time: 222s

Please wait for more pages or press "Stop crawling".

NBA7we src	d4i55	RthDt
https://lh3.googleusercontent.com/a-/ALV-UJ7U7i Sid Kel		Local Guide · 160 reviews · 24
https://lh3.googleusercontent.com/a-/ACg8ocIGf Manasvi Mathur		Local Guide · 9 reviews · 15 pr
https://lh3.googleusercontent.com/a-/ALV-UJXR; Swarupa Morje		Local Guide · 195 reviews · 1,6
https://lh3.googleusercontent.com/a-/ALV-UJWn Ranajit Das		Local Guide · 192 reviews · 1,6
https://lh3.googleusercontent.com/a-/ALV-UJW9 Shrutika Manrai		4 reviews · 3 photos
https://lh3.googleusercontent.com/a-/ALV-UJv6t Yogesh Merugu		1 review
https://lh3.googleusercontent.com/a-/ACg8ocLVj Winona Dmello		6 reviews · 3 photos
https://lh3.googleusercontent.com/a-/ALV-UJUL; SP Fernandes		Local Guide · 35 reviews · 1.5
https://lh3.googleusercontent.com/a-/ACg8ocIU SHWETA GOWDA		Local Guide · 31 reviews · 1 pr
https://lh3.googleusercontent.com/a-/ACg8ocKC HETAL SAYAM		Local Guide · 10 reviews · 53 p



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
26	https://h3.google	namia Poojan	1 review			2 months ago	Always a go-to place					Share			
29	https://h3.google	Sharukh Bambo	Local Guide · 50			a year ago	McDonald's is a	More		Like		Share			
30	https://h3.google	Hrudh Pandya	1 review			3 weeks ago	It was amazing &	More		Like		Share	Response from 13 weeks ago		Hi ,
31	https://h3.google	tanish chhadwa	6 reviews			2 months ago	All the staff is ve	More		Like		Share			Thanks for the appre
32	https://h3.google	Rupesh	Local Guide · 21			3 weeks ago	This place is goc	More		Like		Share	Response from 13 weeks ago		Hi ,
33	https://h3.google	Raju Pillai	Local Guide · 14			11 months ago	The worst experi	More				Share	Response from 11 months ago		Thanks for the appre
34	https://h3.google	Somnath Adhya	2 reviews			2 months ago	Food quality is t	More		Like		Share			Hi Raju, ...
35	https://h3.google	Shreya Sharma	11 reviews · 2 p			a month ago	Regularly visit th	More		Like		Share	Response from 1 month ago		Hi ,
36	https://h3.google	Neha Jha	1 review			2 months ago	I loved the food .	More		Like		Share			Thanks for the appre
37	https://h3.google	Miss . Anonymou	2 reviews			2 months ago	We enjoyed our	More		Like		Share			
38	https://h3.google	Binay Bhardwaj	1 review			2 months ago	Nice one Service	More		Like		Share			
39	https://h3.google	Kalpesh Mistry (2 reviews			2 months ago	The food is awer	More		Like		Share			
40	https://h3.google	Natasha D'souza	1 review			2 months ago	I am loving it trul	More		Like		Share			
41	https://h3.google	Vinit Rathod	Local Guide · 76			2 days ago	Had a great time	More		Like		Share	Response from 12 days ago		Hi ,
42	https://h3.google	Zenia D'Abreo	Local Guide · 56			2 years ago	Overall Experien	More				Share			Thanks for the appre
43	https://h3.google	Sushma Karnati	Local Guide · 13			4 months ago	Food quality is great no doubt but					Share			Hi Sushma,
							Very irresponsibl	More		Like		Share	Response from 13 months ago		We're sorry for the di
															Hi

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



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

HomeAboutResources For...HelpMy DatasetsNew DatasetMy AccountLog Out

Twitter(discontinued)YouTubeGoogle SheetsText FileRSSReddit/Telegram (via Communalytic.org)

YouTube API information and limitations

The YouTube API limits the number of comments Netlytic can collect daily. If you encounter issues, you can use third-party tools like [YouTube Data Tools](#) to collect comments from YouTube as a CSV file and then import it into Netlytic for further analysis.

OnePlus
Wt8e-vBnDvQ(No Special Characters)

Enter the ID of the YouTube video as follows:
<https://www.youtube.com/watch?v=9bZkp7q19f0>

On the video page,
copy the code after "v=" in the URL

Important notices:

- Please don't close the browser once you click the "Import" button below.
- Netlytic collects top-level comments + up to 5 replies per comment. Replies to replies are not collected.
- Since YouTube API only permits storage of public data for up to 30 days, this dataset will be automatically deleted 30 days after its collection unless it's updated within the 30-day period.

You are using 1/3 of your permitted datasets - [Get More](#)

ImportGo Back (No Action)

netlytic

HomeAboutResources For...HelpMy DatasetsNew DatasetMy AccountLog Out

100%

> Retrieving data... **Don't close the browser!**
> Retrieving up to 2500 top-level comments and up to 5 one-level replies per comment.
> Processing retrieved data.
> Retrieved 355 top-level comments + 82 one-level replies.
> Saving data...
> Saved updated 437 records.
You can now close the browser!

<Go BackNext Step>



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

id	author	description	guild	to	likecount	link	pubdate	replycount	title	authorChannelUrl
1	@neelmd5467	This is not the Si	UgzJfIHUck4MJo1n4AaABAg		0	https://www.yout	2024-03-19 16:5	0	This is not the Si	http://www.youtube.com/@neelmd5467
2	@Abhishek123	सं भी विहार से हूँ।	Ugy0X1_j-CB04rCo3F4AaABAg		1	https://www.yout	2024-03-19 12:1	0	सं भी विहार से हूँ।	http://www.youtube.com/@Abhishek123
3	@parmarRamesh	Nice	Ugz6E89ZMJvdSG1nt54AaABAg		0	https://www.yout	2024-03-19 8:21	0	Nice	http://www.youtube.com/@parmarRamesh-m6je
4	@Bhargav_Gad	"Why TF are YO	UgwdJ4wQo8qA70_KITV4AaABAg		0	https://www.yout	2024-03-19 6:53	0	"Why TF are YO	http://www.youtube.com/@Bhargav_Gadekar
5	@nandeesh788	kat gaya bhai, at	UgzfVPrBqsJWxbxp494AaABAg		0	https://www.yout	2024-03-19 5:55	0	kat gaya bhai, at	http://www.youtube.com/@nandeesh788
6	@neymarjunior4	love from durga	UgzKBzVV25BujWuL0I4AaABAg		0	https://www.yout	2024-03-19 3:30	0	love from durga	http://www.youtube.com/@neymarjunior4512
7	@TheTeaCompi	When did Ahem	UgwtetJnMaVqZfobwHJ4AaABAg		0	https://www.yout	2024-03-18 2:43	0	When did Ahem	http://www.youtube.com/@TheTeaCompany
8	@akhlantal87	Namita and Ama	UgzKH0cNcepczLkEmgB4AaABAg		0	https://www.yout	2024-03-17 23:3	0	Namita and Ama	http://www.youtube.com/@akhlantal87
9	@mickeysam97	Ismain anupam	I UgzVWReSAFCypcWP64AaABAg		0	https://www.yout	2024-03-17 10:0	0	Ismain anupam	http://www.youtube.com/@mickeysam9752
10	@InnovativeChai	Namita just neec	UgwQhLDJQUadv9Cj14AaABAg		0	https://www.yout	2024-03-16 22:1	0	Namita just neec	http://www.youtube.com/@InnovativeChangazi
11	@EyeOTuber	Vahi yeh so calle	UgzWfZg-jPjWLU9HB4AaABAg		0	https://www.yout	2024-03-16 22:0	0	Vahi yeh so calle	http://www.youtube.com/@EyeOTuber
12	@vaishnavikumi	Mujihe yesa lgt	UgzG1gtcMR_jIP5hezp4AaABAg		3	https://www.yout	2024-03-16 13:0	1	Mujihe yesa lgt	http://www.youtube.com/@vaishnavikumad5343
13	@merjob2236		UgzG1gtcMR_jIP5hezp4AaABAg		1	https://www.yout	2024-03-16 11:1	0		http://www.youtube.com/@merjob2236
14	@pranshubudhr	Sony please red	Ugy1TPuB6IMdQ2UnV4AaABAg		0	https://www.yout	2024-03-16 10:5	0	Sony please red	http://www.youtube.com/@pranshubudhran8959
15	@AnuragKumar		UgwFXLDQdNHuBcjinH4AaABAg		0	https://www.yout	2024-03-16 10:3	0		http://www.youtube.com/@AnuragKumarPal-sn5cd
16	@CocoBebo	Why sharks don	UgyuqTstTYKZs6mim14AaABAg		0	https://www.yout	2024-03-15 7:59	0	Why sharks don	http://www.youtube.com/@CocoBebo
17	@naveenranhoti	Plz old contact n	UgxczHnPa_aP7XmDBSx4AaABAg		0	https://www.yout	2024-03-15 8:12	0	Plz old contact n	http://www.youtube.com/@naveenranhoti3914
18	@meenakshiyac	WELL DONE	UgxsqXnb45puyW59MX54AaABAg		0	https://www.yout	2024-03-15 7:14	0	WELL DONE	http://www.youtube.com/@meenakshiyac4642
19	@TruthOnly-md	It's a snapchat f	Ugy6Q19HPhBFARU9DV4AaABAg		0	https://www.yout	2024-03-15 5:14	0	It's a snapchat f	http://www.youtube.com/@TruthOnly-md4px
20	@SaahilChavan	Fuck, endobat is	UgzqR3cYuiZQa0WPM14AaABAg		0	https://www.yout	2024-03-14 15:2	0	Fuck, endobat is	http://www.youtube.com/@SaahilChavan
21	@sameersingh6	Kaash ki alsa re	UgzWV2QTxpBE9BPomZ4AaABAg		0	https://www.yout	2024-03-14 14:1	0	Kaash ki alsa re	http://www.youtube.com/@sameersingh6402
22	@Arttime-mu4y	Linda	UgzNwa1JnGICAT7QqB14AaABAg		0	https://www.yout	2024-03-14 11:3	0	Linda	http://www.youtube.com/...
23	@overunltyliver	They have not h	UgzMcoz8sB6WJeFzK494AaABAg		1	https://www.yout	2024-03-14 4:05	0	They have not h	http://www.youtube.com/@roopeshmeena6061
24	@roopeshmeeni	Ashneer fan	Ugz7QoX-igtY1AHrQ5B4AaABAg		0	https://www.yout	2024-03-14 3:18	0	Ashneer fan	http://www.youtube.com/@roopeshmeena6061
25	@aryansinghraj	BIHAR	UgzDooq-13mjpkQoGMR4AaABAg		0	https://www.yout	2024-03-14 3:09	0	BIHAR	http://www.youtube.com/@aryansinghrajput9909

Web Scrapping using Octoparse



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

Step 1 : Go to web page

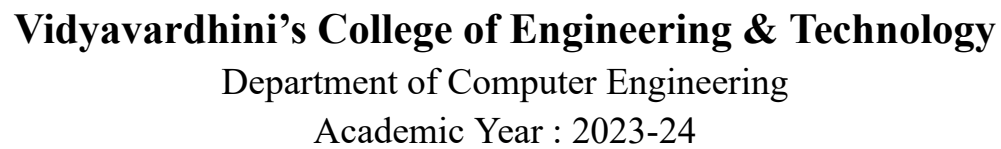
Step 2 : Create pagination

Step 3 : Build a loop item

Step 4 : Extract the data

Step 5 : Run the task and get the data

The screenshot shows the 'New Task' window of the Upgrow application. On the left is a dark sidebar with icons for Home, New, Task List, Templates, Tools, Pricing, Automation, Data Service, Inbox, Help, and Settings. The main area has a top bar with 'Home' and 'New Task' tabs. Below the tabs, there's a 'Task Group' section with a dropdown menu set to 'My Group' and a 'New Group' button. The 'URL Source' section has four buttons: 'Enter manually' (highlighted in blue), 'Import from file', 'Import from task', and 'Batch generate'. The 'URL Input' section contains a large text area with the URL: `https://www.amazon.in/s?k=redmi+note+13+pro&rid=1ENTAFXU08IVI&prefix=redmi%2Caps%2C211&ref=nb_sb_ss_ts-doa-p_1_5`. Below the text area is a small note: 'Please enter no more than 10K URLs.' and a blue 'Save' button.



The screenshot shows the Flipkart website with the Redmi 12 5G product page. The page displays the product image, specifications (4 GB RAM, 128 GB ROM, 17.12 cm display, 50MP camera, 5000 mAh battery, 1 Year warranty), and price (₹10,929). The 'Filters' section on the left shows categories like Electronics, TVs & Appliances, Men, Women, Baby & Kids, Home & Furniture, Sports, Books & More. The 'Data Preview' table at the bottom shows the extracted data for the product.

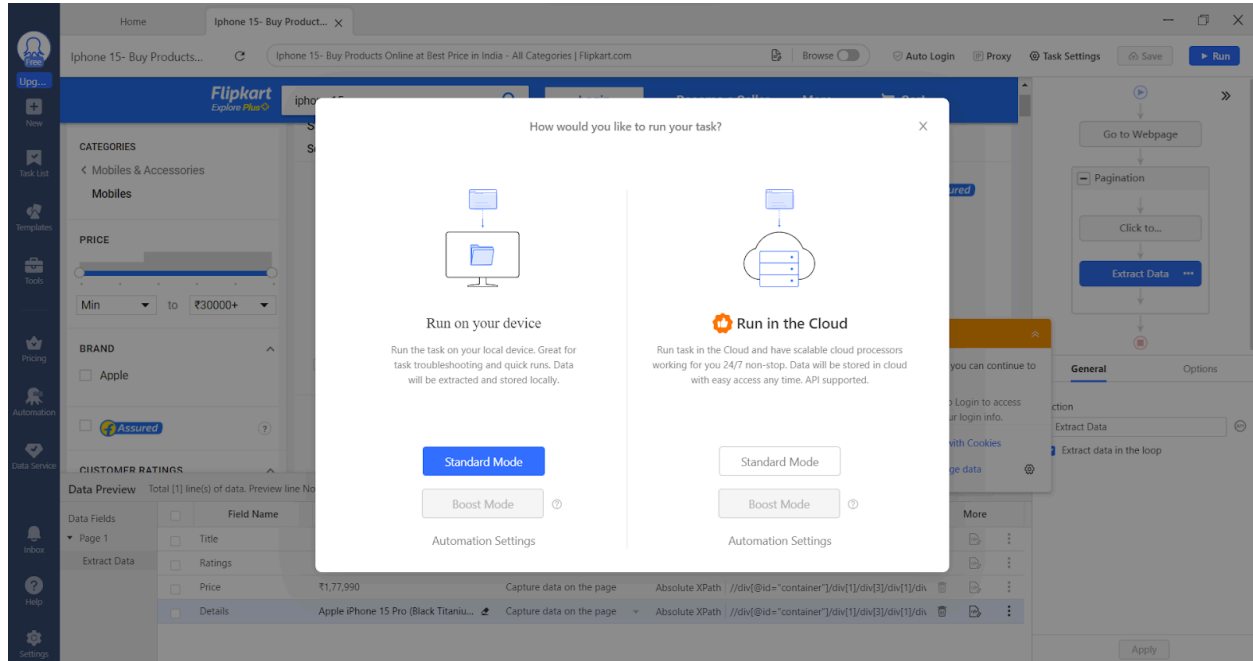
Data Fields	Field Name	Content	Field Settings	More
Page 1	Product	REDMI 13c 5G (Starlight Black, 128 GB)	Capture data on the page	Absolute XPath //div[@id="container"]/div[1]/div[3]/div[1]/div[2]/div[2]/div
Extract Data	Ratings	1,809 Ratings	Capture data on the page	Absolute XPath //div[@id="container"]/div[1]/div[3]/div[1]/div[2]/div[2]/div
	Hardware	4 GB RAM 128 GB ROM	Capture data on the page	Absolute XPath //div[@id="container"]/div[1]/div[3]/div[1]/div[2]/div[2]/div
	Price	₹10,929	Capture data on the page	Absolute XPath //div[@id="container"]/div[1]/div[3]/div[1]/div[2]/div[2]/div



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24





Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

Home | Iphone 15- Buy Product... X

Iphone 15- Buy Products Online at Best Price in India - All Categories | Flipkart.com

20 Data Extracted

Completed

Task completed

Duplicates: 2 lines Time Spent: 24s Avg. Speed: 49 lines/min

Run Export

Task Overview Data List Event Log Recent Runs

#	Title	Ratings	Price	Details
10	Apple iPhone 13 Pro Max (Graphite, 256 GB)	4.6	₹1,39,900	Apple iPhone 13 Pro Max (Graphite, 256 GB) 4.62,06 ... 15 Bionic Chip ...
11	Apple iPhone XR (Blue, 64 GB) (Includes EarPods, Power Adapter)	4.6	₹47,900	Apple iPhone XR (Blue, 64 GB) (Includes EarPods, P... ocessor iOS 13 C...
12	Apple iPhone XR (White, 128 GB)	4.6	₹52,900	Apple iPhone XR (White, 128 GB) 4.61,00,954 Rating ... ation Fast-charg...
13	Apple iPhone 13 mini (Starlight, 256 GB)	4.5	₹74,900	Apple iPhone 13 mini (Starlight, 256 GB) 4.53,798 ... 15 Bionic Chip Pro...
14	Apple iPhone SE (Red, 256 GB)	4.5	₹54,900	Apple iPhone SE (Red, 256 GB) 4.51,54,661 Ratings & ... rgers are Sold Se...
15	Apple iPhone 8 (Silver, 64 GB)	4.5	₹39,900	Apple iPhone 8 (Silver, 64 GB) 4.511,379 Ratings & ... ocessor iOS 13 C...
16	Apple iPhone 14 Pro (Deep Purple, 512 GB)	4.6	₹1,49,900	Apple iPhone 14 Pro (Deep Purple, 512 GB) 4.62,145 ... Phone and 6 M...
17	Apple iPhone 14 Pro (Deep Purple, 512 GB)	4.6	₹1,49,900	Apple iPhone 14 Pro (Deep Purple, 512 GB) 4.62,145 ... Phone and 6 M...
18	Apple iPhone 14 Pro Max (Space Black, 1 TB)	4.6	₹1,77,999	Apple iPhone 14 Pro Max (Space Black, 1 TB) 4.62,3 ... Phone and 6 Mo...
19	Apple iPhone 14 Pro (Silver, 256 GB)	4.6	₹1,29,900	Apple iPhone 14 Pro (Silver, 256 GB) 4.62,145 Rati ... Phone and 6 Mon...
20	Apple iPhone 13 mini (Starlight, 128 GB)	4.5	₹64,900	Apple iPhone 13 mini (Starlight, 128 GB) 4.53,798 ... 15 Bionic Chip Pro...

< 1 > Go to Page

Redmi 12 5g- Buy Products Online at Best Price in India - All Categories _ Flipkart.com

File Edit View Insert Format Data Tools Extensions Help

100% 123 Default... 10 B I A

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	ProductName	Rating	Price	Hardware											
2				6 GB RAM 128 GB ROM Expandable Upto 1 TB 17.25 cm (6.79 inch) Full HD+ Display 50MP + 2MP 8MP Front Camera 5000 mAh Battery Snapdragon 4 Gen 2 Processor 1 Year Manufacturer Warranty for Phone and 6 Months Warranty for In the Box Accessories											
3	REDMI 12 5G (F	4.3	₹13,499												
4	REDMI 13c 5G (4.2	₹12,188												
5				4 GB RAM 128 GB ROM Expandable Upto 1 TB 16.56 cm (6.52 inch) HD+ Display 8MP + 8MP 5MP + 5MP Dual Front Camera 5000 mAh Battery OCTA CORE Processor 1 Year and 6 months for Battery,Accessories											
6	REDMI A2+ (Cla	4.1	₹7,950												
7				4 GB RAM 128 GB ROM Expandable Upto 1 TB 16.94 cm (6.67 inch) Full HD+ AMOLED Display 48MP + 8MP + 2MP 13MP Front Camera 5000 mAh Battery Qualcomm Snapdragon 4 Gen 1 Processor 1 Year Manufacturer Warranty for Phone and 6 Months Warranty for In the Box Accessories											
8	REDMI Note 12	3.9	₹15,499												
9				4 GB RAM 128 GB ROM Expandable Upto 1 TB 17.04 cm (6.71 inch) HD+ Display 50MP Rear Camera 5MP Front Camera 5000 mAh Battery Helio G85 Processor 1 Year Manufacturer Warranty for Phone and 6 Months Warranty for In the Box Accessories											
10	REDMI 12C (Ma	4.2	₹9,380												
11				4 GB RAM 64 GB ROM Expandable Upto 512 GB 16.59 cm (6.53 inch) HD+ Display											

Redmi 12 5g- Buy Products Online at Best Price in India - All Categories _ Flipkart.com



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

Academic Year : 2023-24

Conclusion: In conclusion, this experiment showcased the practical application of web crawling, scraping, and parsing techniques using Instant Data Scraper, Netlytic, and Octoparse. Instant Data Scraper proved useful for extracting Google reviews with its user-friendly interface, while Netlytic demonstrated its efficiency in analyzing social media data by extracting YouTube comments effectively. Octoparse's flexibility and automation features made it ideal for complex web scraping tasks, including pagination and data extraction from multiple pages. These tools collectively offer a range of capabilities for web data extraction, catering to different needs and skill levels, and can be valuable assets in research, analysis, and data-driven decision-making processes.