



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:

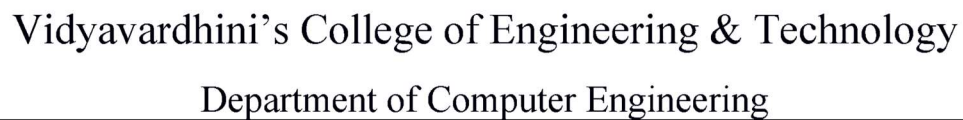
Web Crawling: Web crawling involves systematically browsing through web pages and following hyperlinks to discover and index content. It starts with a seed URL and then traverses through links to other pages, recursively fetching and indexing data. This process is essential for gathering a comprehensive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



Implementation and Output:

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

CSDL8023: Social Media Analytics Lab





Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

media-thumball-wrapper href																												
1	media-thu	n	business-business-n	categories	categories	categories	categories	track-vi	track-vi	track-map	track-map	menu	menu	href	years-in-bi	count	label	count	2	label	2	amenities	amenities	amenities	amenities	ph		
2	https://w	https://13.1	https://w	Mark A. Vi	Clinics	https://w	Dental Clin	https://w	Website	http://w	Directions	https://w	Services	https://w	https://13.38	Years	https://w	https://13.38	Years	In Practice with	Yellow Pag	Amenities:	object Ob	Wheelchai	(9			
3	https://w	https://13.2	https://w	Buchanan, Im	plant De	https://w	Cosmetic C	https://w	Website	http://w	Directions	https://w	Services	https://w	https://13.35	Years	https://w	https://13.35	Years	In Business with		Amenities:	object Ob	Wheelchai	(2			
4	https://w	https://13.3	https://w	Peter Holte	Teeth Whi	https://w	Prosthodo	https://w		http://w	Directions	https://w	Services	https://w	https://13.47	Years	https://w	https://13.47	Years	In Business		Amenities:	object Ob	Wheelchai	(2			
5	https://w	https://13.4	https://w	Columbia C	othing Al	https://w	Oral & Ma	https://w	Website	http://w	Directions	https://w	Services	https://w	https://13.14	Years	https://w	https://13.14	Years	In Business	(1)	Accredited		Amenities:	object Ob	Wheelchai	(2	
6	https://w	https://13.5	https://w	David D. Fj	Dental Clin	https://w	Clinics	https://w	Website	http://w	Directions	https://w	Services	https://w	https://13.41	Years	https://w	https://13.41	Years	In Business		Accredited		Amenities:	object Ob	Wheelchai	(2	
7	https://w	https://13.6	https://w	Valley View	Pediatric C	https://w	Dental H	https://w	Website	http://w	Directions	https://w	Services	https://w	https://13.56	Years	https://w	https://13.56	Years	In Business	(13)	Accredited		Amenities:	object Ob	Wheelchai	(9	
8	https://w	https://13.7	https://w	Preston Dk	Implant De	https://w	Teeth Whi	https://w	Website	http://w	Directions	https://w	Services	https://w	https://13.28	Years	https://w	https://13.28	Years	In Business	(1)	Accredited		Amenities:	object Ob	Wheelchai	(9	
9	https://w	https://13.8	https://w	Rio Dental	Pediatric C	https://w	Clinics	https://w	Website	http://w	Directions	https://w	Services	https://w	https://13.19	Years	https://w	https://13.19	Years	In Business		Accredited		Amenities:	object Ob	Wheelchai	(9	
10	https://w	https://13.9	https://w	Preston H		Website				http://w				https://13.11			https://13.11		In Practice	0 Years		Amenities:	object Ob	Wheelchai	(2			
11	https://w	https://13.10	https://w	Fadi Thama		Website				http://w				https://13.158			https://13.158		In Business	37 Years		Amenities:	object Ob	Wheelchai	(9			
12	https://w	https://13.11	https://w	DNA Dental	Cosmetic C	https://w				http://w				https://13.26			https://13.26		In Business	49 Years		Amenities:	object Ob	Wheelchai	(2			
13	https://w	https://13.12	https://w	Dr. Chung		Website				http://w				https://13.166			https://13.166		In Practice	55 Years		Amenities:	object Ob	Wheelchai	(9			
14	https://w	https://13.13	https://w	Dr. Harry L	Orthodont	https://w	Cosmetic C	https://w	Website	http://w				https://13.17			https://13.17		In Business	37 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
15	https://w	https://13.14	https://w	Superior S	Endodont	https://w	Orthodont	https://w	Website	http://w				https://13.12			https://13.12		In Business	30 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
16	https://w	https://13.15	https://w	Great Cgr	Dental Hyg	https://w	Teeth Whi	https://w	Website	http://w				https://13.1			https://13.1		In Business	28 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
17	https://w	https://13.16	https://w	Lake Highl	Implant De	https://w	Periodont	https://w	Website	http://w				https://13.11			https://13.11		In Business	41 Years		Accredited		Amenities:	object Ob	Wheelchai	(9	
18	https://w	https://13.17	https://w	ART OF DE	Dental Hyg	https://w	Prosthodo	https://w	Website	http://w				https://13.33			https://13.33		In Business	30 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
19	https://w	https://13.18	https://w	Smiley Den	Cosmetic C	https://w	Teeth Whi	https://w	Website	http://w				https://13.1			https://13.1		In Business	28 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
20	https://w	https://13.19	https://w	Monarch		Website				http://w				https://13.11			https://13.11		In Business	41 Years		Accredited		Amenities:	object Ob	Wheelchai	(9	
21	https://w	https://13.20	https://w	Lopez Rosi	Implant De	https://w	Teeth Whi	https://w	Website	http://w				https://13.33			https://13.33		In Business	33 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
22	https://w	https://13.21	https://w	James Loui	Teeth Whi	https://w	Oral & Ma	https://w	Website	http://w				https://13.31			https://13.31		In Business	31 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
23	https://w	https://13.22	https://w	Lynn Denti	Dental Hyg	https://w	Dental Clin	https://w	Website	http://w				https://13.31			https://13.31		In Business	44 Years		Accredited		Amenities:	object Ob	Wheelchai	(9	
24	https://w	https://13.23	https://w	McWorthe	Pediatric C	https://w			Website	http://w				https://13.31			https://13.31		In Business	44 Years		Accredited		Amenities:	object Ob	Wheelchai	(9	
25	https://w	https://13.24	https://w	Montoya S	Dental Hyg	https://w	Dental Clin	https://w	Website	http://w				https://13.31			https://13.31		In Business	24 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
26	https://w	https://13.25	https://w	Palmer Ma	Oral & Ma	https://w	Implant De	https://w	Website	http://w				https://13.31			https://13.31		In Business	24 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
27	https://w	https://13.26	https://w	Dr. Connie		Website				http://w				https://13.31			https://13.31		In Business	44 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
28	https://w	https://13.27	https://w	Tucker All	Dental Hyg	https://w	Dental Clin	https://w	Website	http://w				https://13.31			https://13.31		In Business	44 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
29	https://w	https://13.28	https://w	Deepti S M	Teeth Whi	https://w	Orthodont	https://w	Website	http://w				https://13.31			https://13.31		In Business	44 Years		Accredited		Amenities:	object Ob	Wheelchai	(2	
30	https://w	https://13.29	https://w	Evans J Tay	Teeth Whi	https://w	Orthodont	https://w	Website	http://w				https://13.31			https://13.31		In Business	23 Years		Accredited		Amenities:	object Ob	Wheelchai	(9	
31	https://w	https://13.30	https://w	Marsh Lan	Implant De	https://w	Teeth Whi	https://w	Website	http://w				https://13.31			https://13.31		In Business	23 Years		Accredited		Amenities:	object Ob	Wheelchai	(9	

H93																											
1	media-thumball-wrapper hrefmedia-thumball																										

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



[Home](#) [About](#) [Resources For...](#) [Help](#) [My Datasets](#) [New Dataset](#) [My Account](#) [Log Out](#)

[1. Edit](#) / [2. Preview](#) / [3. Text Analysis](#) / [4. Network Analysis](#) / [5. Report](#)

[Twitter\(discontinued\)](#) [YouTube](#) [Google Sheets](#) [Text File](#) [RSS](#) [Reddit/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.

(No Special Characters)

Enter the ID of the YouTube video as follows:

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



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering

netlytic

Home About Resources For... Help My Datasets New Dataset My Account Log Out

1. Edit / 2. Preview / 3. Text Analysis / 4. Network Analysis / 5. Report

26%

- > Total records in the dataset before the import: 0
- > Retrieving data... **Don't close the browser!**
- > Retrieving **up to 2500** top-level comments and up to 5 one-level replies per comment.
- > Warning: Data collection was stopped as the number of retrieved comments+replies exceeds the limit of 2500.
- > Processing retrieved data.
- > Retrieved 2200 top-level comments + 308 one-level replies.
- > **There are more comments than what can be stored in your account. - Get More Space**
- > Saving data...

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	id	author	description	guid	to	likecount	link	pubdate	replycount	title	authorChannelId						
1	1	@Samsung	Got what iGai Ugg27UTX07l_K0yUB4AAf1116				https://www.yi2024-01-17 1:145			Got what iGai http://www.youtube.com/@Samsung							
2	2	@leosroth	iPhone 15 prc Ugg27UTX07l_@Samsung			66	https://www.yi2024-01-17 1:0			iPhone 15 prc http://www.youtube.com/@leosroth/rontofsalear813							
3	3	@jessaid	@@Leosr Ugg27UTX07l_@Samsung			15	https://www.yi2024-01-17 1:0			@@Leosr http://www.youtube.com/@jessaid							
4	4	@weinserr	Great phone i Ugg27UTX07l_@Samsung			4	https://www.yi2024-01-17 1:0			Great phone i http://www.youtube.com/@weinserr7278							
5	5	@MyBusiness	yes even in fr Ugg27UTX07l_@Samsung			0	https://www.yi2024-01-17 1:0			yes even in fr http://www.youtube.com/@MyBusinessTrips							
6	6	@winwison	translate Ugg27UTX07l_@Samsung			13	https://www.yi2024-01-17 1:0			translate http://www.youtube.com/@winwisoncherry							
7	7	@spal9003	S25 Launch c Ugg1567w6R9ySvubm4A0				https://www.yi2024-03-30 0:0			S25 Launch c http://www.youtube.com/@spal9003							
8	8	@festimhafe	Best smartph UgySn656RmsR0r0B4A0				https://www.yi2024-03-29 2:0			Best smartph http://www.youtube.com/@festimhafe9280							
9	9	@mohamedfessiou	Nice smartph UgyAN7255U-In0Zn7Z4A0				https://www.yi2024-03-29 2:0			Nice smartph http://www.youtube.com/@mohamedfessiou3379							
10	10	@lewakdarius	Please Samsi Ugyw2VNF1QIFng2pln4A0				https://www.yi2024-03-28 1:0			Please Samsi http://www.youtube.com/@lewakdarius3636							
11	11	@Travelandthe	Time to replac Ugy5J2Hedp2NwR4A0			1	https://www.yi2024-03-28 1:0			Time to replac http://www.youtube.com/@Travelandthe1							
12	12	@desirerangha	We need this Ugy9y21GaiKvnhleN4A0				https://www.yi2024-03-28 0:0			We need this http://www.youtube.com/@desirerangha3463							
13	13	@danny-eq8rk	apple where y UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-28 0:1			apple where y http://www.youtube.com/@danny-eq8rk							
14	14	@zorred519	Miles ahead UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-28 0:0			Miles ahead http://www.youtube.com/@zorred5191							
15	15	@yadedimpr	got the phone Ugya258R9GnkVof-zbM4A0			2	https://www.yi2024-03-28 0:0			got the phone http://www.youtube.com/@yadedimprmacapugy354							
16	16	@noobhonts	SAMSUNG A UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-27 2:0			SAMSUNG A http://www.youtube.com/@noobhonts2510							
17	17	@thanshghv	I thought iPh UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-27 2:0			I thought iPh http://www.youtube.com/@thanshghv6201							
18	18	@user-j5f5y9e	The speaker i Ugy2GdFuzseeYqYnp4A0				https://www.yi2024-03-27 2:0			The speaker i http://www.youtube.com/@user-j5f5y9e							
19	19	@ninomunies	happy! UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-27 1:0			happy! http://www.youtube.com/@ninomunies27401							
20	20	@kokoijames	I got an idea Ugyw2VNF1QIFng2pln4A0				https://www.yi2024-03-27 1:0			I got an idea http://www.youtube.com/@kokoijames1768							
21	21	@5G5gamer42	Scratch resist UgyCzRyR0vHvSAIKV4A0				https://www.yi2024-03-26 0:0			Scratch resist http://www.youtube.com/@5G5gamer42							
22	22	@the_trouble_makers103	Galaxy Ai is a UgyEQ4Jv0xH5QIHV1k4A0				https://www.yi2024-03-26 0:0			Galaxy Ai is a http://www.youtube.com/@the_trouble_makers103							
23	23	@quao0680	Samsung galli Ugy23Gz5NDeLwV_4H4A0				https://www.yi2024-03-25 0:0			Samsung galli http://www.youtube.com/@quao0680							
24	24	@mecon0805	But it will com UgytlnkGm9Gd4A0				https://www.yi2024-03-25 0:0			But it will com http://www.youtube.com/@mecon0805							
25	25	@Verbrecher3000	You can call i Ugydm12-vS8JC8YV4A0				https://www.yi2024-03-24 1:0			You can call i http://www.youtube.com/@Verbrecher3000							
26	26	@vekkhesoch3239	Im here to we UgyS858PvQVMauCE354A0				https://www.yi2024-03-24 1:0			Im here to we http://www.youtube.com/@vekkhesoch3239							
27	27	@gunuganer8632	Small sensor UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-24 0:0			Small sensor http://www.youtube.com/@gunuganer8632							
28	28	@skintusw01	LOVE IT! UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-24 0:0			LOVE IT! http://www.youtube.com/@skintusw01348							
29	29	@fardidafad2701	I have purcha UgyRyCyFvFZSLM4A0				https://www.yi2024-03-24 0:0			I have purcha http://www.youtube.com/@fardidafad2701							
30	30	@user-ef9v5j8c	Green line ph UgydGakFzV7AUh-DZ4A0				https://www.yi2024-03-23 2:0			Green line ph http://www.youtube.com/@user-ef9v5j8c							

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

Home New Task X

Task Group: My Group New Group

URL Source: Enter manually Import from file Import from task Batch generate

URL Input: https://www.amazon.in/s?k=samsung+s24&crd=18DC2697WE1GK&prefix=sama%2Caps%2C256&ref=nb_sb_ss_doa-p_5_4

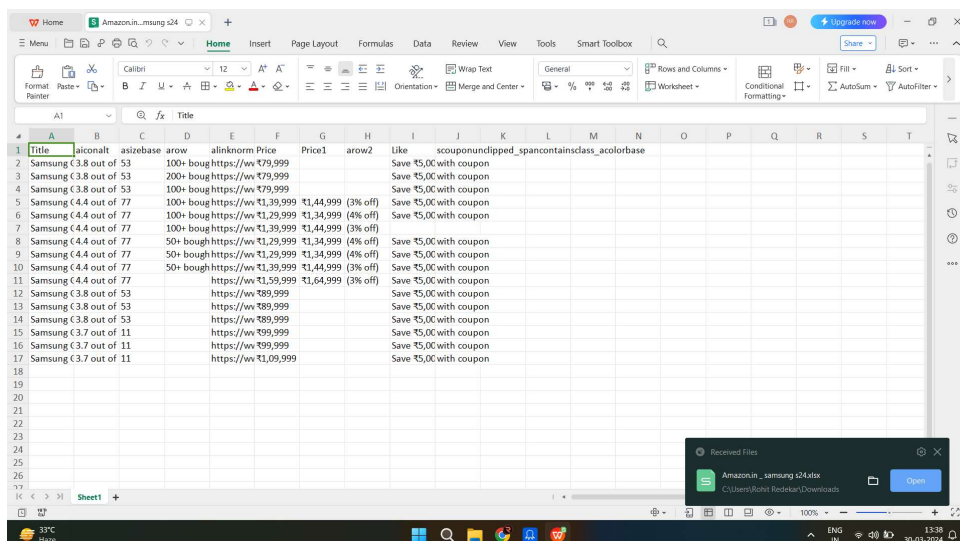
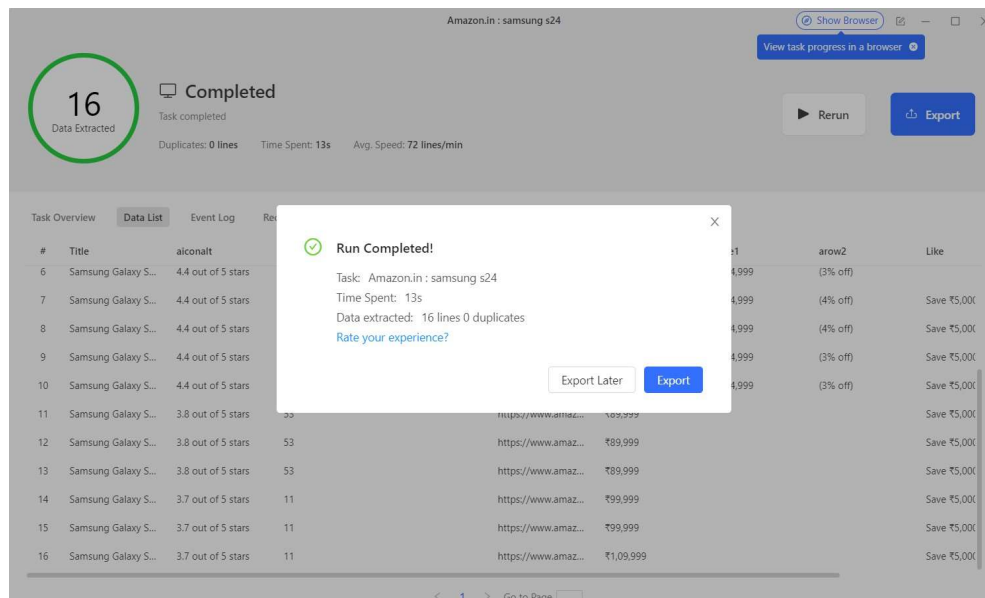
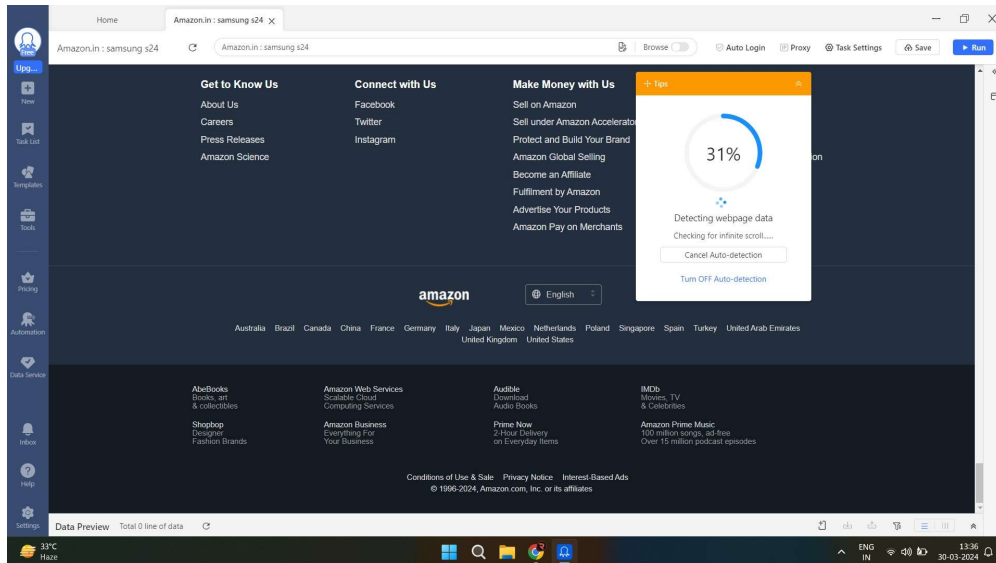
Please enter no more than 10K URLs.

Save



Vidyavardhini's College of Engineering & Technology

Department of Computer Engineering





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.