

## SMA EXPERIMENT NO. 2

**Roll No.: B856**

**Date:**

**Aim:** To perform data collection- select the social media platform of your choice (Facebook, Twitter, Youtube, Instagram) connect to and capture social media data for business using octoparse (scrapping, crawling and parsing)

### **Theory:**

Social media platforms like Twitter, Facebook, LinkedIn, YouTube, and web blogs have become invaluable sources of data for businesses looking to gain insights into customer behavior, market trends, and brand perception. With millions of users sharing their thoughts, opinions, and experiences daily, businesses can leverage this data to make informed decisions, optimize marketing strategies, and improve customer engagement. To extract meaningful information from these vast data pools, businesses often rely on various data collection methods, including scraping, crawling, and parsing. Each method plays a critical role in gathering, organizing, and analyzing social media data. In this article, we will explore these methods in detail and their applications for businesses.

### **1. Scraping**

Scraping is the process of extracting data directly from websites or social media platforms using automated tools. It involves accessing a website, identifying the data of interest, and pulling it into a structured format for analysis. Scraping is one of the most commonly used methods for collecting data from social media platforms like Twitter, Facebook, and LinkedIn, as it allows businesses to capture real-time information such as user posts, comments, likes, shares, and more.

### **2. Crawling**

Crawling refers to the process of discovering and indexing web pages through automated tools or bots. This method is often used to scan websites and social media platforms to identify new content. Unlike scraping, which focuses on extracting specific data from a page, crawling is about navigating the web, discovering new pages, and indexing content for further processing.

In social media data collection, crawling is primarily used to gather content from sources like web blogs, news websites, and sometimes even from public social media profiles or forums. Crawlers visit web pages, follow links, and gather information from multiple sources, enabling businesses to monitor and track changes across the web.


### **3. Parsing**

Parsing is the process of analyzing and extracting structured or unstructured data from the raw content of web pages. It involves using tools or software to identify relevant patterns, tags, or attributes in the HTML (or other formats like XML) of a page. Parsing is often combined with scraping to extract specific data from social media platforms, such as Twitter or Facebook.

### **Output:**

Hello, ayushijii. Glad to see you here!


Search for task templates or type webpage URL(s)→ Start



**What's a custom task?**

Auto-detect webpages and set up custom workflows to extract data from any website.

[+ New Task](#)[Tutorial](#)



**What's a task template?**


Choose from a selection of prebuilt task templates to get data instantly with zero setup.

[88 Templates](#)[Tutorial](#)


Popular Templates

E-Commerce   Lead Generation   Social Media   Real Estate   Jobs   Maps


[See All >](#)



Xiaohongshu Scraper  
[Template Bundle](#)

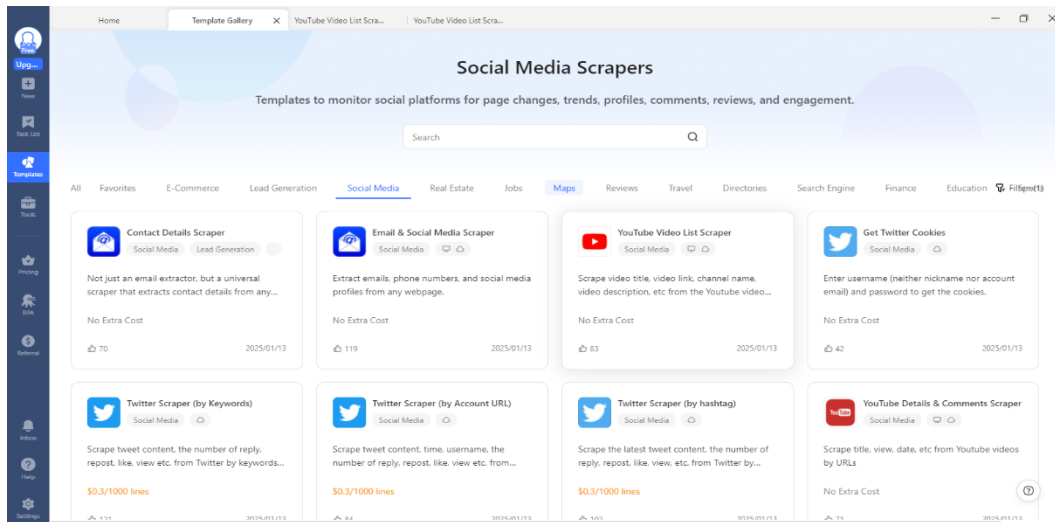


Twitter Scraper  
[Template Bundle](#)

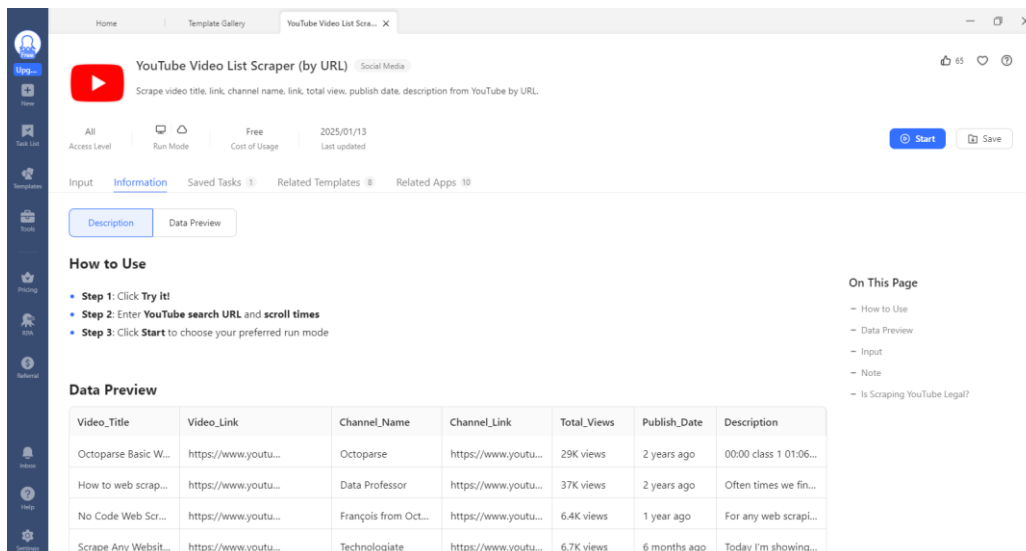


Google Maps Scraper  
[Template Bundle](#)

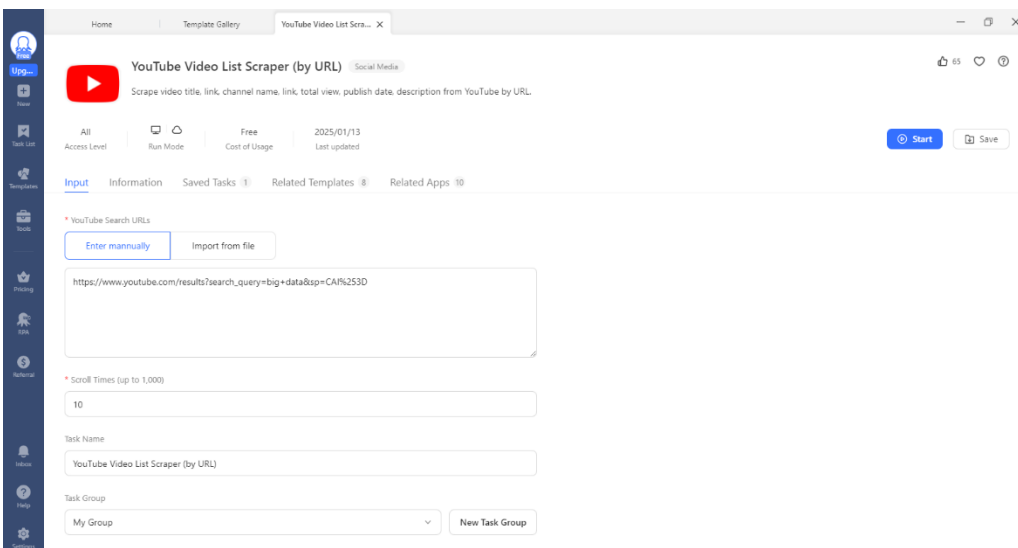
## Step1: Select a social media platform for data collection.



## Step2: Connect to the platform using Octoparse.



## Step3: Use scraping to extract the required data.



#### Step4: Employ crawling to discover and index web pages.

The screenshot shows the YouTube Video List Scraper interface. The browser address bar displays 'big data - YouTube'. The main content area shows a video player for a video titled '#AI (Artificial Intelligence)#MachineLearning#IoT (Internet of Things)#BigData#TechNews'. Below the video player, there is a table with columns: #, Keyword, Video\_Title, Video\_Link, Channel\_Name, Channel\_Link, Total\_Views, Publish\_Date, and Description. The table lists 5 videos.

#	Keyword	Video_Title	Video_Link	Channel_Name	Channel_Link	Total_Views	Publish_Date	Description
1	big data	Diving into the worl...	https://www.youtub...	Roberto C Santos	https://www.youtub...	No views	1 hour ago	Experimento de ediç...
2	big data	Big data activity Pow...	https://www.youtub...	Erika Peñaranda Mejia	https://www.youtub...	No views	2 hours ago	
3	big data	Simple Guide to ...	https://www.youtub...	智慧岛	https://www.youtub...	No views	4 hours ago	#SQL #StructuredDa...
4	big data	Diving into the worl...	https://www.youtub...	Roberto C Santos	https://www.youtub...	No views	6 hours ago	Experimento criado ...
5	big data	Presentasi UTS Mata...	https://www.youtub...	Titus tri	https://www.youtub...	No views	6 hours ago	Titus tri purbo asmo...

#### Step5: Parse the collected data for meaningful insights.

The screenshot shows the YouTube Video List Scraper interface. The top status bar indicates '40 Data Extracted' and 'Completed'. Below this, there is a table with columns: #, Keyword, Video\_Title, Video\_Link, Channel\_Name, Channel\_Link, Total\_Views, Publish\_Date, and Description. The table lists 20 videos.

#	Keyword	Video_Title	Video_Link	Channel_Name	Channel_Link	Total_Views	Publish_Date	Description
10	big data	UAS Big Data Analys...	https://www.youtub...	Arkan Hilman Hakim	https://www.youtub...	6 views	16 hours ago	Nama : Arkan Hilma...
11	big data	Creating And Queryi...	https://www.youtub...	Sasta Engineer Tutor...	https://www.youtub...	No views	17 hours ago	mongodb #indexesi...
12	big data	Big Data Analytics a...	https://www.youtub...	M. Yaman Darmawan	https://www.youtub...	No views	17 hours ago	
13	big data	Tugas UAS Big Data ...	https://www.youtub...	SUBHAN ALGIFARI	https://www.youtub...	No views	18 hours ago	Vidio ini dibuat unt...
14	big data	(UAS) BIG DATA AN...	https://www.youtub...	ADITYA YOGI PRATA...	https://www.youtub...	18 views	19 hours ago	Nama : Aditya Yogi ...
15	big data	Presentasi Bedah Pa...	https://www.youtub...	Ali Ridho Wahyu Fitr...	https://www.youtub...	No views	19 hours ago	202131420004 - Ali ...
16	big data	Presentasi APACHE ...	https://www.youtub...	Ali Ridho Wahyu Fitr...	https://www.youtub...	No views	20 hours ago	202131420004 - Ali ...
17	big data	Presentasi Big Data ...	https://www.youtub...	Ali Ridho Wahyu Fitr...	https://www.youtub...	No views	20 hours ago	202131420004 - Ali ...
18	big data	UAS Big Data Analys...	https://www.youtub...	Oka Muhamad Nurf...	https://www.youtub...	No views	21 hours ago	Nama : Oka Muham...
19	big data	[UAS] Big Data Anal...	https://www.youtub...	Henry Adam	https://www.youtub...	No views	22 hours ago	Materi yang ada di ...
20	big data	Phone dekhte hai us...	https://www.youtub...	Vihaan Maahir Warri...	https://www.youtub...	No views	22 hours ago	Phone dekhte hai us...

**Conclusion:** Octoparse enables businesses to efficiently collect social media data from platforms like Twitter, Facebook, YouTube, and Instagram through scraping, crawling, and parsing. This automation helps businesses gain valuable insights into customer behavior, trends, and brand perception. However, it's essential to adhere to platform policies and ethical guidelines during the data collection process.