

Experiment No. 2

Aim : Data collection from different social media platform using Octoparse web scrapping tool.

Theory :

Octoparse is a modern visual web data extraction software. Both experienced and inexperienced users would find it easy to bulk extract information from websites with it. For most scraping tasks, no coding is needed. Octoparse supports Windows XP, 7, 8, 10. It works well for both static and dynamic websites, including those web pages using Ajax. To export the data, there are various data formats of your choice like CSV, EXCEL, HTML, TXT, and databases (MySQL, SQL Server, and Oracle via API). Octoparse simulates human operation to interact with web pages. Octoparse provides a visual operation pane, which is very user-friendly and straightforward. It simulates human web browsing behavior like opening a web page, logging into an account, entering text, pointing-and-clicking the web element, etc. Just click the information on the website in the built-in browser and start the extraction, and you will get the structured data you need. There are 2 extraction modes (Task Template and Advanced Mode) in Octoparse. It takes you only half an hour to get started with Octoparse, and people who have programming experience would spend less time getting familiar with Octoparse.

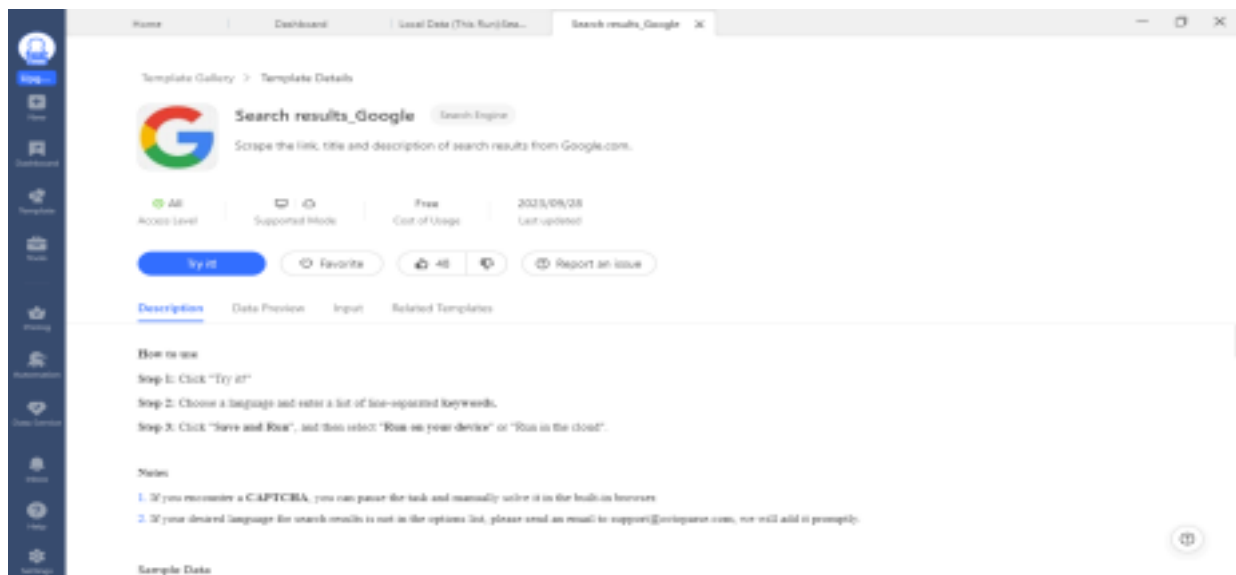


Fig 1. Target social media platform – Search results_google

#	Language	Keyword	Title	Detail URL	Description	Source	ErrorMessage
1	简体中文	file	【File75】 【File75】...	https://forums.x-plane.co...	2天前 — File Informatio...	https://forums.x-plane.co...	
2	简体中文	file	Sun Country (NEZ) ...	https://theflightline.co.uk...	12小时前 — RAR file, en...	https://theflightline.co...	
3	简体中文	file	Railroads Enhanced - Im...	https://thegta5-mods.co...	15小时前 — Find the fi...	https://thegta5-mods.co...	
4	简体中文	file	在浏览器中打开文件，保...	https://www.lovepdf.co...	在浏览器中打开文件可保...	https://www.lovepdf.com	
5	简体中文	file	金碧山景酒店：房价最...	https://www.ubs.com.au...	3天前 — Coulburn Sum...	https://www.ubs.com.au	
6	简体中文	file	Hbo hd m3u8	https://dtkillas.pl/hbo-h...	If you have an M3U8 U...	https://dtkillas.pl	
7	简体中文	file	Python file read教程, m...	https://blog.51cto.com/...	5天前 — Python file rea...	https://blog.51cto.com	
8	简体中文	file	Dj Okawari Power Danc...	https://den-software.pl/...	Multiple file transfer. Pu...	https://den-software.pl	
9	简体中文	file	[Perrin] PwID 设置教程...	https://www.tpoint360.c...	12小时前 — The Office ...	https://www.tpoint360.c...	
10	简体中文	file	【AUX.TSX】 【Zao7】...	https://x-plane.to/file/1...	1天前 — ... file, © Alex...	https://x-plane.to	
11	简体中文	file	解构tech的烦恼：Dwa...	https://segmentfault.co...	18小时前 — 简介、源码...	https://segmentfault.co...	

Fig 2(1). Data collection using Octoparse

#	Language	Keyword	Title	Detail URL	Description	Source	ErrorMessage
1	简体中文	human	Inhibitory effect of luteal...	https://www.zgggw.com...	作者：XYS - 2018 — Co...	https://www.zgggw.com	
2	简体中文	human	Recombinant Human S...	https://www.peprotech.co...	BCA-1/5UC, a CXC chem...	https://www.peprotech...	
3	简体中文	human	Key Takeaways from Chi...	https://www.usu.edu/v/...	2019年6月18日 — 虽然...	https://www.usu.edu	
4	简体中文	human	Human Resources & Or...	https://scholar.google.c...	S.The International Jou...	https://scholar.google.c...	
5	简体中文	human	BT Generation Robot V...	https://www.mouser.co...	Table of Contents Hum...	https://www.mouser.com	

Fig 2(2). Data collection using Octoparse

#	Language	Keyword	Title	Detail URL	Description	Source
1	简体中文	file	[BT] [Keep original format] Co...	https://www.cnblogs.com/alex-b...	10分钟前 — If the existing Execf ...	https://www.cnblogs.com
2	简体中文	file	Winb m3u8	https://dckillas.pl/winb-m3u8.ht...	... file, and a file with m3u8.txt...	https://dckillas.pl
3	简体中文	file	Shared Chest - Palworld	https://www.nexusmods.com/pal...	5天前 — ... files are needed. Files...	https://www.nexusmods.com
4	简体中文	file	Youtube tv m3u - domena mam...	https://mamweb.pl/youtube-tv-...	Download M3U8 file, PL5 file, XS...	https://mamweb.pl
5	简体中文	file	CVE-2024-25113: diffoscope GP...	https://vuldb.com/cv/5id/250403	22小时前 — 导出漏洞影响的...	https://vuldb.com
6	简体中文	file	M3u8 cp24 - Granulatin	https://granulatin.pl/m3u8-cp24...	... file, and a file with it relies on ...	https://granulatin.pl
7	简体中文	file	X-UI, a multi-user XRay graphical...	https://seekfind.github.io/2021/...	2021年10月10日 — ... file /root/...	https://seekfind.github.io
8	简体中文	file	Apttool 366v - Fast	https://testfima.pl/apktool.html	now you get a file folder in that f...	https://testfima.pl
9	简体中文	file	Fgt mko - email.pl	https://email.pl/fgs-mko.html	mko File Size: 24. Waj. GGP3. x2...	https://email.pl
10	简体中文	file	【AlexTex】《ZiboT38》Shando...	https://forums.e-plane.org/index...	2天前 — File Information, View...	https://forums.e-plane.org
11	简体中文	file	San Country (N6385Y) - PMD0 T...	https://ch.rihtaim.to/file/69065...	12小时前 — RAR File, extract the ...	https://ch.rihtaim.to
12	简体中文	file	Railroads Enhanced - Improved r...	https://ch.gta3-mods.com/maps...	13小时前 — Find the file diclist...	https://ch.gta3-mods.com
13	简体中文	file	在线压缩PDF文件, 保持与源PDF...	https://www.lovepdf.com/zh-cn...	压缩PDF文件时可保持与源PDF文...	https://www.lovepdf.com
14	简体中文	file	金銀信譽鑄造: 澳洲央行鑄造...	https://www.sbs.com.au/languag...	5天前 — Doublemint Sunrise (File L...	https://www.sbs.com.au
15	简体中文	file	Hbo hd m3u8	https://dckillas.pl/Hbo-hd-m3u8...	If you have an M3U8 URL, or an ...	https://dckillas.pl
16	简体中文	file	Python file read读_mob648e5...	https://blog.51cto.com/u_16175...	5天前 — Python file read读... ..	https://blog.51cto.com

Fig 3. Collected Data in CSV format

Conclusion: In conclusion, our exploration into data collection through social media platforms utilizing the Octoparse web scraping tool has yielded invaluable insights. Through meticulous extraction and analysis, we have unearthed a wealth of information crucial for understanding user behaviors, preferences, and trends. This process has not only enhanced our comprehension of social media dynamics but also provided a foundation for informed decision-making in various fields, from marketing strategies to sociological research. As we navigate the ever-evolving digital landscape, the integration of sophisticated tools like Octoparse promises to continually refine our ability to harness data effectively, unlocking new opportunities for innovation and understanding.