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.

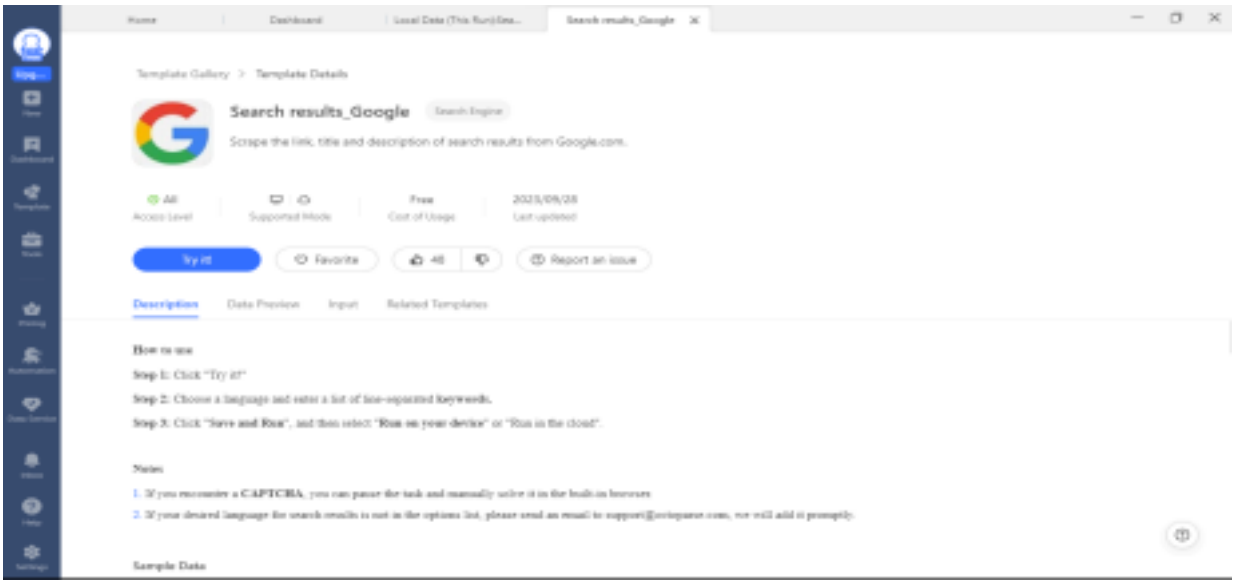

Fig 1. Target social media platform – Search results_google

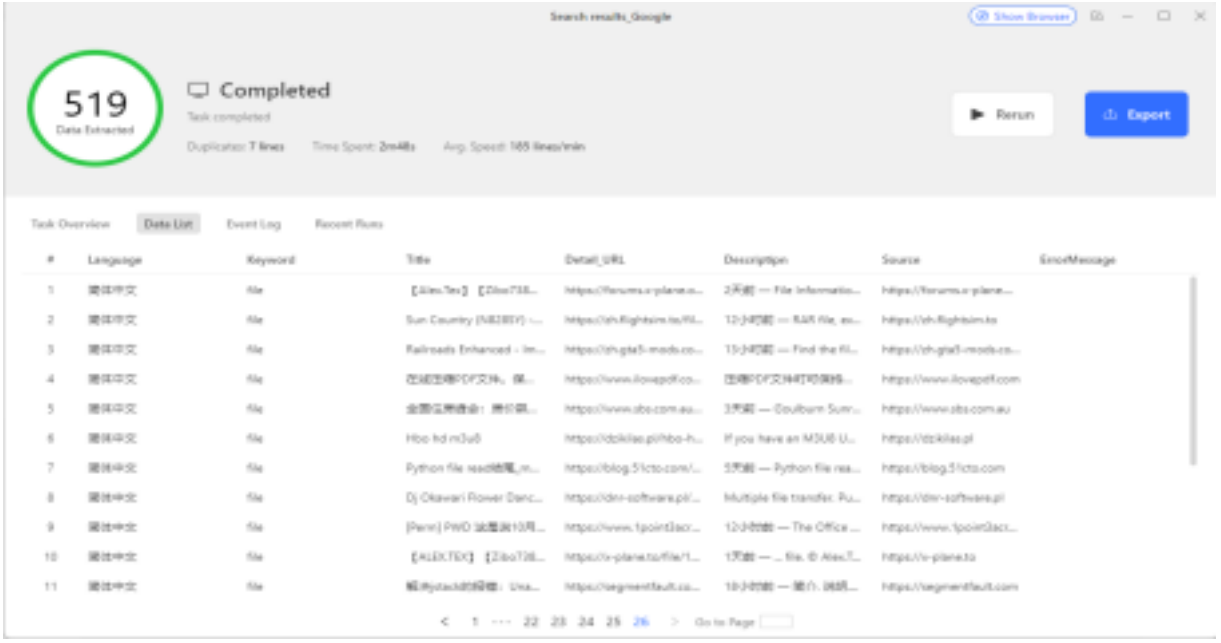

Fig 2(1). Data collection using Octoparse

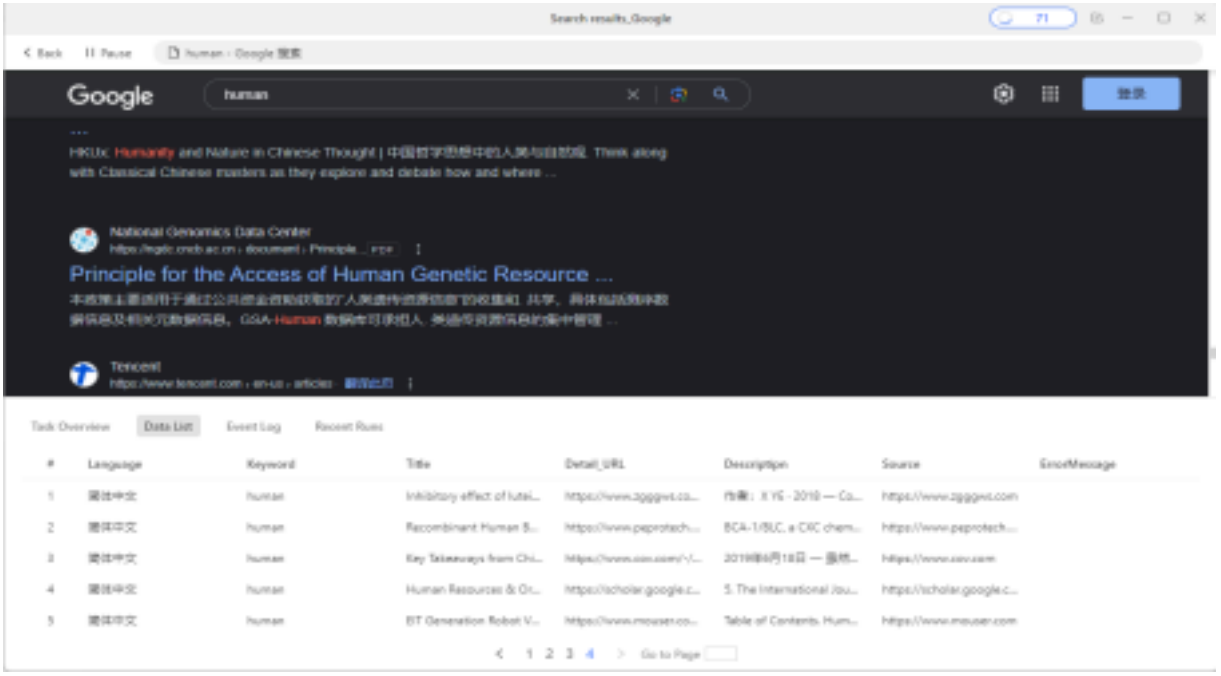

Fig 2(2). Data collection using Octoparse

\

#	Language	Keyword	Title	Detail_URL	Description	Source
1	简体中文	file	[971] [Keep original formats] Co...	https://www.cnblogs.com/alex-b...	10分钟前 — If the existing Excel ...	https://www.cnblogs.com
2	简体中文	file	Wnbc m3u8	https://dzikilas.pl/wnbc-m3u8.ht...	... file, and a file with .m3u","cont...	https://dzikilas.pl
3	简体中文	file	Shared Chest - Palworld	https://www.nexusmods.com/pal...	5天前 — ... files are needed. Plea...	https://www.nexusmods.com
4	简体中文	file	Youtube tv m3u - domena mam...	https://mamweb.pl/youtube-tv-...	Download M3U8 file, PLS file, XS...	https://mamweb.pl
5	简体中文	file	CVE-2024-25711: diffoscope GP...	https://vuldb.com/zh/?id.255405	22小时前 — [illegible]	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://seakfind.github.io/2021/...	2021年10月10日 — ... file /root/...	https://seakfind.github.io
8	简体中文	file	Apktool 368w - Fast	https://fastfirma.pl/apktool.html	now you get a file folder in that f...	https://fastfirma.pl
9	简体中文	file	Fgt mkv - ermal.pl	https://ermal.pl/fgt-mkv.html	mkv File Size: 24. Way. DDP5. x2...	https://ermal.pl
10	简体中文	file	【Alex.Tex】【Dbo738】Shando...	https://forums.x-plane.org/index...	2天前 — File Information. Views ...	https://forums.x-plane.org
11	简体中文	file	Sun Country (N8288Y) - PMDG 7...	https://zh.flightsim.to/file/69065...	12小时前 — RAR file, extract the ...	https://zh.flightsim.to
12	简体中文	file	Railroads Enhanced - Improved r...	https://zh.gta5-mods.com/maps...	13小时前 — Find the file dlclist.x...	https://zh.gta5-mods.com
13	简体中文	file	在线压缩PDF文件，保持质量PDF...	https://www.ilovepdf.com/zh-cn...	压缩PDF文件时可保持质量PDF文...	https://www.ilovepdf.com
14	简体中文	file	[illegible]	https://www.sbs.com.au/languag...	5天前 — Goulburn Sunrise (File I...	https://www.sbs.com.au
15	简体中文	file	Hbo hd m3u8	https://dzikilas.pl/hbo-hd-m3u8...	If you have an M3U8 URL or an ...	https://dzikilas.pl
16	简体中文	file	Python file read[illegible]_mob648e5...	https://blog.51cto.com/u_16175...	5天前 — Python file read[illegible]...	https://blog.51cto.com

Fig 3. Collected Data in CSV format

<u>Conclusion</u> : 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.