

Лекція 9. Парсинг даних: поняття та інструменти для вебскрейпінгу



Валентина Корольчук, Таїсія Саяпіна,
Тетяна Волошина

кафедра інформаційних систем і технологій

Збирання даних вебсайту



сторінка сайту



технологія вебскрейпінгу



структуровані дані

Навіщо використовувати інструменти вебзбирання?



Дослідження ринку



Витяг контактної інформації



Фінансові дані



Пошук роботи та співробітників



Відстеження цін у різних магазинах



Страховання

Вебсканування

Вебсканування (індексація) - це процес, який використовує ботів, також відомих як сканери, для індексації вмісту вебсайту

Вебсканери використовуються



пошуковими системами (Google, Bing)



статистичними організаціями



вебагрегаторами

вебсканування збирає
загальні дані

Основні переваги сканування в Інтернеті



Аналіз і курація контенту:

відстежуючи активність користувачів, вебсканери можуть використовуватися для кращого вивчення поведінки користувачів

збираючи різні дані, вебсканери відстежують поведінку користувачів, усвідомити дії читачів

Основні переваги сканування в Інтернеті



Цільовий список:

вебсканери дозволяють створити цільовий список компаній або окремих контактів для різних цілей

сканер дає змогу отримувати таку інформацію, як номери телефонів, адреси та адреси електронної пошти

скласти список цільових вебсайтів, які надають відповідні списки компаній

Основні переваги сканування в Інтернеті



Отримати інформацію про те, що говорять про Вас і Ваших колег у соціальних мережах

Вебсканери можуть допомогти:



отримати інформацію про те, що говорять про Вас у соціальних мережах



відстежувати коментарі користувачів (клієнтів), зроблені на інших вебсайтах

Основні переваги сканування в Інтернеті



Підтримка поточних тенденцій галузі

маючи доступ до величезної кількості даних з різних джерел вебсканери дозволяють відстежувати тенденції галузі

Вебскрейпінг

Вебскрейпінг - це процес, який автоматизує вилучення певних наборів даних за допомогою ботів, часто відомих як «скрепери»

вебскрейпінг фокусується на окремих фрагментах набору даних, які можна використовувати для:



порівняння



аналізу відповідно до вимог і цілей

Переваги вебскрейпінгу



Ефективне управління даними

змога отримувати дані з численних вебсайтів

заощадження часу на копіювання та структурування даних

збереження та захист отриманих даних

Переваги вебскрейпінгу



Точність та швидкість збору даних

правильне вилучення даних має вирішальне значення для подання будь-якої інформації

залежить від складності використовуваних проєктів, ресурсів і технологій

Переваги вебскрейпінгу



Низький рівень обслуговування та простота реалізації

онлайн-методи вебскрейпінгу не потребують обслуговування

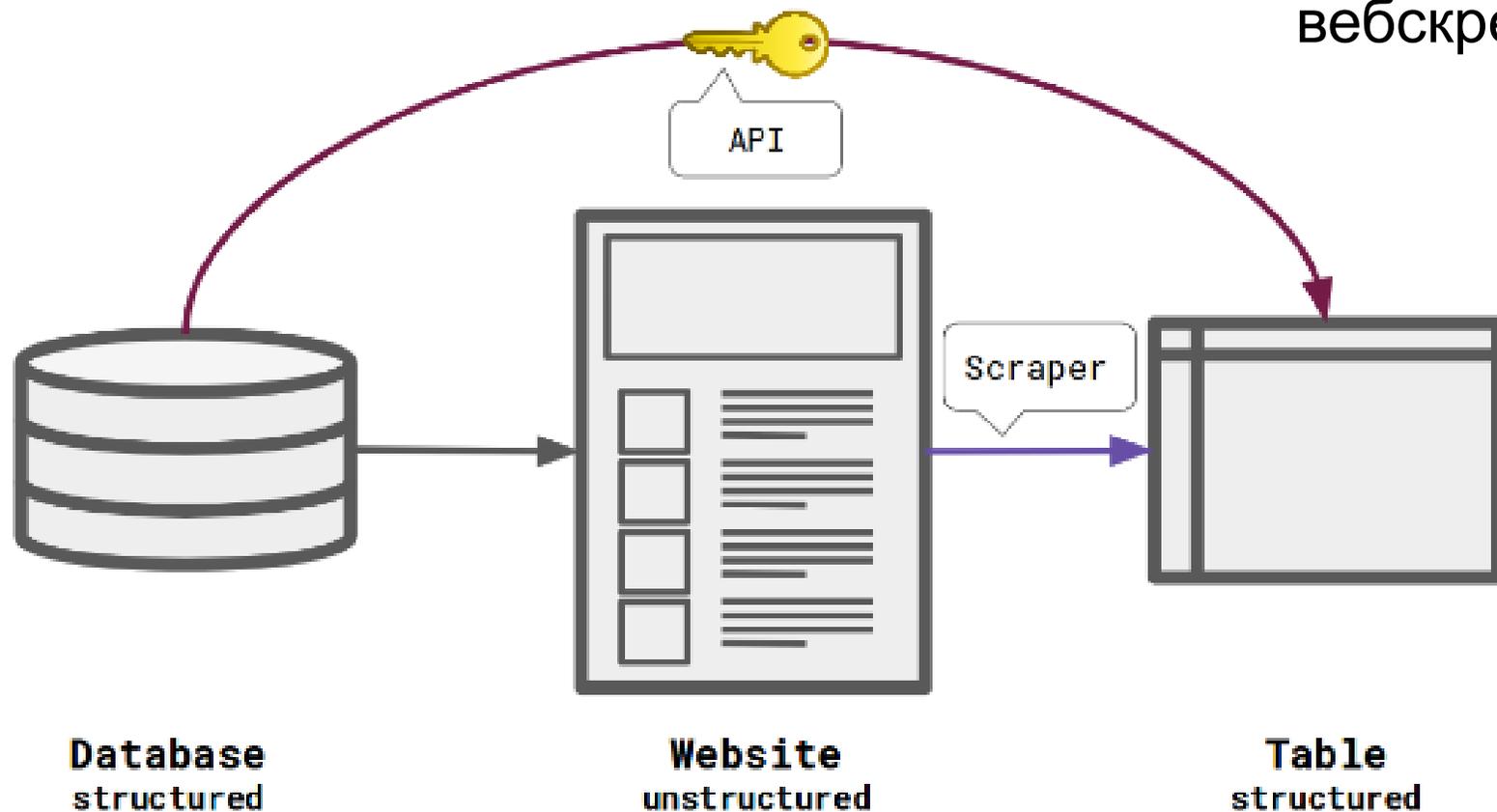
можна зібрати велику кількість даних з мінімальними витратами та відповідно максимальну цінність

Переваги вебскрейпінгу



Економічно вигідно

мінімізація витрат, завдяки API
вебскрейпінгу



Вебскребкування

Вебскребкування (вебзбір або вилучення вебданих) - різновидом скрейпінгу даних, який використовується для збору інформації з вебсайтів

Програмне забезпечення для вебскрейпінгу може отримати прямий доступ до всесвітньої мережі через HTTP (HTTPS) або веббраузер

Переваги вебскребкування



Швидкість: можливість обробити багато даних за короткий час



Автоматизація: можливість заощадити ручну роботу



Повторення: можливість повторно використовувати через регулярні проміжки часу, коли вебсайти оновлюються

Створення власних наборів даних за допомогою Google Таблиць



Запуск нової електронної таблиці



Пошук надійних даних



Усунення несправностей і повідомлення про помилки



Імпорт даних до Google Таблиць



Відображення даних

```
fx | =importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films", "table", 1)
```

	A	B	C	D	E	F	G
1	Rank	Peak	Title	Worldwide gross	Year	Reference(s)	
2	1		1 *Avatar*	\$2,787,965,087	2009	[# 1][# 2]	
3		2	1 *Titanic*	\$2,187,463,944	1997	[# 3][# 4]	
4		3	3 *Star Wars: The	\$2,068,223,624	2015	[# 5][# 6]	

Імпорт даних до Google Таблиць



Імпортує дані з таблиці або списку на сторінці HTML в електронну таблицю Google (формула **importHTML**)

потрібні три параметри:



URL-адреса



тип даних, які збираємо, таблиця чи список



число, що представляє позицію таблиці чи списку в кодї HTML

Синтаксис

```
IMPORTHTML(url, query, index)
```



`url` - URL-адреса сторінки для перевірки, включаючи протокол (наприклад, `http://`)



`Query` - або «список» або «таблиця», залежно від того, який тип структури містить потрібні дані



`Index` - індекс, починаючи з 1, який визначає, яку таблицю або список, як визначено в джерелі HTML, потрібно повернути

Імпорт даних до Google Таблиць



```
=importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films"; "table"; 1)
```

Rank	Peak	Title	Worldwide gross	Year	Reference(s)
1	1	*Avatar*	\$2,906,256,019	2009	[# 1][# 2]
2	2	*Avengers: Endgame*	\$2,797,501,328	2019	[# 3][# 4]
3	3	*Titanic*	\$2,187,535,296	1997	[# 5][# 6]
4	4	*Star Wars: The Force Awakens*	\$2,068,223,624	2015	[# 7][# 8]
5	5	*Avengers: Infinity War*	\$2,048,359,754	2018	[# 9][# 10]
6	6	*Spider-Man: No Way Home*	\$1,916,044,248	2021	[# 11][# 12]
7	7	*Jurassic World: Dominion*	\$1,671,537,444	2015	[# 13][# 14]

ERROR! -
переконайтеся,
що лапки є
подвійними
лапками

```
=importHTML("https://index.minfin.com.ua/ua/economy/index/inflation/";  
"table"; 1)
```

VALUE! -
переконайтеся, що
у клітинці немає
зайвих дужок або
лапок

Дані доступні для редагування



«Очищення даних» означає зробити їх доступними для роботи

google табл ☆ 📄 🌐

Файл Змінити Вигляд Вставити Формат Дані Інструменти Розширення Довідка Шойно змінено

100% | грн. % .00 .00 123 | За умовча... | 10 | B I U A | 📄 📊 📉 📈 📉 📈 📉 📈 📉 📈

С:Н | fx | =importHTML("https://en.wikipedia.org/wiki/List_of_highest-grossing_films"; "table"; 1)

Rank	Peak	Title	Worldwide gross	Year	Reference(s)
1	1	*Avatar*	\$2,906,256,019	2009	[# 1][# 2]
2	1	*Avengers: Endgame*	\$2,798,948,689	2019	[# 1][# 2]
3	1	*Titanic*	\$2,120,738,116	1997	[# 1][# 2]
4	3	*Star Wars: The Force Awakens*	\$2,069,495,043	2015	[# 1][# 2]
5	4	*Avengers: Infinity War*	\$2,048,360,679	2018	[# 1][# 2]
6	6	*Spider-Man: No Way Home*	\$1,921,843,057	2021	[# 1][# 2]
7	3	*Jurassic World*	\$1,671,713,244	2015	[# 1][# 2]
8	7	*The Lion King*	\$1,663,270,000	1994	[# 1][# 2]
9	3	*The Avengers*	\$1,518,815,014	2011	[# 1][# 2]
10	4	*Furious 7*	\$1,515,370,000	2015	[# 1][# 2]
11	11	*Top Gun: Maverick*	\$1,485,000,000	2022	[# 1][# 2]
12	10	*Frozen II*	\$1,453,551,000	2019	[# 1][# 2]
13	5	*Avengers: Age of Ultron*	\$1,449,100,000	2015	[# 1][# 2]
14	9	*Black Panther*	\$1,347,000,000	2018	[# 1][# 2]
15	3	*Harry Potter and the Chamber of Secrets*	\$1,342,000,000	2002	[# 1][# 2]
16	9	*Star Wars: The Force Awakens*	\$1,332,537,000	2015	[# 1][# 2]
17	12	*Jurassic World: Dominion*	\$1,330,000,000	2022	[# 1][# 2]
18	5	*Frozen*	\$1,290,000,000	2013	[# 1][# 2]
19	10	*Beauty and the Beast*	\$1,260,000,000	2017	[# 1][# 2]
20	15	*Incredibles 2*	\$1,242,000,000	2018	[# 1][# 2]
21	11	*The Fate of the Furious*	\$1,238,000,000	2017	[# 1][# 2]
22	5	*Iron Man 3*	\$1,214,000,000	2012	[# 1][# 2]
23	10	*Minions*	\$1,159,000,000	2015	[# 1][# 2]
24	12	*Captain America: Civil War*	\$1,153,000,000	2016	[# 1][# 2]
25	20	*Aquaman*	\$1,148,000,000	2018	[# 1][# 2]
26	2	*The Lord of the Rings: The Two Towers*	\$1,143,000,000	2002	[# 1][# 2]
27	24RK	*Spider-Man: Far From Home*	\$1,131,000,000	2019	[# 1][# 2]
28	23RK	*Captain Marvel*	\$1,113,000,000	2019	[# 1][# 2]
29	5RK	*Transformers: The Last Knight*	\$1,103,000,000	2017	[# 1][# 2]
30	7	*Skyfall*	\$1,103,000,000	2012	[# 1][# 2]

☒ Вирізати Ctrl+X
☒ Копіювати Ctrl+C
☒ Вставити Ctrl+V
☒ Спеціальна вставка
+ Вставити стовпців ліворуч: 6
+ Вставити стовпців праворуч: 6
🗑️ Видалити стовпці C – H
✖️ Очистити стовпці C – H
🔒 Приховати стовпці C – H
📏 Змінити розмір стовпців C–H
🔤 Сортувати аркуш у порядку Від "А" до "Я"
🔤 Сортувати аркуш у порядку Від "Я" до "А"
🔤 Умове форматування
🔤 Перевірка даних
🔤 Статистика за стовпцями
🔤 Конвертувати в чип "Люди"
⋮ Переглянути інші дії зі стовпцями

Лише значення Ctrl+Shift+V
Лише формат Ctrl+Alt+V
Лише формула
Лише умовне форматування
Лише перевірка даних
Як текст
CSV як стовпці
Транспозиція
Лише по ширині стовпця
Усе, крім меж



Виділити вихідну таблицю



«Спеціальна вставка» >
«Лише значення»



Закріпити рядок із назвами стовпців > Наведіть курсор миші на рядок трохи вище рядка 1 над сірою смугою (курсор перетворився на руку) > Перетягніть панель вниз рядка 1 і залиште її там > Верхній рядок закріплено

Очищення даних



Пакетне редагування за допомогою функції пошуку та заміни

The screenshot displays a Google Sheets interface with a data table and the 'Find and Replace' dialog box open. The data table contains the following information:

	E	F	G	H	
	Title	Worldwide gross	Year	Reference(s)	
1	*Avatar*	\$2,906,666,303	2009	[# 1][# 2]	\$2,906,666,303
1	*Avengers: Endg	\$2,797,501,328	2019	[# 3][# 4]	\$2,797,501,328
1	*Titanic*	\$2,187,535,296	1997	[# 5][# 6]	\$2,187,535,296
3	*Star Wars: The	\$2,068,223,624	2015	[# 7][# 8]	\$2,068,223,624
4	*Avengers: Infini	\$2,048,359,754	2018	[# 9][# 10]	\$2,048,359,754
6	*Spider-Man: No	\$1,916,044,248	2021	[# 11][# 12]	\$1,916,044,248
3	*Jurassic World*	\$1,671,537,444	2015	[# 13][# 14]	\$1,671,537,444
7	*The Lion King*	\$1,656,943,394	2019	[# 15][# 4]	\$1,656,943,394
3	*The Avengers*	\$1,518,812,988	2012	[# 16][# 17]	\$1,518,812,988
4	*Furious 7*	\$1,516,045,911	2015	[# 18][# 19]	\$1,516,045,911
11	*Top Gun: Mave	\$1,476,420,739	2022	[# 20]	\$1,476,420,739
10	*Frozen II*	\$1,450,026,933	2019	[# 21][# 22]	\$1,450,026,933
5	*Avengers: Age	\$1,402,809,540	2015	[# 23][# 19]	\$1,402,809,540
9	*Black Panther*	\$1,347,280,838	2018	[# 24][# 25]	\$1,347,280,838

The 'Find and Replace' dialog box is open, showing the following settings:

- Find: Av
- Replace: Avatar
- Search: Певний діапазон (Selected)
- Search range: 'Аркуш1':C:H
- Match case: (Checked)
- Match entire contents of cells: (Unchecked)
- Search using wildcards: (Unchecked)
- Also search in formulas: (Unchecked)
- Also search in links: (Unchecked)

Buttons at the bottom of the dialog: Знайти, Замінити, Замінити всі, Готово.

Вибір інструментів

Фактори, які необхідно враховувати при виборі інструментів



Якість даних



Доставка даних



Масштабованість



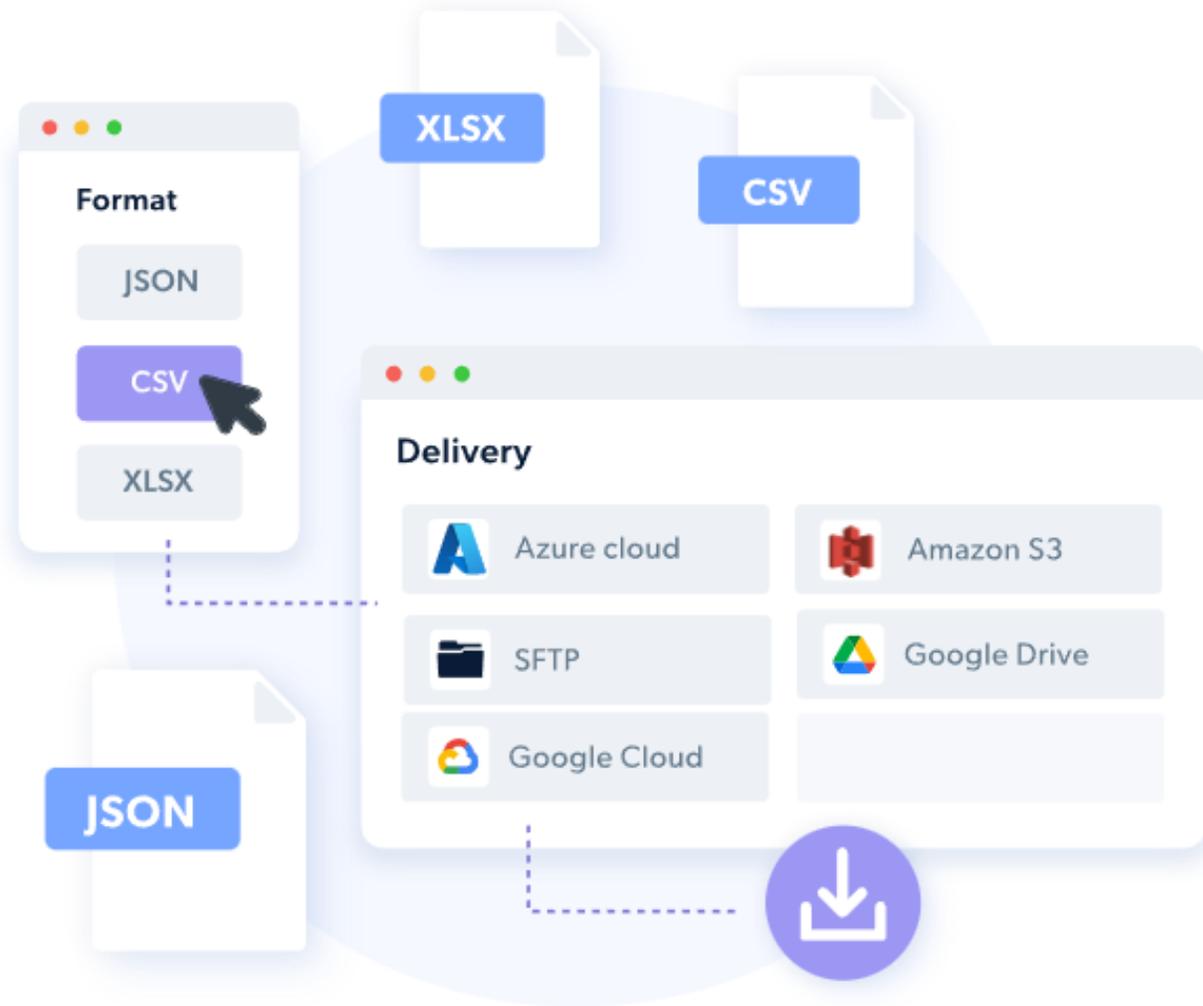
Ціна



Підтримка клієнтів

Bright Data

інструмент для автоматичного сканування будь-якого вебсайту

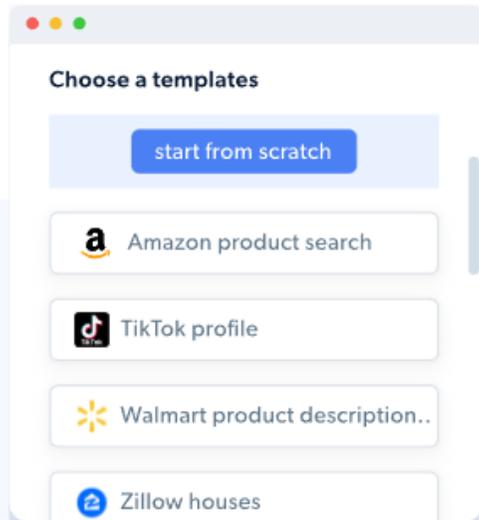


<https://brightdata.com/>

Як зробити вебскребок

КРОК 1

Виберіть із готових шаблонів коду або почніть з нуля



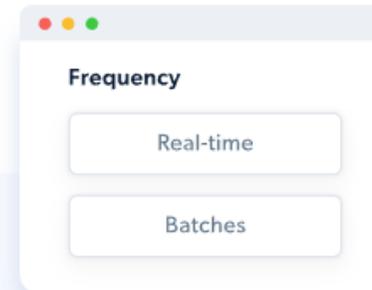
КРОК 2

Розробіть і налаштуйте свій скребок за допомогою готових функцій скрапінгу Bright Data



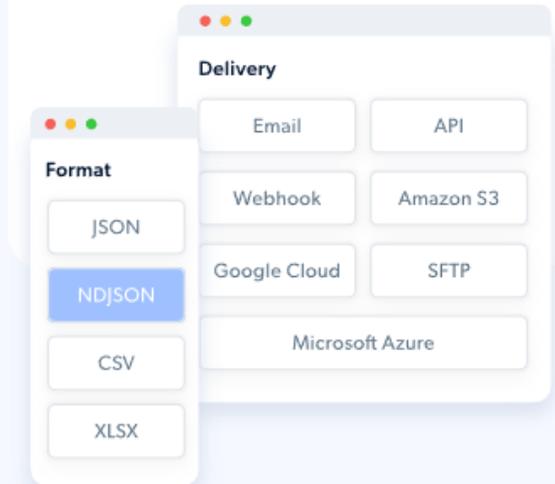
КРОК 3

Виберіть, коли отримувати дані: у режимі реального часу чи пакетами



КРОК 4

Виберіть формат файлу та куди надсилати дані



ProWebScraper



Отримати дані з динамічних вебсайтів
(витягувати дані із сайту за допомогою кількох рівнів навігації - будь то категорії, підкатегорії, розбивка на сторінки чи сторінки продуктів)



Витягувати будь-що з вебсторінок
(текст, посилання, дані таблиці або високоякісні зображення тощо)



<https://prowebscraper.com/>

Web Scraper



вилучення вебданих простим і доступним способом



доступний для браузера Google Chrome



можна автоматизувати процес вилучення, запланувавши його

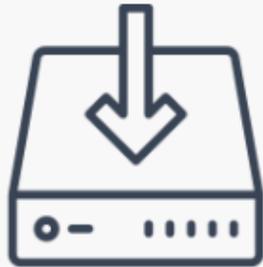


експорт даних у форматах CSV, XLSX і JSON

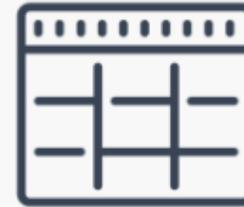


<https://webscraper.io/>

Як зробити вебскребок в Web Scraper



Встановіть **Web Scraper**



Скребіть свій перший сайт



Розширення Chrome



Надбудова Firefox

Основні кроки скребування

1



Вставка

Введіть URL-адресу веб-сайту, з якого ви хочете отримати дані.

2



Налаштувати

Налаштуйте скребок за допомогою елементів «точка-на-ціль».

3



Завантажити

Доступ до даних через JSON і API за лічені секунди.

Прості інструменти для вебскрейпінгу

Розширення для Google Chrome



Email Extractor

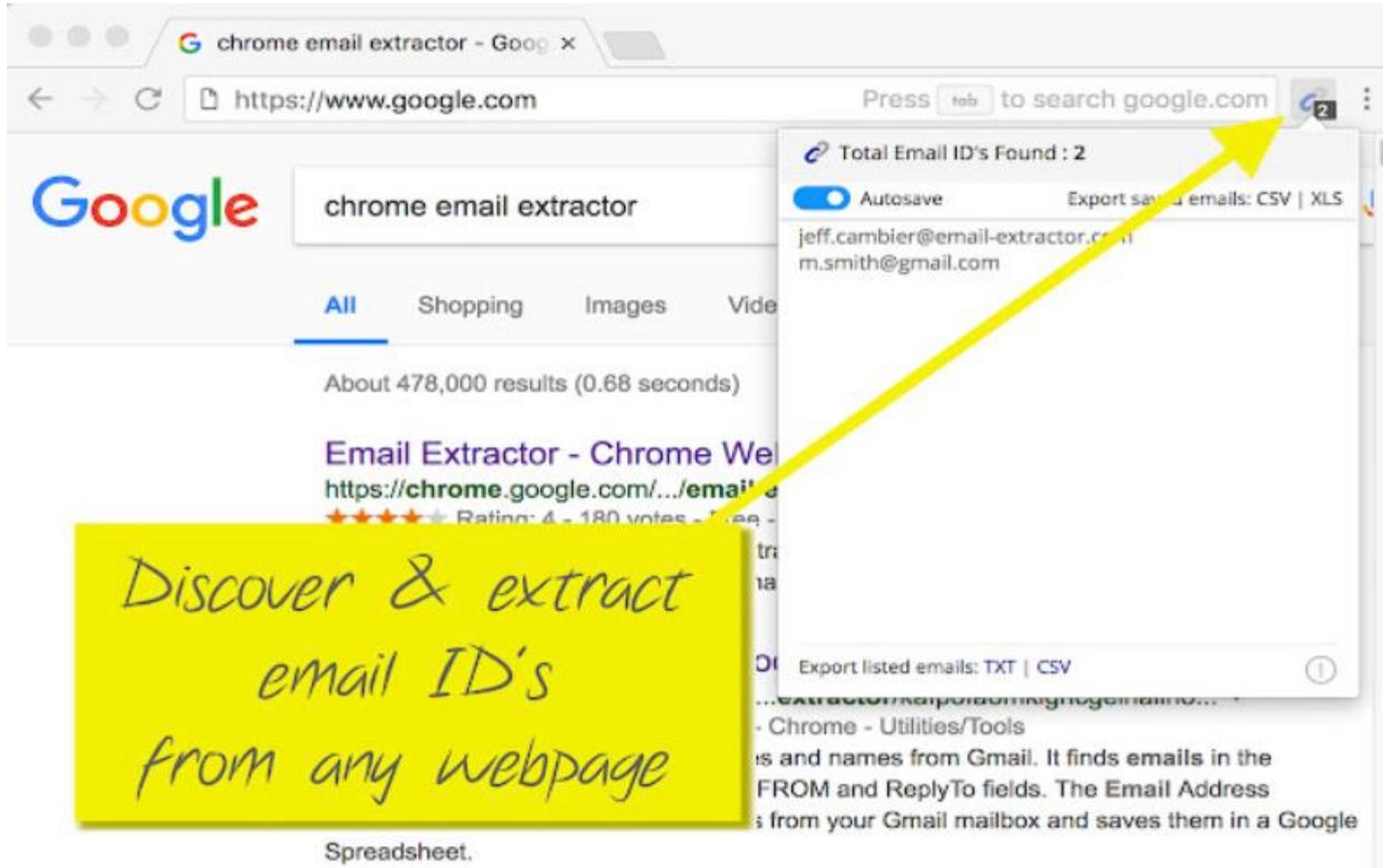


Table Capture



Instant Data Scraper

Email Extractor



chrome email extractor - Google x

https://www.google.com

Press **tab** to search google.com

Google

chrome email extractor

All Shopping Images Video

About 478,000 results (0.68 seconds)

Email Extractor - Chrome Web Store

https://chrome.google.com/.../email-extractor

★★★★★ Rating: 4 - 180 votes

Total Email ID's Found : 2

Autosave Export saved emails: CSV | XLS

jeff.cambier@email-extractor.com

m.smith@gmail.com

Export listed emails: TXT | CSV

Chrome - Utilities/Tools

is and names from Gmail. It finds emails in the FROM and ReplyTo fields. The Email Address is from your Gmail mailbox and saves them in a Google Spreadsheet.

Discover & extract email ID's from any webpage



<http://surl.li/dnmxy>

розширення для вилучення електронної пошти

Table Capture

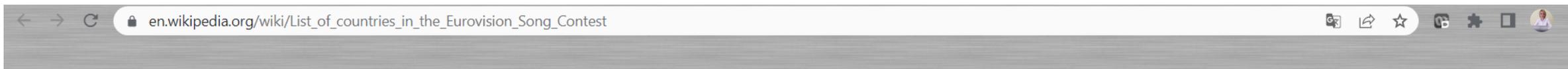
Bulgaria †	RNT	2005	2022	14	5
Croatia			2023	27	18
Cyprus			2023	38	31
Czech Republic			2023	10	4
Denmark			2023	50	44
Estonia			2023	27	17
Finland			2023	55	47
France			2023	64	64
Georgia					
Germany	BR (1979–1991) (ARD) MDR (1992–1995) (ARD)	1956	2023	65	65



копіює таблиці HTML у буфер обміну або експортує їх у Microsoft Excel, CSV, Google Sheets, Microsoft 365 тощо

<http://surl.li/dnxee>

Table Capture



 Bulgaria †	BNT	2005	2022	14	5	5/14	2021	0	N/A
 Croatia	HRT	1993	2023	27	18	6/15	2017	0	N/A
 Cyprus	CyBC	1981	2023	38	31	9/16	2021	0	N/A
 Czech Republic	ČT	2007	2023	10	4	4/10	2022	0	N/A
 Denmark	DR	1957	2023	50	44	10/16	2019	3	2013
 Estonia	ERR	1994	2023	27	17	8/18	2022	1	2001
 Finland	Yle	1961	2023	55	47	8/16	2022	1	2006
 France	RTF (1956–1964) ORTF (1965–1974) TF1 (1975–1981) Antenne 2 (1983–1992)	1956	2023	64	64	N/A	2022	5	1977



Options

Ignore image + icon attributes

Delete empty rows

Ignore hidden elements

Paged & Dynamic Tables



Try **Table Capture Pro** to be able to capture multi-page tables ([demo](#)) and tables whose content changes dynamically ([demo](#)). Also, try **Table Capture Cloud (Beta)** for a bunch of next-level features.

Preview Data · Rows: 53, Columns: 10

Country	Broadcaster(s) [12]	Debut year	Most recent entry	...
Albania	RTSH	2004	2023	...
Andorra †	RTVA	2004	2009	...
...				
Yugoslavia [e] ‡	JRT	1961	1992	...



Table Capture

особливості:



копіювання таблиці в буфер обміну з належними роздільниками рядків і стовпців



експорт в Google Таблиці



пакетний експорт таблиць у буфер обміну, Excel і Google Таблиці



захоплення таблиць `<div>` (або будь-яких повторюваних елементів на вебсайті)

Instant data scraper



автоматично знаходить і витягує дані з вебсторінок



експортує файли Excel або CSV



працює у браузері Chrome або Edge як розширення



<http://surl.li/bkntm>

Розширені інструменти для вебскрейпінгу



Data Miner Pro

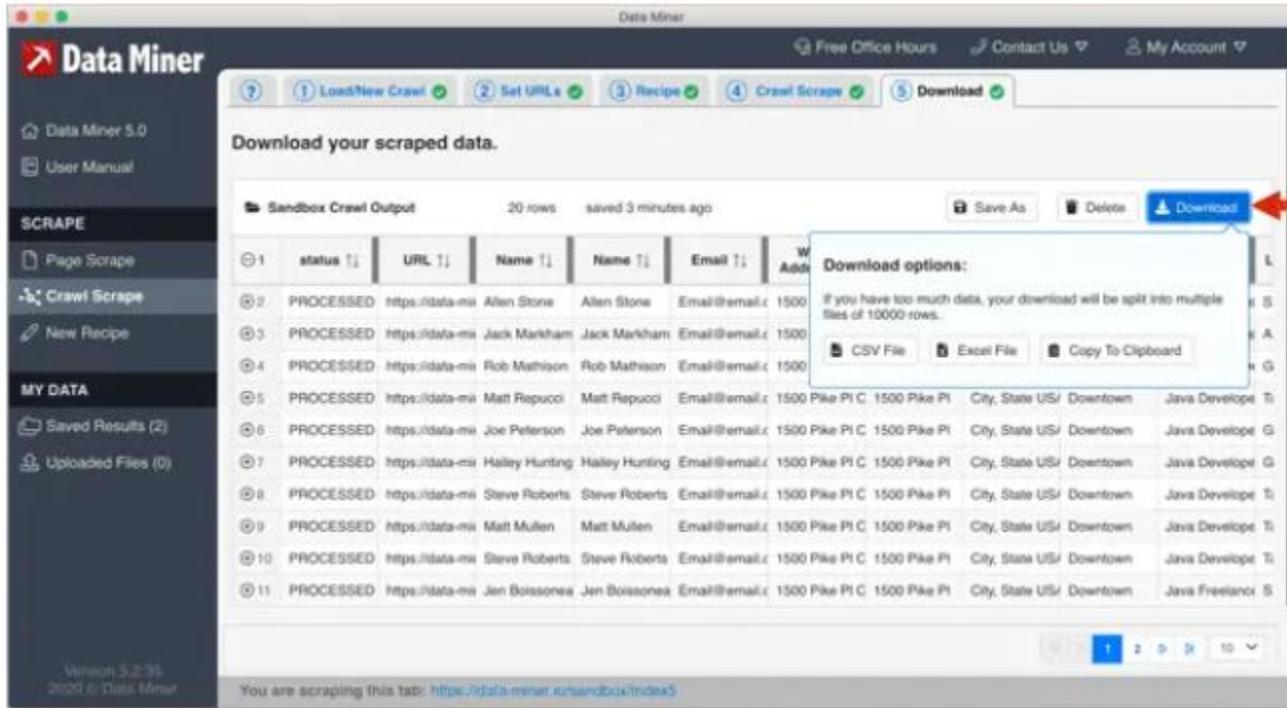


OUTWIT
TECHNOLOGIES



PhantomBuster

Dataminer



The screenshot shows the Data Miner web interface. At the top, there's a navigation bar with "Data Miner" logo, "Free Office Hours", "Contact Us", and "My Account". Below that, a progress bar shows five steps: "Load New Crawl", "Set URLs", "Recipe", "Crawl Scrape", and "Download". The main content area is titled "Download your scraped data." and shows a table of data from a "Sandbox Crawl Output". The table has columns for status, URL, Name, Email, and other details. A "Download" button is highlighted with a red circle, and a tooltip menu is open over it, showing options: "CSV File", "Excel File", and "Copy To Clipboard".

status	URL	Name	Name	Email	W Add
PROCESSED	https://data-mi	Allen Stone	Allen Stone	Email@email.c	1500
PROCESSED	https://data-mi	Jack Markham	Jack Markham	Email@email.c	1500
PROCESSED	https://data-mi	Rob Mathison	Rob Mathison	Email@email.c	1500
PROCESSED	https://data-mi	Matt Repucci	Matt Repucci	Email@email.c	1500 Pike Pl C
PROCESSED	https://data-mi	Joe Peterson	Joe Peterson	Email@email.c	1500 Pike Pl C
PROCESSED	https://data-mi	Halley Hunting	Halley Hunting	Email@email.c	1500 Pike Pl C
PROCESSED	https://data-mi	Steve Roberts	Steve Roberts	Email@email.c	1500 Pike Pl C
PROCESSED	https://data-mi	Matt Mullen	Matt Mullen	Email@email.c	1500 Pike Pl C
PROCESSED	https://data-mi	Steve Roberts	Steve Roberts	Email@email.c	1500 Pike Pl C
PROCESSED	https://data-mi	Jen Boissonea	Jen Boissonea	Email@email.c	1500 Pike Pl C



перетворює HTML-сторінки у файли CSV або Excel

<http://surl.li/dnzem>

Dataminer

The image shows a screenshot of the Dataminer web interface with several annotations. At the top right, there are three navigation links: "Recipe Creator" (with a pencil icon), "Privacy" (with a shield icon), and "Account" (with a person icon). Below these is the "Data Miner" logo and a "Start Scraping" button. Two green checkmarks indicate site compatibility: "Data Miner can scrape this site." and "You can [make your own recipe](#) for this page." Below this is a "Pinned Recipes" section with two items: "Email Scraper" (12 results) and "Person Contact Scraper" (4 results), each with a "Scrape" button. At the bottom, it says "Logged in as [dataminervideo@gmail.com](#)".

Recipe Creator
Privacy
Account

Data Miner

Site Compatibility → ✓ Data Miner can scrape this site. **Start Scraping** ← Launch Scraper
✓ You can [make your own recipe](#) for this page.

Quick Scrape →

Pinned Recipes

Email Scraper	12	Scrape
Person Contact Scraper	4	Scrape

Logged in as [dataminervideo@gmail.com](#) ← Account Email

Outwit Hub

The screenshot displays the Outwit Hub application window. The top menu bar includes 'OutWit Hub', 'File', 'Edit', 'View', 'Navigation', 'Tools', 'Help', and 'Registration'. The address bar shows the URL 'https://www.newyorker.com/magazine'. The main content area features the 'THE NEW YORKER' logo and a navigation menu with categories like News, Culture, Books, Business & Tech, Humor, Cartoons, Magazine, Video, Podcasts, Archive, and Goings On. A prominent banner for 'LET'S MAKE PAID PATERNITY LEAVE THE NEW STANDARD' is visible. Below the banner, the text 'THE MAGAZINE' is displayed. The bottom section of the window shows a table of RSS feed items with columns for ID, Source Uri, Feed Title, Title, Article Uri, Date, and Ab. The table contains 13 rows of data, including articles from June 10 & 17, 2019, and June 6, 2019. The interface also includes a sidebar on the left with various tool options and a bottom status bar with version information and search controls.

Local IP:	5.62.63.183	Remote IP:	151.101.16.239			
ID	Source Uri	Feed Title	Title	Article Uri	Date	Ab
814	http://www.newyorker.com/feed/magazine/rss	The New Yorker June 10 & 17, 2019 Issue	"Big Little Lies" Season 2, Reviewed: Meryl, M...	https://www.newyorker.com/culture/on-televi...	09/06/2019 11:00:00	Doree
815	http://www.newyorker.com/feed/magazine/rss	The New Yorker June 10 & 17, 2019 Issue	Sunday Reading: Pride and the Fiftieth Anniver...	https://www.newyorker.com/books/double-tak...	09/06/2019 09:00:00	From
816	http://www.newyorker.com/feed/magazine/rss	The New Yorker June 10 & 17, 2019 Issue	More People Should Know the Name of Ashle...	https://www.newyorker.com/sports/sporting-s...	08/06/2019 22:45:00	Lou
817						
818	http://www.newyorker.com/feed/news	News, Politics, Opinion, Commentary, and Anal...	A D Day Journey in the Spirit of A. J. Liebling	https://www.newyorker.com/news/letter-from-...	07/06/2019 15:22:12	Ref
819	http://www.newyorker.com/feed/news	News, Politics, Opinion, Commentary, and Anal...	Ross Douthat on the Crisis of the Conservative...	https://www.newyorker.com/news/q-and-a/ros...	07/06/2019 09:00:00	Isaac
820	http://www.newyorker.com/feed/news	News, Politics, Opinion, Commentary, and Anal...	A Gobshite American President in Ireland	https://www.newyorker.com/news/our-columni...	06/06/2019 18:20:07	John C
821	http://www.newyorker.com/feed/news	News, Politics, Opinion, Commentary, and Anal...	Why Famous, Powerful Presidential Candidate...	https://www.newyorker.com/news/news-desk/...	10/06/2019 09:00:00	Eric Le
822	http://www.newyorker.com/feed/news	News, Politics, Opinion, Commentary, and Anal...	Ta-Nehisi Coates Revisits the Case for Repara...	https://www.newyorker.com/news/the-new-yo...	10/06/2019 09:00:00	In an it
823	http://www.newyorker.com/feed/news	News, Politics, Opinion, Commentary, and Anal...	A Weakening Economy May Be the Biggest Thr...	https://www.newyorker.com/news/our-columni...	08/06/2019 18:45:00	John C
R74	http://www.newyorker.com/feed/news	News, Politics, Opinion, Commentary, and Anal...	"1984" at Seventy: Why We Still Read Orwell's...	https://www.newyorker.com/news/daily-comm...	08/06/2019 09:00:00	Iouis t



<https://www.outwit.com/>

Outwit Hub



Конструктор пошукових запитів дозволяє генерувати, редагувати та надсилати пошукові URL-адреси за кількома критеріями для найбільш використовуваних пошукових систем



Панель генерації рядків - це редактор рядків, за допомогою якого ви можете генерувати, редагувати URL-адреси чи будь-які інші рядки рядків, використовуючи простий синтаксис для визначення діапазонів цифр або літер, списків значень тощо



Перегляд новин - вивчає поточну сторінку та відповідні посилання на ній, щоб знайти канали RSS і відобразити їх як записи в таблиці даних подання

PhantomBuster



доступне розширення для Google Chrome



можливість перетворити найпопулярніші вебсайти у базу даних



можливість знайти контакти, компанії тощо у своїй ніші



<https://phantombuster.com/>

Скрейпінг з Python

Що потрібно знати?



базові знання мови програмування Python



встановити Anaconda



редактор кода Jupyter Notebook або інший



завантажити або оновити до самої останньої версії
браузера Chrome

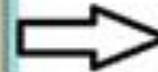
Вебскрейпінг



сторінка сайту



вебскрейпінг



візуалізація
структурованих даних

КОРИСНІ РЕСУРСИ



[Скрейпимо публічні дані, або Як я робив мапу АЗС](#)



[Що таке веб-скрейпінг і як він пов'язаний з проксі](#)



[Інструменти збирання даних](#)