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
- Страхування

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

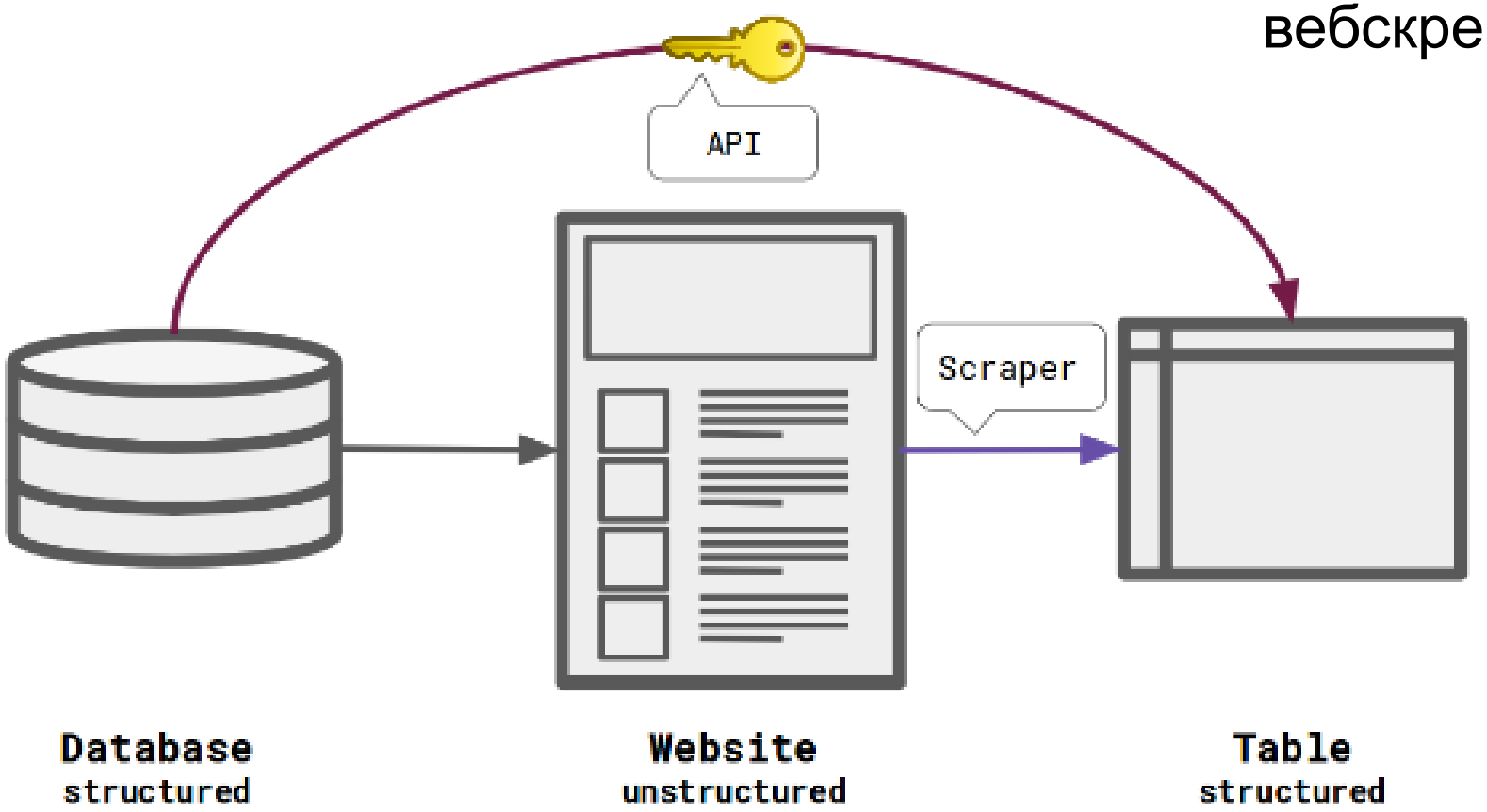
онлайн-методи вебскрейпінгу не потребують обслуговування

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
- Імпорт даних до Google Таблиць
- Усунення несправностей і повідомлення про помилки
- Відображення даних

```
=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 коді HTML

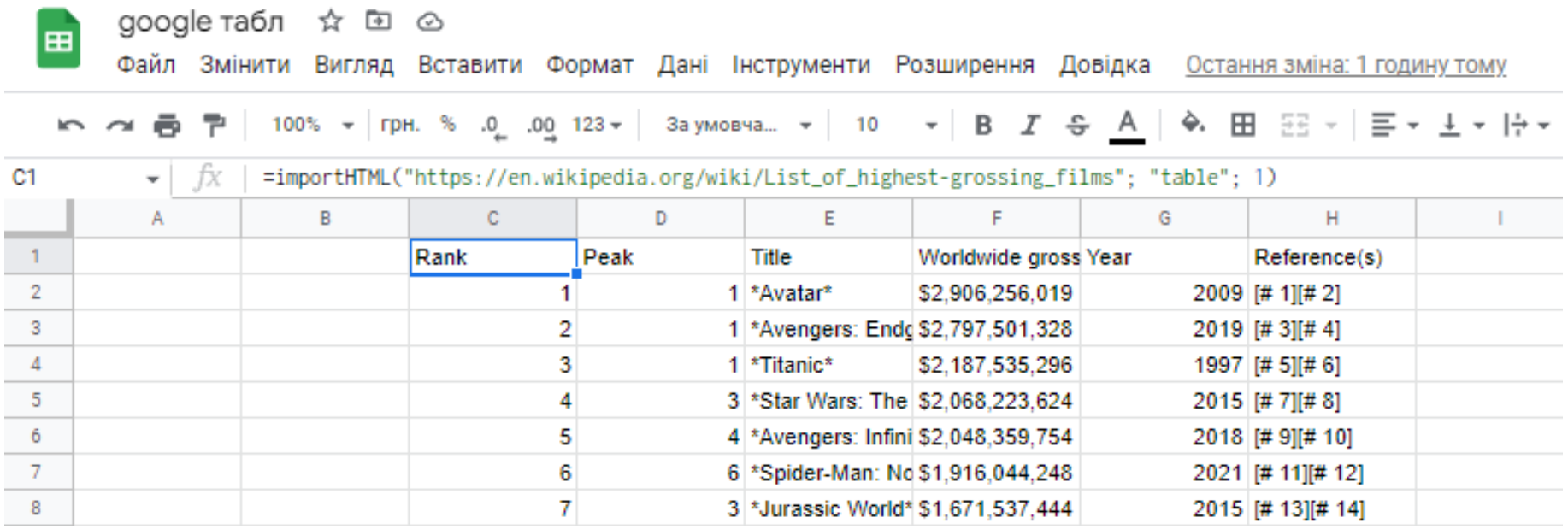
Синтаксис

`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)



	A	B	C	D	E	F	G	H	I
1			Rank	Peak	Title	Worldwide gross	Year	Reference(s)	
2			1	1	*Avatar*	$2,906,256,019	2009	[# 1][# 2]	
3			2	1	*Avengers: Endg	$2,797,501,328	2019	[# 3][# 4]	
4			3	1	*Titanic*	$2,187,535,296	1997	[# 5][# 6]	
5			4	3	*Star Wars: The	$2,068,223,624	2015	[# 7][# 8]	
6			5	4	*Avengers: Infini	$2,048,359,754	2018	[# 9][# 10]	
7			6	6	*Spider-Man: No	$1,916,044,248	2021	[# 11][# 12]	
8			7	3	*Jurassic World*	$1,671,537,444	2015	[# 13][# 14]	

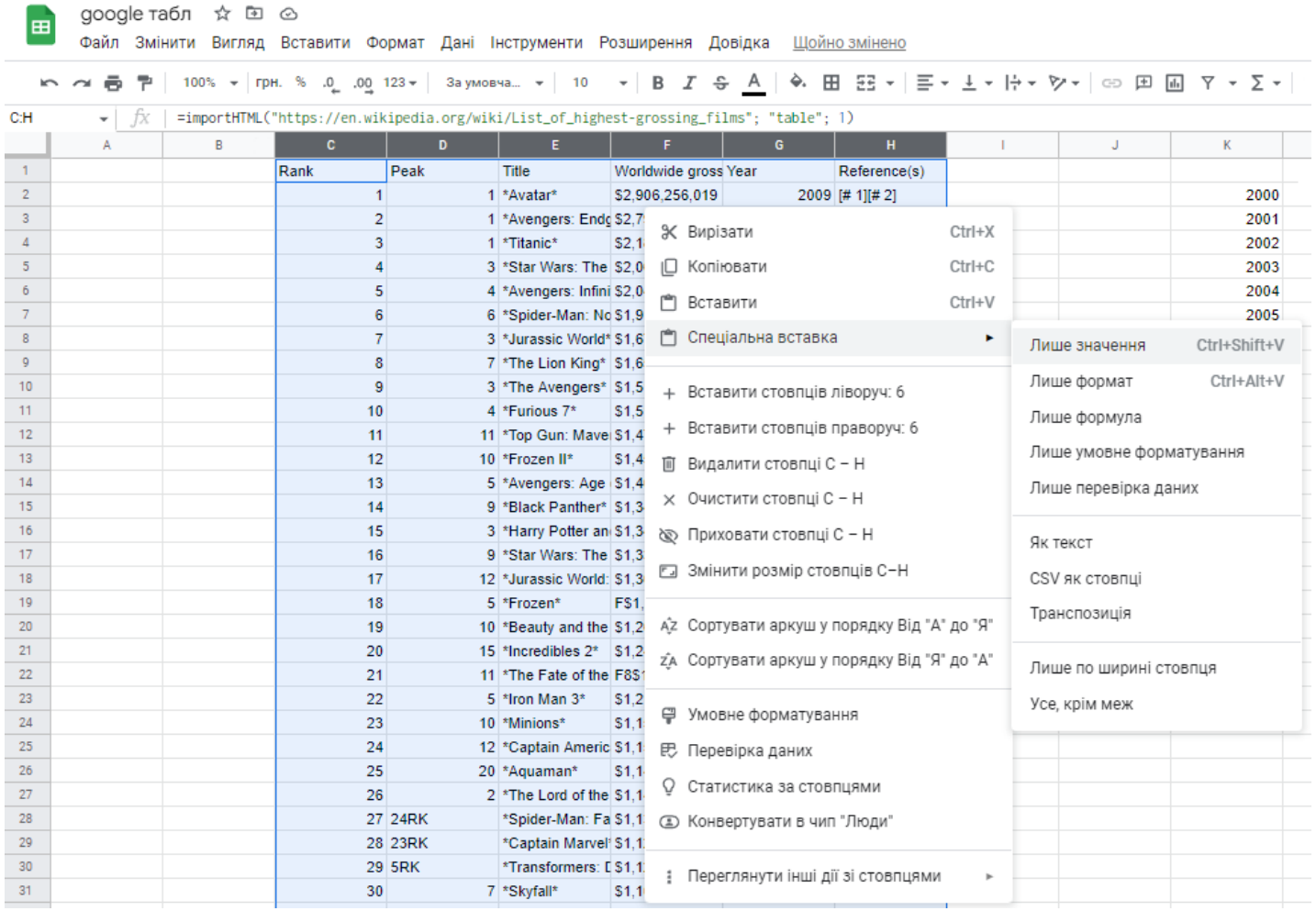
=importHTML("https://index.minfin.com.ua/ua/economy/index/inflation/"; "table"; 1)

ERROR! - переконайтеся, що лапки є подвійними лапками

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

Дані доступні для редагування

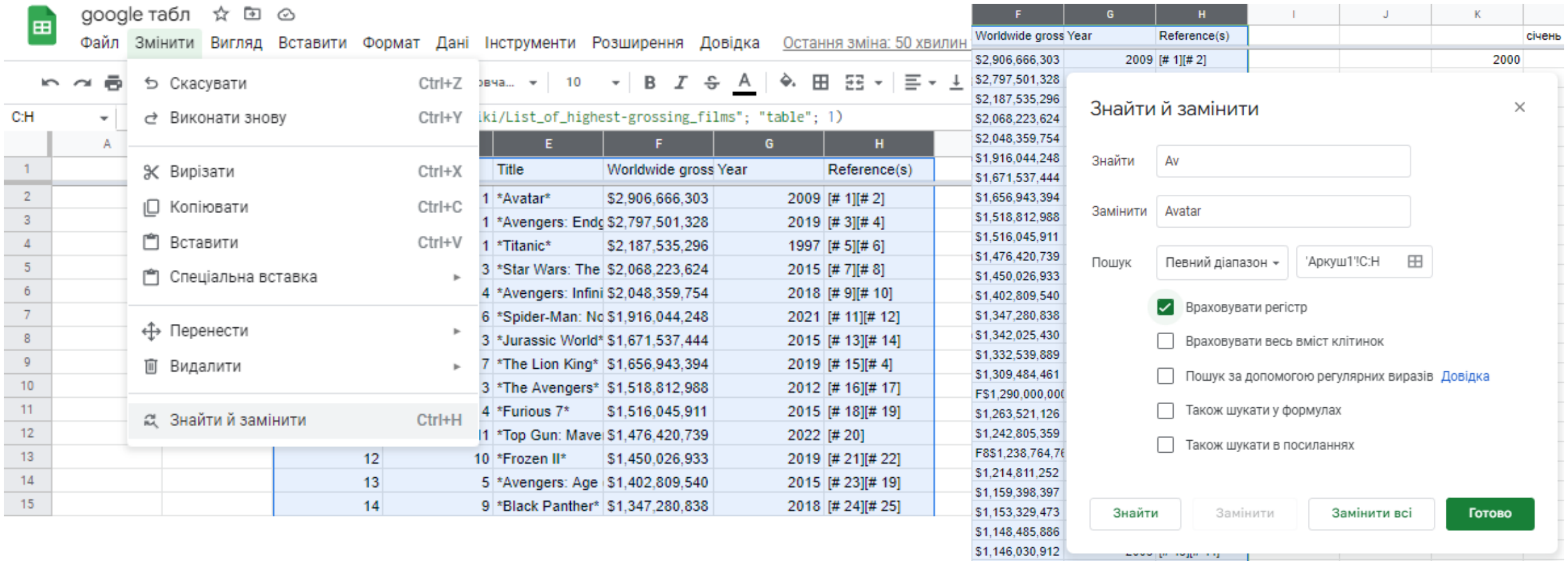
«Очищення даних» означає зробити їх доступними для роботи



- 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



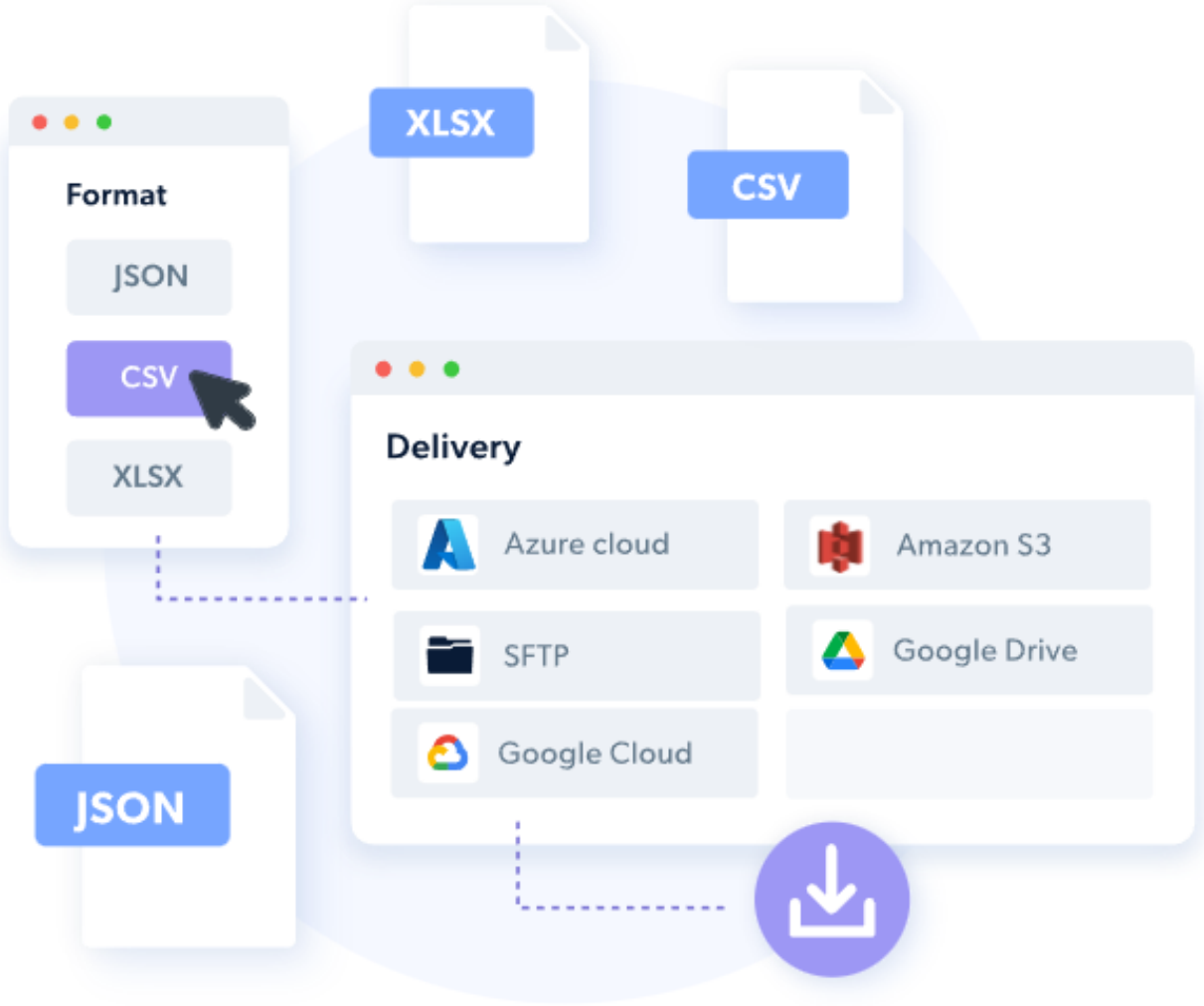
Вибір інструментів

Фактори, які необхідно враховувати при виборі інструментів

- Якість даних
- Доставка даних
- Масштабованість
- Ціна
- Підтримка клієнтів

Bright Data

інструмент для автоматичного сканування будь-якого вебсайту

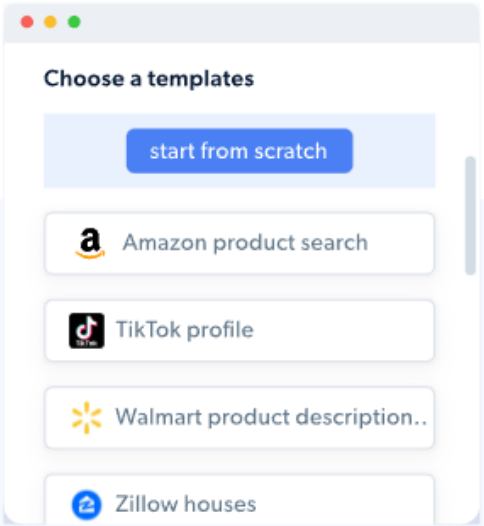


https://brightdata.com/

Як зробити вебскребок

КРОК 1

Виберіть із готових шаблонів коду або почніть з нуля

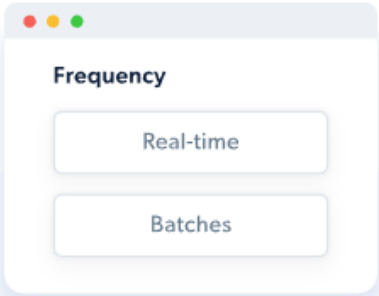


КРОК 2

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

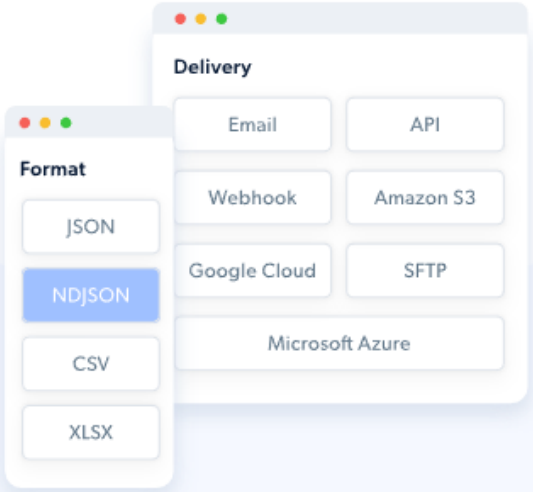
КРОК 3

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



КРОК 4

Виберіть формат файлу та куди надсилати дані



https://brightdata.com/products/data-collector

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

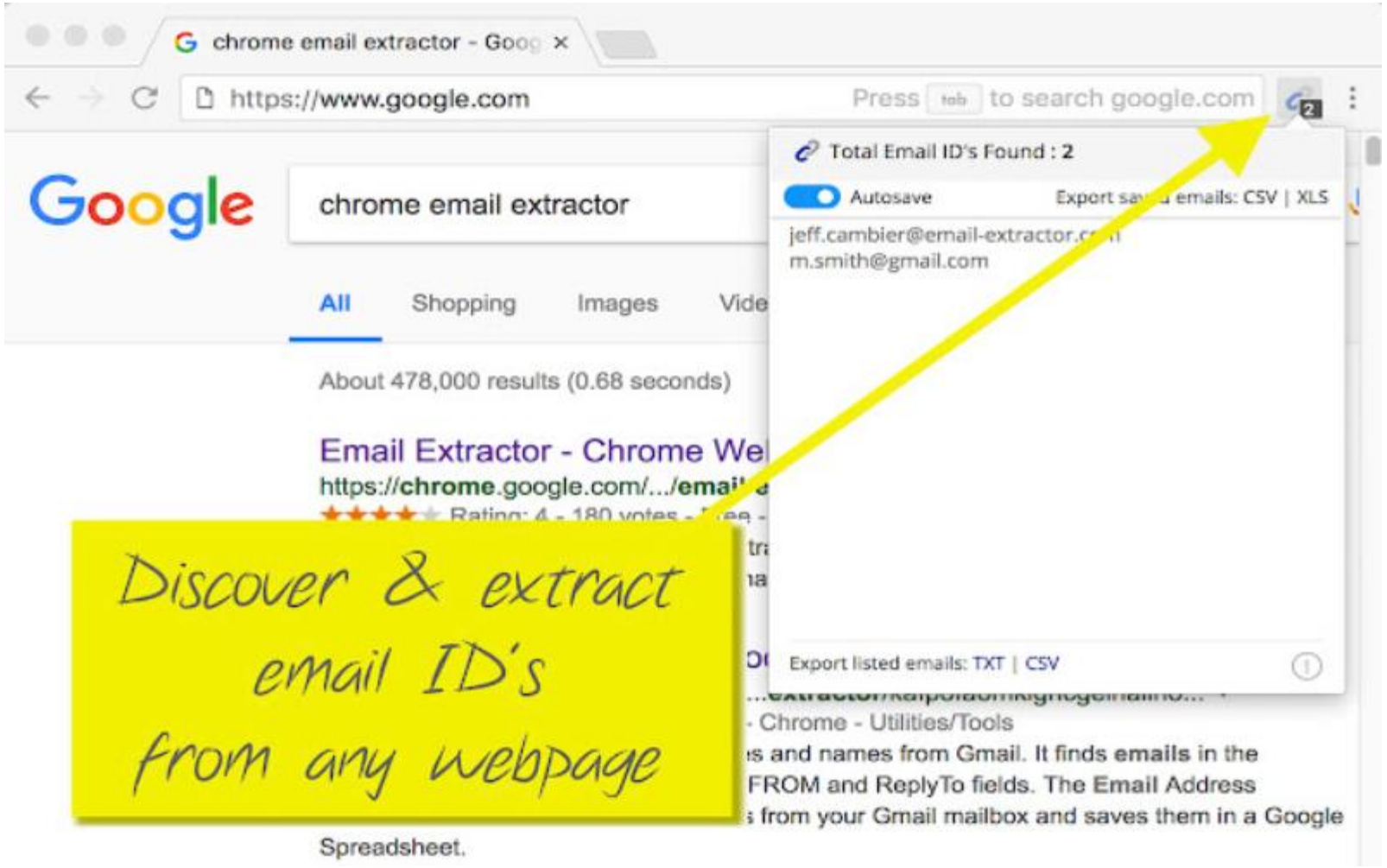
Розширення для Google Chrome

Email Extractor

Table Capture

Instant Data Scraper

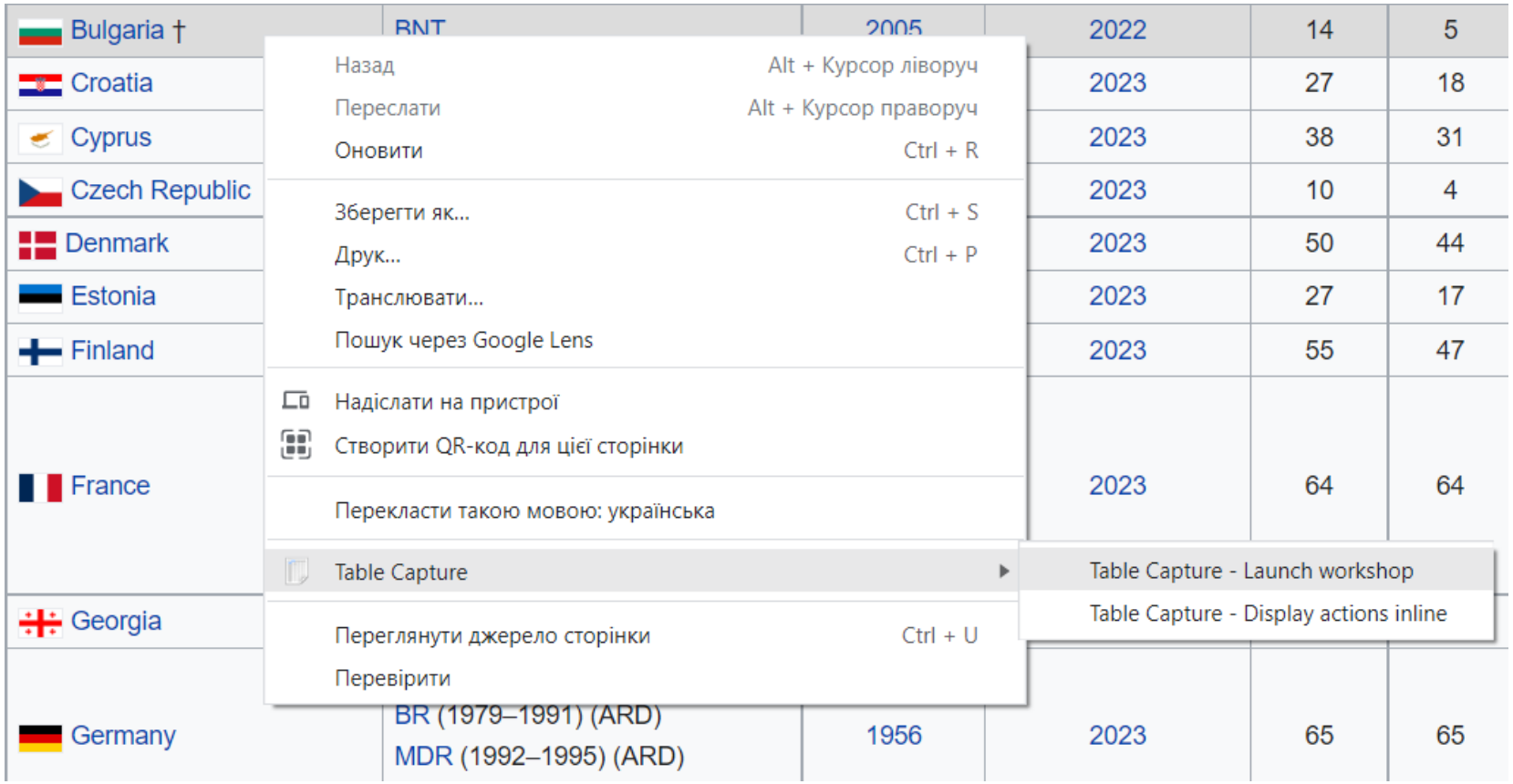
Email Extractor



http://surl.li/dnmxy

розширення для вилучення електронної пошти

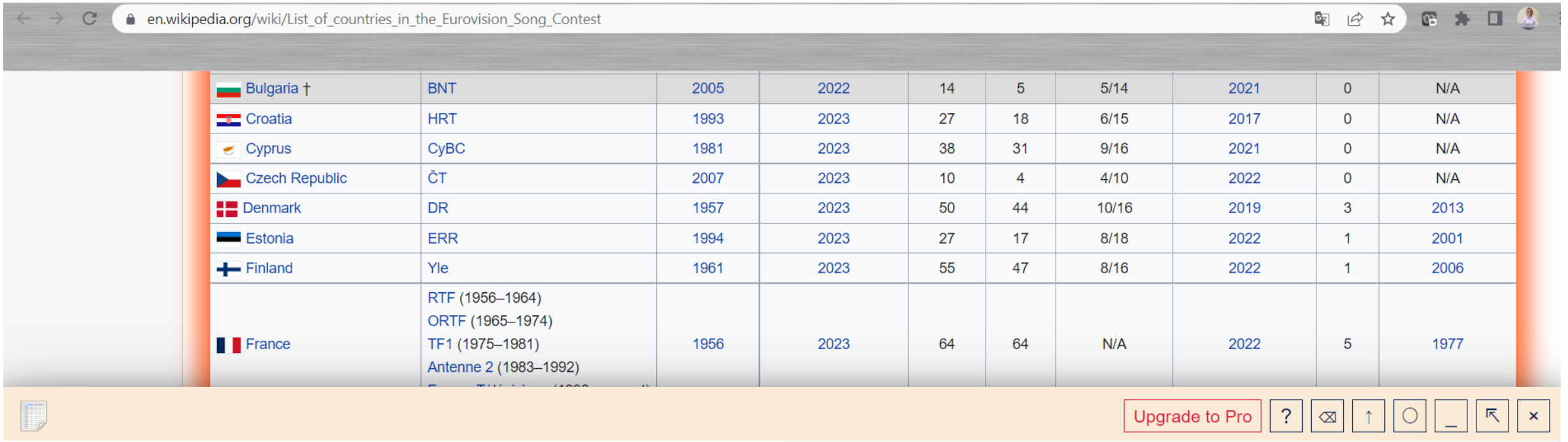
Table Capture



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

http://surl.li/dnxee

Table Capture



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	...

CSV

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 (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

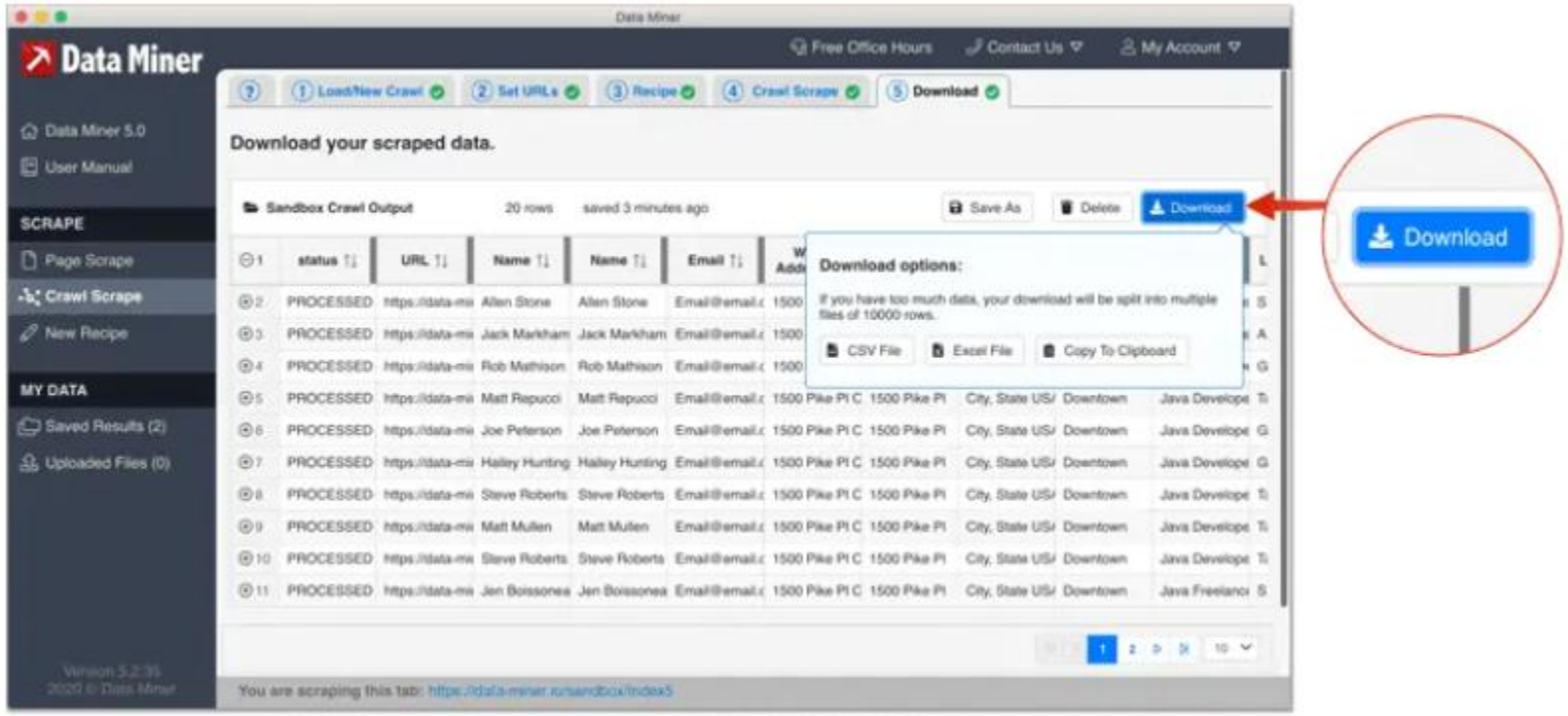
Розширені інструменти для вебскрейпінгу

Data Miner Pro

OUTWIT TECHNOLOGIES

PhantomBuster

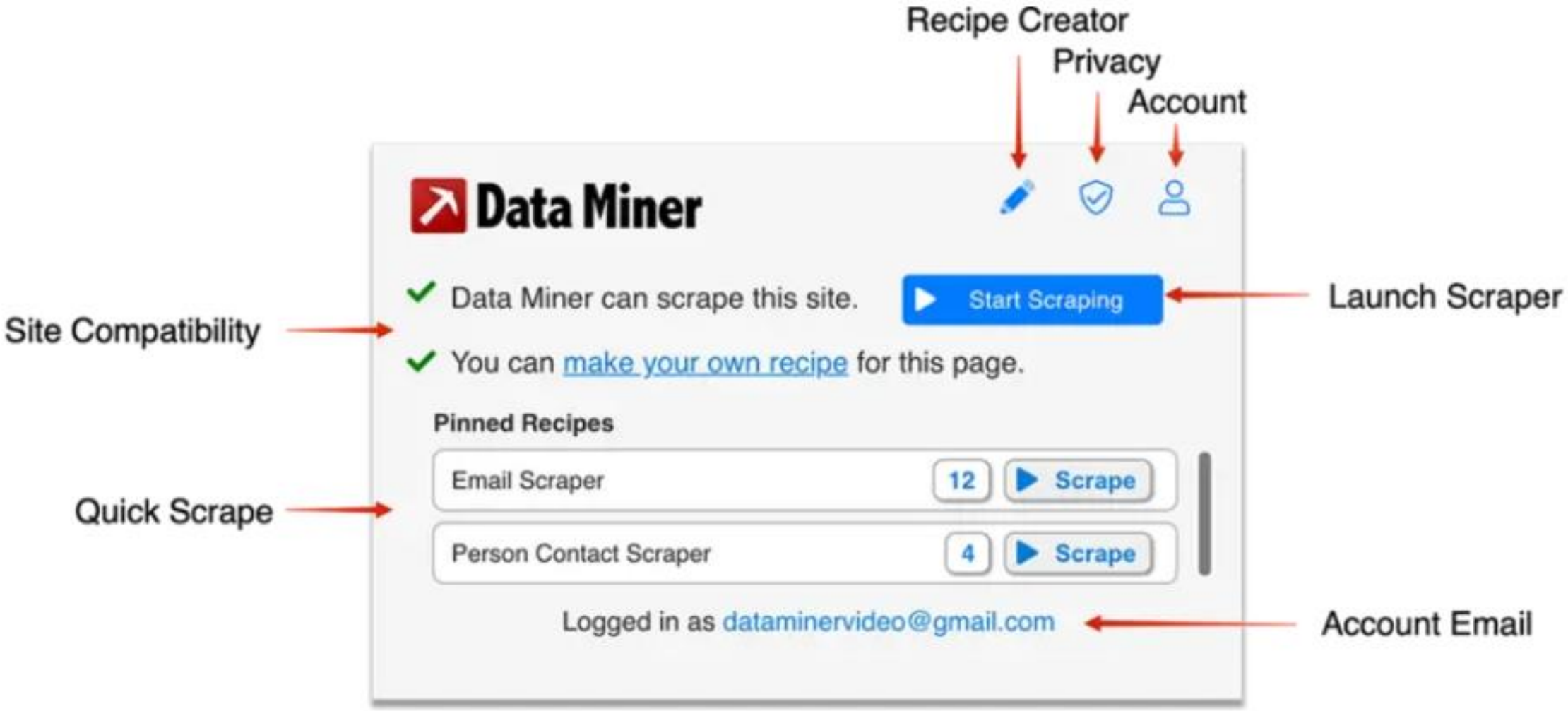
Dataminer



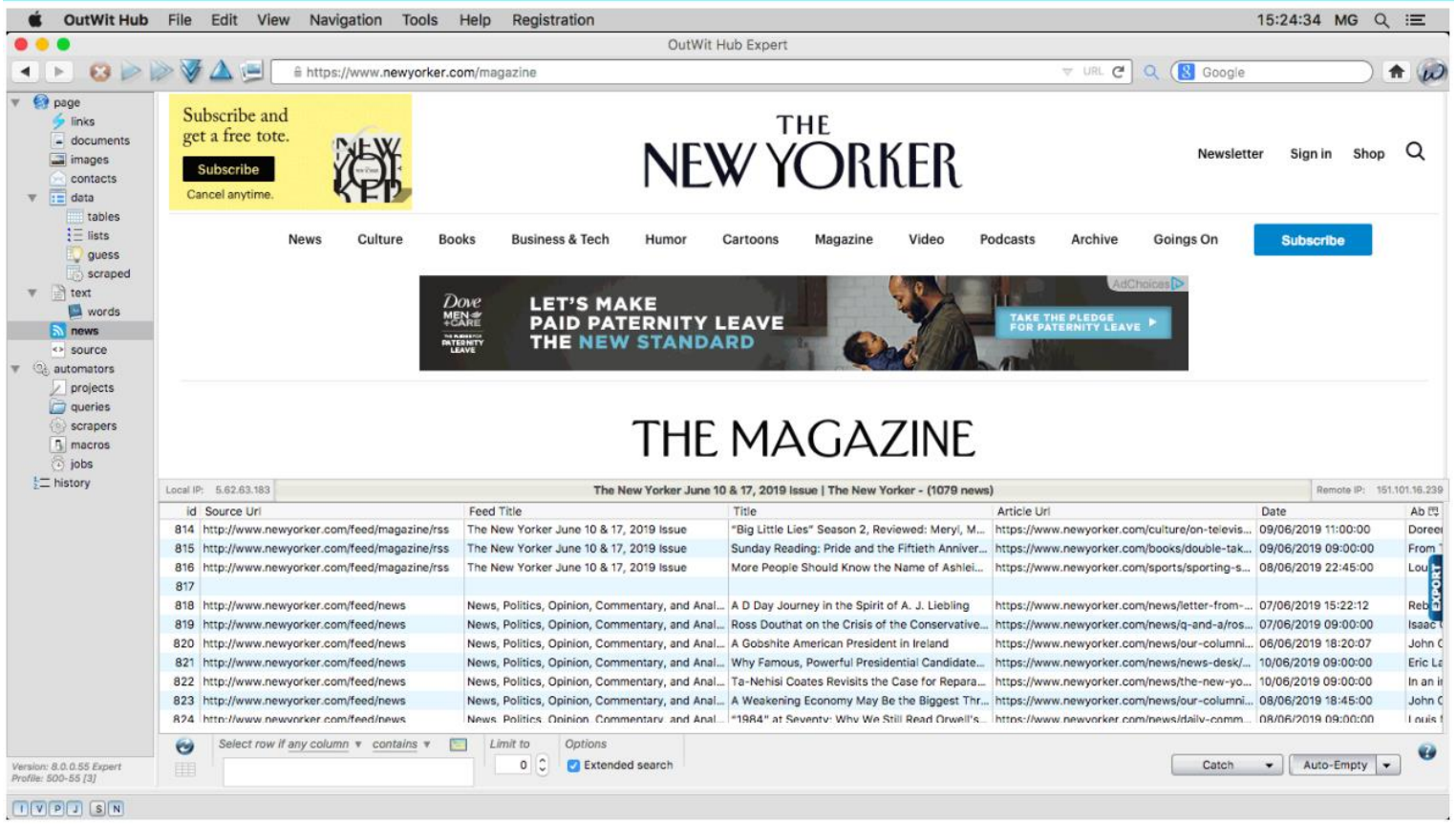
перетворює HTML-сторінки у файли CSV або Excel

http://surl.li/dnzem

Dataminer


Recipe Creator
Privacy
Account
Data Miner
Data Miner can scrape this site.
Start Scraping
Launch Scraper
Site Compatibility
You can make your own recipe for this page.
Pinned Recipes
Email Scraper
12
Scrape
Quick Scrape
Person Contact Scraper
4
Scrape
Logged in as dataminervideo@gmail.com
Account Email

Outwit Hub



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