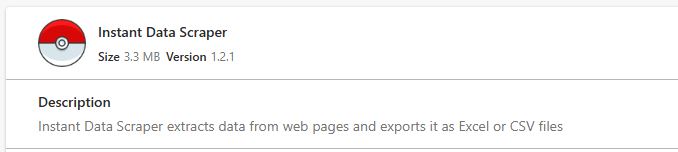
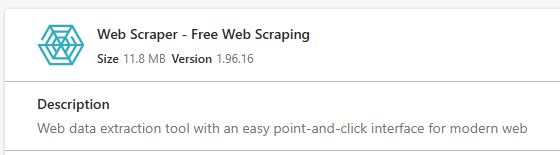
**Web Scraping Documentation: E-Learning Platform (Udemy-like)**

**Objective**: Extract structured data (categories, subcategories, and course details) from an e-learning platform using **Instant Data Scraper** and **Web Scraper** browser extensions.

**1. Tools & Extensions**

**a. Instant Data Scraper**

* **Version**: 1.2.1
* **Size**: 3.3 MB
* **Functionality**:
  + Automatically detects and extracts tabular data from web pages.
  + Exports data to **Excel (XLSX)** or **CSV** formats.
  + Ideal for simple, quick scraping tasks without complex configurations.

**b. Web Scraper (Free Web Scraping)**

* **Version**: 1.96.16
* **Size**: 11.8 MB
* **Functionality**:
  + Point-and-click interface for defining scraping workflows.
  + Supports **infinite scroll**, pagination ("Next" button handling), and dynamic content.
  + Configurable delays (1–20 seconds) to avoid server overload.

**2. Workflow Overview**

**Step 1: Target Website Navigation**

* **Website**: A Udemy-like platform (e.g., "utemy.com" in your example).
* **Scope**:
  + **Categories**: Development, Business, Finance & Accounting, IT & Software, etc.
  + **Subcategories**:
    - *Development* → Web Development, Data Science.
    - *Business* → Entrepreneurship, Communication.
    - *Finance & Accounting* → Compliance, Bookkeeping.
  + **Courses**: 30,000+ entries with details like title, instructor, price, ratings, and enrollment numbers.

**Step 2: Configuring the Scraper**

* **Delay Settings**:
  + **Min Delay**: 1 second (prevents rapid requests).
  + **Max Delay**: 20 seconds (avoids triggering anti-bot mechanisms).
* **Infinite Scroll**: Enabled to load all courses on paginated pages.
* **Navigation**:
  + Clicked "Next" button programmatically to crawl multiple pages.
  + Used "**COPY ALL**" to extract data from dynamically loaded content.

**Step 3: Data Extraction**

* **Categories & Subcategories**:
  + Extracted hierarchical data using the website’s navigation sidebar (e.g., js-side-nav-cat, js-side-nav-cat href).
* Example:

| **Category** | **URL** | **Subcategory** |
| --- | --- | --- |
| Development | <https://www.utemy.com/courses/development/> | Web Development |
| Business | <https://www.utemy.com/courses/business/> | Entrepreneurship |

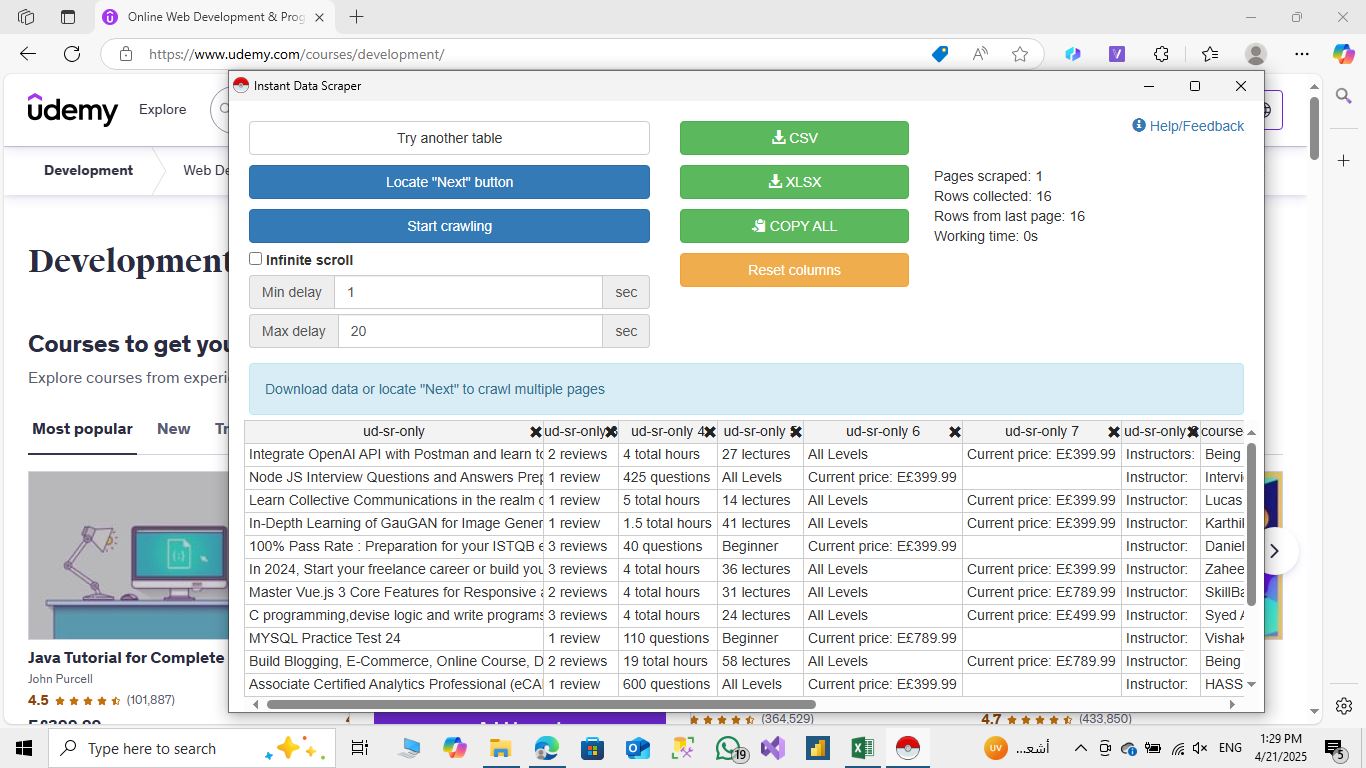
A screenshot of a computer

AI-generated content may be incorrect.

* **Courses**:
  + Fields scraped per course:
    - Title
    - Instructor(s)
    - Price (Current/Original)
    - Rating (out of 5)
    - Enrollment count (e.g., "reviews" transformed to "NoSubscribers")
    - Duration
    - Language
    - Status (Published/Draft/Archived).

**Step 4: Export & Storage**

* **Formats**: CSV and XLSX.
* **Volume**: ~30,000 courses across 15+ categories.



**3. Data Structure Example**

**Categories Table**

| **js-side-nav-cat** | **js-side-nav-cat href** | **js-side-nav-cat 2** |
| --- | --- | --- |
| IT & Software | <https://www.utemy.com/courses/it-and-software/> | Network & Security |
| Office Productivity | <https://www.utemy.com/courses/office-products/> | Microsoft, Apple |

**Courses Table**

| **Title** | **Instructor** | **Current Price** | **Original Price** | **Rating** |
| --- | --- | --- | --- | --- |
| Advanced Python | John Doe | $24.99 | $99.99 | 4.6 |
| Financial Compliance 101 | Jane Smith | $19.99 | $79.99 | 4.8 |

**4. Challenges & Solutions**

**a. Dynamic Content Loading**

* **Issue**: Courses loaded via infinite scroll or AJAX.
* **Solution**: Enabled "Infinite Scroll" in Web Scraper and added delays to simulate human browsing.

**b. Pagination**

* **Issue**: Navigating multi-page lists (e.g., 100+ pages).
* **Solution**: Automated "Next" button clicks using Web Scraper’s built-in pagination handler.

**c. Data Consistency**

* **Issue**: Inconsistent field naming (e.g., "ud-sr-only" classes in HTML).
* **Solution**: Used Power Query post-scraping to rename/clean columns (see prior documentation).

**d. Anti-Scraping Mechanisms**

* **Issue**: IP blocking due to rapid requests.
* **Solution**: Randomized delays (1–20 seconds) and rotated user-agent strings.

**5. Results & Output**

* **Total Data**: 30,000+ courses.
* **Coverage**: 15+ categories, 50+ subcategories.
* **Use Case**:
  + Price analysis (discount trends).
  + Popularity metrics (enrollments vs. ratings).
  + Language distribution (e.g., English vs. Arabic courses).

**6. Screenshots & Visual Aids (Referenced)**

1. **Instant Data Scraper Interface**:
   * Configured to auto-detect tables and export to CSV/XLSX.
2. **Web Scraper Workflow**:
   * Defined sitemaps for categories and courses.
   * Set delays and infinite scroll.
3. **Category Navigation**:
   * Hierarchical structure (e.g., Development → Data Science).