Окунькова О.О., старший викладач,
Дячук О.Ю., старший викладач,
Колощук М.С., асистент
Державний університет «Житомирська політехніка»

ПЕРЕВАГИ ТА НЕДОЛІКИ ПАРСИНГУ САЙТІВ ЗАСОБАМИ POWER QUERY

Парсинг сайтів - це процес автоматичного збору інформації з веб-сторінок. Парсинг відіграє критичну роль у сучасному цифровому світі, де дані є ключовим активом, ставши невід'ємною частиною сучасного цифрового світу, надаючи можливість отримувати та аналізувати великі обсяги даних для різних цілей, таких як: бізнес, фінанси, інвестиції, наукові дослідження та багато інших [1].

Для отримання та перетворення даних доцільно скористатися потужним інструмент у Microsoft Excel та Power BI - Power Query. Він має інтуїтивно зрозумілий графічний інтерфейс, який дозволяє користувачам без навичок програмування легко створювати запити для парсингу веб-сайтів.

Метою дослідження є аналіз сильних та слабких сторін використання інструменту Power Query для парсингу вебсайтів.

Power Query дозволяє об'єднувати інформацію з різних джерел, наприклад об'єднати дані з сайту, з файлу Excel, та з бази даних, що забезпечує реалізацію комплексного аналізу даних.

Після імпорту даних редактор Power Query, дозволяє виконати різноманітні операції для обробки отриманої інформації: фільтрація рядків, редагування стовпців, перетворення типів даних, операції з текстом, числами і датами, видалення дублікатів, обробка HTML-розмітки, дозволяючи користувачу витягувати та структурувати дані у потрібному форматі без необхідності писати складний код

Power Query автоматично розпізнає всі таблиці, що є на веб-сторінці, і виведить їх список. На жаль, часто зустрічаються сайти, де при спробі такого завантаження Power Query "не бачить" таблиць. Найчастіше причина в тому, що веб-дизайнер при створенні таблиці використовував у HTML-коді сторінки не стандартну конструкцію з тегом <TABLE>, та її аналог - вкладені теги-контейнери <DIV>. В такому випадку після підключення до веб-джерела необхідно завантажувати інформацію у вигляді текстового файлу, що дозволить завантажити дані як простий текст, а не як веб-сторінку. В такому випадку при роботі в Power Query весь процес отримання необхідних даних будується на використанні фільтрів, введені параметрів та вилученні зайвого контенту [2].

Слід зауважити, що надбудова Power Query - це інтерпретатор нової, скриптової, спеціалізованої для роботи з даними мови програмування M. Тому на кожну дію, яка виконується над даними у графічному інтерфейсі Power Query, у скрипт пишеться новий рядок коду. В результаті користувач отримує в панелі послідовність виконаних дій, що дозволяє маніпулювати виконаними кроками.

Після налаштування запиту Power Query може автоматично оновлювати дані з веб-сайту за розкладом або за запитом. Це значно спрощує процес збору даних та підвищує його ефективність, заощаджуючи час і зусилля. Висока продуктивність забезпечується використанням спеціальних алгоритмів для вилучення та трансформації даних [2].

В процесі аналізу парсингу сайтів за допомогою Power Query були виявлені деякі недоліки, а саме: обмежені можливості при роботі з динамічним контентом - складно парсити сайти, що використовують JavaScript, відсутність вбудованих інструментів для обходу захисту від парсингу, проблеми з сайтами, що мають CAPTCHA (автоматизований тест, який використовується для того, щоб відрізнити людину від комп'ютера) або обмеження доступу, можливі проблеми з продуктивністю при обробці дуже великих обсягів даних, обмежені можливості для паралельної обробки даних, складність налаштування парсингу складних структур даних та вкладених елементів, залежність від структури HTML, при зміні розмітки сайту потрібно переналаштовувати парсинг.

Незважаючи на виявлені недоліки можна з впевненістю стверджувати, що парсинг сайтів за допомогою Power Query є ефективним та зручним способом збору та аналізу веб-даних. Його переваги роблять використання Power Query особливо цінним для користувачів, які працюють з платформами Excel та Power BI і не мають глибоких знань в програмуванні.

Список використаних джерел:

1. Парсинг сайтів: що це і навіщо він потрібен URL: https://web-promo.ua/ua/blog/parsing-sajtov-chto-eto-i-zachem-nuzhen (дата звернення: 12.03.2025).
2. Равив Гил Power Query в Excel і Power BI: збирання, об'єднання та перетворення даних. Microsoft, 2021.378с.