УДК 004.85

# WEB SCRAPING ЯК ІНСТРУМЕНТ ОТРИМАННЯ ДАНИХ В МЕРЕЖІ ІНТЕРНЕТ

Студ. Яковець Р.І.
Наук. керівник доц. Пономаренко І.В.
Київський національний університет технологій та дизайну

Сучасний розвиток людства характеризується активним запровадженням інноваційних технологій, інформатизацією та віртуалізацією багатьох процесів. Одночасно суспільство накопичує великі обсяги різноманітної інформації, яка може бути використана окремими особами, компаніями або державами у якості цінного ресурсу для аналізу актуальних проблем та розробки на основі отриманих результатів ефективних управлінських рішень. Джерелами інформації можуть виступати різноманітні технологічні процеси, соціально-демографічні явища та ін. Особливе місце серед джерел даних посідає мережа Інтернет, яка вважається найбільшим генератором інформації структурованого, напівструктурованого та неструктурованого характеру.

В останні роки більшість сайтів підтримує API-інтерфейси, які дозволяють комп'ютерам збирати великі обсяги інформації. Для досягнення поставлених завдань використовується технологія Web scraping, яка представляє собою синтаксичне перетворення HTML-сторінок до зручного для вилучення інформації стану.

Розглянемо деякі з прикладів отримання інформації для потреб різних економічних суб'єктів. За допомогою API-інтерфейсу Yelp можна вилучати різноманітну бізнес-інформацію. Зазначений сервіс дозволяє здійснювати пошук компаній за певними характеристиками:

- вид економічної діяльності;
- географічне розташування;
- відстань компанії до певного об'єкта, міста чи адреси тощо.

Отримана у результаті збору інформація записується у форматі JSON та містить відомості про знайдені компанії, що відповідають заданим критеріям. Серед цінних даних, що отримується у результаті Web scraping, слід виділити інформацію про адреси, географічне положення та відстані, рейтинги, показники економічної діяльності, а також URL-адреси для інформації інших типів (хактеристики бізнесу, інформацію у мобільному форматі та ін.).

Технологія Web scraping також використовується для аналізу LinkedIn – соціальної мережі, яка дозволяє контактувати фахівцям у різних галузях для спілкування за професійними інтересами, пошуку нових вакансій, отримання інформації про діяльність компаній та ін. Користувачі LinkedIn можуть звернутися до API-інтерфейсів сайту (на основі REST і на основі JavaScript) з метою отримання інформації, яка також доступна для читання людиною на самому веб-сайті: контактна інформація, потоки соціальної інформації, переданої в колективне користування, категорії контенту, спілкування (повідомлення і запрошення до контакту) і т.д.

Реалізації підходу Web scraping дозволяє отримати унікальний контент, який може мати значну економічну цінність. Швидкість збору інформації та можливість вибору великої кількості параметрів для відбору робить зазначений алгоритм важливим інструментом в компанії, яка функціонує в умовах конкурентного середовища. Поряд з цим, слід пам'ятати про юридичні аспекти отримання інформації подібними методами з сайтів інших компаній. Не дивлячись на загрози порушення чинного законодавства існує потужний ринок послуг стосовно отримання цінних даних про діяльність конкурентів.