ЛІТЕРАТУРА
1. Офіційний сайт Python. URL: https://www.python.org/ (дата звернення: 22.01.2022).
2. Офіційний сайт бібліотеки для створення нейронних мереж scikit-learn.URL: https://scikit-learn.org/stable/ (дата звернення: 22.02.2022).
3. Офіційний сайт реляційної бази даних з відкритим кодом PostgreSQL. URL: https://www.postgresql.org/ (дата звернення 22.02.2022).
4. Офіційний сайт бібліотеки потоків введення/виведення аудіо PyAudio.URL: https://pypi.org/project/PyAudio/ (дата звернення: 22.01.2022).
5. Бібліотека pyttsx3 перетворення тексту на мову Text to Speech (TTS) для Python 2 та 3 та більш нових версій. URL: https://pypi.org/project/pyttsx3/ (дата звернення: 22.02.2022).
6. Офіційний сайт інтерфейсу Web Speech APISpeechRecognition URL: https://developer.mozilla.org/en-US/docs/Web/API/SpeechRecognition (дата звернення 22.01.2022).
7. Офіційний сайт розпізнавання мовлення Vosk Offline API. URL: https://alphacephei.com/vosk/ (дата звернення 22.01.2022).

Слюсар М. О.,
Криворізький національний університет
Купін А. І.
д.т.н., професор, Криворізький національний університет

ІНФОРМАЦІЙНА СИСТЕМА ДЛЯ АВТОМАТИЗОВАНОГО ПАРСИНГУАНКЕТНИХДАНИХВИПУСКНИКІВ У СОЦІАЛЬНИХ МЕРЕЖАХ

Зроблений опису поняття парсеру, веб-скрейпінгу та краулінгу. Розглянуто варіанті сучасні інструменти для веб-скрейпінгу. Створено програмне забезпечення парсеру для моніторингу кар'єри студентів-випускників Криворізького національного університету.

Сьогодні, в епоху інформаційних технологій, знання є найважливішим фактором успіху в суспільстві. Завдяки Інтернету кожен має можливість здобувати знання. Хоча формально Інтернет є глобальною мережею обчислювальних ресурсів, колективний доступ яких базується на використанні єдиної стандартної схеми адресації, потужної магістралі та високошвидкісних ліній зв'язку з головними комп'ютерами мережі. Сьогодні Інтернет - це не просто засіб спілкування між людьми, він також є найбільшим у світі джерелом інформації. І це джерело швидко зростає, адже щодня створюється багато веб-сайтів та порталів з важливим та цікавим контентом. Однак людина не може оволодіти такою кількістю інформації, і саме тут з'являється допомога комп'ютерів, особливо технології веб-вишкрібання. [1]

Веб-скрапінг - відносно новий винахід, який вплинув на життя кожного, хто в одному. Так чи інакше, стикається з необхідністю збору даних з Інтернету значно спрощується. Вишкрібання - це технологія, яка використовує сценарії для входу на веб-сайт, який видається звичайним користувачем, та збору інформації відповідно до заздалегідь визначених параметрів. Таким чином, дані з тисяч веб-сайтів можна отримати, обробити, систематизувати та зберегти у форматі простого тексту протягом декількох хвилин. Ця технологія особливо затребувана в галузі журналістики та статистики, що підтверджує її актуальність.

Парс - збирати і систематизувати інформацію, розміщену на певних сайтах, за допомогою спеціальних програм, що автоматизують процес. Парсинг законний, якщо він стосується збору інформації, що знаходиться у відкритому доступі. Тобто все, що можна і так зібрати вручну.

Парсери просто дозволяють прискорити процес і уникнути помилок через людський фактор.

Інша справа, як власник свіжозібраної бази розпорядиться подібною інформацією. Відповідальність може наступити саме за наступні дії.

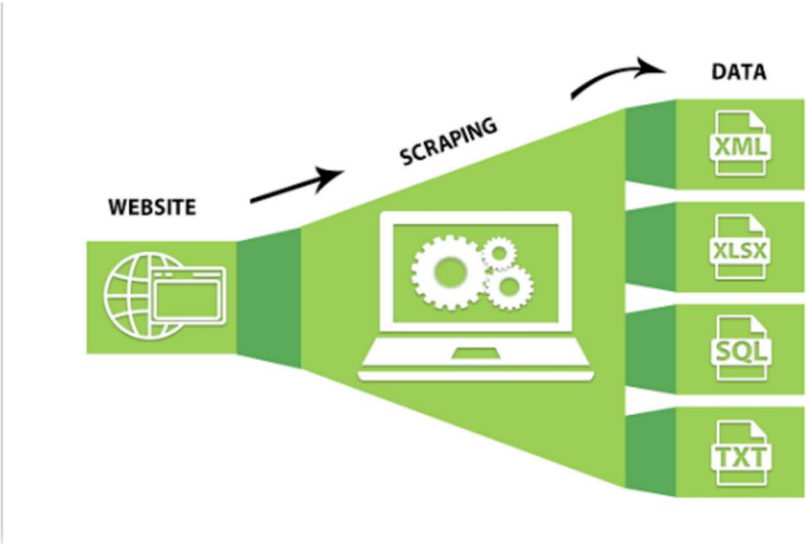

Рис.1 – Парсинг даних

Парсер - це програма, створена для автоматизації обробки поста отриманої с веб-сайту інформації. Іншими словами, парсер - це програма з аналізу та перетворення тексту з метою виділити з нього певні фрагменти або видалити зайве. На відміну від людини, парсер швидко і безпомилково відокремлює потрібну і відкидає зайву інформацію з html документа, і ефективно упаковує результат в певному форматі. [2]

Існує кілька способів отримання даних:

• Аналіз DOM дерева html сторінки. DOM - це уявлення HTML / XML документа у вигляді дерева об'єктів, яке дозволяє скриптам змінювати вміст і структуру документа. Дані в такому підході виходять з атрибуту елемента дерева або при відсутності таких, спускаючись вниз по DOM дереву. Отримані дані можуть бути будь-якої структури, а для отримання значення елемента, достатньо знати його розташування. Однак скрипт, який використовує даний метод потрібно прив'язувати до движку, а при зміні розташування елемента втрачається доступ до нього;

• Парсинг рядків. Цей спосіб парсинга має вузьку сферу застосування, так як отримання даних відбувається шляхом парсинга окремих рядків, що, в свою чергу, можливо лише в разі чітко фіксованого формату даних;

• Використання регулярних виразів.Цей метод, в основному, використовують для вирішення невеликих завдань або для написання власних процедур;

• XML парсинг. Ще одним підходом є розгляд HTML як XML дані. Причина в тому, що HTML рідко буває дійсним, під валідність HTML, в даному випадку, розуміється відповідність XML стандартам. Бібліотеки, які реалізували такий підхід, більше часу приділяли перетворенню HTML в XML, ніж безпосередньо парсингу даних;

• Візуальний підхід. Суть підходу в тому, щоб користувач міг без використання програмного мови або API налаштувати систему для отримання потрібних даних будь-якої складності і вкладеності. Однак такий підхід перебуває на початковій стадії розвитку.

Звичайно, парсери не читають тексти, вони всього лише порівнюють текст с запрошеним та діють за програмою. Розглянемо функціонал програми. Програма запускає пошук по випускникам "Криворізького Національного Університету". У знайдених студентам розберемо поля які необхідні в цілях програми.

Розроблене додаток можна розширити функціями, пов'язаними з синтаксичним аналізом, такими як, пошук слів кожної статті або підрахунок пропозицій всіх новинних ресурсів конкретного сайту. Також можна написати інтерфейс для програми або сформувати файл у форматі XML і JSON.

```
Thread.Sleep( millisecondsTimeout: 1000);
string name = _driver // IWebDriver
    .FindElement(By.XPath(
        "//*[@class='text-heading-xlarge inline t-24 v-align-middle break-words']")) // IWebElement
    .GetAttribute("innerText");
string title = _driver.FindElement(By.XPath("//div[@class='text-body-medium break-words']")) // IWebElement
    .GetAttribute("innerText");
string resultStr = $"{name};{title};";
var education IReadOnlyCollection<IWebElement> = _driver.FindElements(By.XPath(
    "//*[@class='pv-entity__school-name t-16 t-black t-bold']"));

foreach (var element in education)
{
```

Рис. 2 – Приклад с коду програми

Алгоритм роботи програми-парсера в кожному випадку приблизно однаковий:

- програма отримує доступ до ресурсу в мережі інтернет;

-завантаження коду сторінки, парсинг якою необхідно провести;

-читання і обробка інформації;

-надання результату в одному із зручних форматів - .html, .xml, .sql та інших.

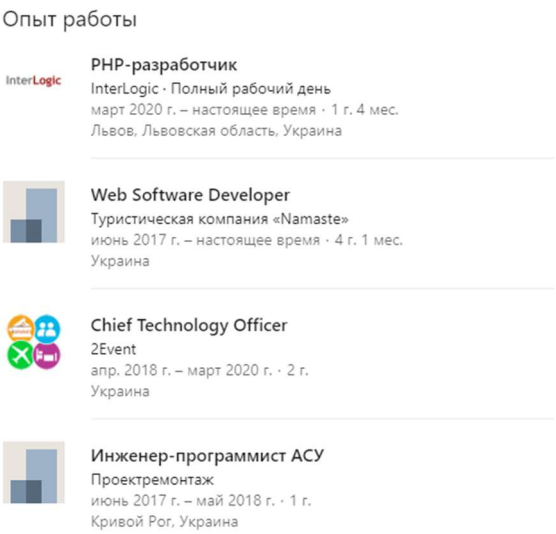

Рис. 3 – Приклад роботи парсингу

ВИСНОВКИ

Таким чином, у роботі досліджувалися питання, пов'язані з вивченням і застосуванням технології веб-скрейпінга, розглянуті види пошукових роботів і способи роботи синтаксичного аналізу. Для цього було розроблено та випробувано додаток, що дозволяє знайти всю потрібну інформацію з веб-ресурсу LinkedId і записати ці дані в реляційної базі даних простої структури.

ЛІТЕРАТУРА

1. Ріхтер Джеффрі. CLR via C #. Програмування на платформі Microsoft.NET Framework 4.5 мовою C #. 4-е изд. / Джеффрі Ріхтер. - Пітер, 2013. - 896 с.
2. Молінаро Ентоні. SQL. Збірник рецептів / Ентоні Молінаро. - O'Reilly, 2009. - 672 с.