Моренець Володимир Ігорович, к.ф.-м.н, асистент
Пономарьов Вадим Дмитрович, к.ф.-м.н, доцент
Шарапов Михайло Михайлович, к.ф.-м.н, доцент

МЕТОДИ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ПРИ АВТОМАТИЗОВАНОМУ ЗБОРІ ІНФОРМАЦІЇ З ВІДКРИТИХ ДЖЕРЕЛ

Роботу присвячено методичним підходам до викладання технологій автоматизованого збору інформації з відкритих вебджерел та методів інтелектуального аналізу даних у процесі підготовки фахівців з аналізу даних. Розглянуто особливості використання вебскрейпінгу, API та засобів попередньої обробки відкритих даних у навчальному процесі. Запропоновано pipeline-орієнтований підхід до організації навчання, що інтегрує етапи збору, очищення, аналізу та візуалізації даних у єдиний практично орієнтований процес. Описано структуру навчального курсу, принципи організації практичних проєктів та підходи до оцінювання результатів навчання. Показано, що використання реальних відкритих даних сприяє формуванню у студентів навичок роботи з гетерогенними даними та підвищує ефективність підготовки у сфері аналізу даних.

Ключові слова: відкриті дані, вебскрейпінг, інтелектуальний аналіз даних, API, pipeline, обробка даних.

The work is devoted to methodological approaches to teaching technologies for automated information collection from open web sources and methods of intelligent data analysis in the training of data analysis specialists. The paper considers the use of web scraping, APIs, and open data preprocessing tools within the educational process. A pipeline-oriented approach to course organization is proposed, integrating data collection, cleaning, analysis, and visualization into a unified practice-oriented workflow. The structure of the educational course, principles of practical project organization, and approaches to learning outcome assessment are described. It is shown that the use of real open datasets contributes to the development of students' skills in working with heterogeneous data and improves the effectiveness of training in the field of data analysis.

Keywords: open data, web scraping, intelligent data analysis, API, pipeline, data processing.

Вступ

Розвиток цифрового суспільства характеризується стрімким зростанням обсягів відкритих даних та активним використанням вебресурсів як джерел інформації для аналітичних досліджень, прогнозування та підтримки прийняття рішень. Значна частина таких даних представлена у вигляді вебсторінок, API-сервісів, відкритих державних порталів та інших гетерогенних джерел, що потребують автоматизованих підходів до збору, обробки та аналізу інформації. У зв'язку з цим особливого значення набувають технології вебскрейпінгу та методи інтелектуального аналізу даних.

Автоматизований збір інформації з відкритих вебджерел активно використовується у сфері соціально-економічного аналізу, маркетингових досліджень, моніторингу інформаційних ресурсів та інших прикладних задачах аналізу даних [1–3]. Водночас робота з відкритими вебданими характеризується гетерогенністю форматів, нестабільністю структури вебресурсів, наявністю пропусків та дублювань. Це вимагає від фахівців не лише знань мов програмування та статистичних методів, але й навичок автоматизованого збору, очищення, інтеграції та аналізу даних.

У підготовці студентів спеціальностей, пов'язаних з аналізом даних та прикладною статистикою, важливим є поєднання теоретичних знань із практично орієнтованими підходами до роботи з реальними відкритими даними. Традиційні навчальні курси часто зосереджуються окремо на програмуванні або статистичному аналізі, тоді як сучасні задачі Data Science потребують комплексного підходу, що інтегрує автоматизований збір даних, їх попередню обробку, аналіз та візуалізацію результатів.

Одним із перспективних підходів до організації навчального процесу є pipeline-орієнтована модель навчання, у межах якої студенти послідовно проходять усі етапи роботи з даними: від пошуку відкритих джерел та вебскрейпінгу до очищення, аналізу та представлення результатів. Такий підхід дозволяє формувати системне розуміння процесів роботи з даними та забезпечує практичну спрямованість навчання.

Метою роботи є розробка методичних підходів до викладання технологій автоматизованого збору інформації з відкритих вебджерел та методів інтелектуального аналізу даних у процесі підготовки фахівців з аналізу даних.

Особливості автоматизованого збору інформації з відкритих вебджерел

Автоматизований збір інформації з відкритих вебджерел є важливою складовою сучасних процесів аналізу даних та побудови інформаційно-аналітичних систем. Зростання кількості відкритих вебресурсів, електронних платформ та державних порталів відкритих даних створює значні можливості для отримання актуальної інформації у різних сферах діяльності. Водночас великі обсяги відкритих даних та постійне оновлення інформації роблять ручний збір даних малоефективним, що зумовлює необхідність використання автоматизованих підходів.

Основними джерелами відкритих даних є вебсайти, API-сервіси, відкриті державні реєстри, електронні каталоги та інформаційні портали. На відміну від традиційних структурованих баз даних, вебджерела часто мають неоднорідну структуру та різні формати представлення інформації. Дані можуть бути представлені у вигляді HTML-документів, JSON- або XML-файлів, текстових блоків, таблиць та мультимедійного контенту. Така гетерогенність ускладнює процес автоматизованого збору та подальшої аналітичної обробки інформації.

Одним із найбільш поширених підходів до автоматизованого збору інформації є вебскрейпінг, що передбачає отримання даних безпосередньо зі структури вебсторінок. Для реалізації вебскрейпінгу використовуються технології аналізу HTML-документів, DOM-структури, CSS-селекторів та XPath-запитів. У сучасних умовах важливу роль також відіграє використання API, які забезпечують доступ до структурованих даних у стандартизованих форматах та значно спрощують процес інтеграції інформації з різних джерел.

Процес автоматизованого збору інформації включає кілька основних етапів: пошук джерел даних, отримання інформації, попередню обробку, очищення, інтеграцію та підготовку даних до подальшого аналізу. Узагальнену pipeline-модель роботи з відкритими вебданими наведено на рис. 1.



Рис.1. Pipeline-модель роботи з відкритими вебданими

Особливістю відкритих вебданих є наявність пропущених значень, дублювань, різних форматів запису та нестабільності структури вебресурсів. Це потребує застосування методів попередньої обробки та очищення даних перед виконанням аналітичних задач. Крім того, важливими аспектами роботи з відкритими вебджерелами є дотримання етичних та правових принципів використання даних, врахування обмежень доступу до вебресурсів та коректне використання API-сервісів.

Таким чином, автоматизований збір інформації з відкритих вебджерел є комплексним процесом, що поєднує використання вебтехнологій, методів обробки даних та аналітичних підходів. Формування у студентів практичних навичок роботи з відкритими вебданими є важливою складовою сучасної підготовки фахівців у сфері аналізу даних та Data Science.

Методи інтелектуального аналізу даних у навчальному процесі

Важливою складовою підготовки фахівців у сфері аналізу даних є використання методів інтелектуального аналізу даних у процесі роботи з відкритими вебджерелами. На відміну від традиційних навчальних прикладів

із невеликими структурованими наборами даних, відкриті вебдані характеризуються значним обсягом, неоднорідністю структури та наявністю шуму, що потребує застосування сучасних методів обробки та аналізу інформації.

Одним із перших етапів роботи з даними є їх попередня обробка. У процесі навчання студенти виконують очищення наборів даних, усунення дублікатів, нормалізацію форматів запису, обробку пропущених значень та інтеграцію інформації з різних джерел. Такий підхід дозволяє сформувати практичні навички підготовки даних до подальшого аналітичного опрацювання та демонструє важливість якості даних для отримання коректних результатів аналізу.

Після етапу попередньої обробки застосовуються методи інтелектуального аналізу даних, що дозволяють виявляти закономірності у великих масивах інформації. У межах навчального процесу студенти опановують методи кореляційного аналізу, кластеризації, виявлення аномалій, аналізу часових рядів та побудови візуалізацій. Використання таких методів дає можливість аналізувати структуру даних, визначати взаємозв'язки між показниками та формувати аналітичні висновки на основі реальних наборів відкритих даних.

Одним із важливих елементів навчального процесу є оцінювання результатів виконання практичних проєктів. Для комплексного оцінювання рівня виконання студентських робіт може використовуватись інтегральний показник якості проєкту:

$$Q(S) = w_1 Q_{code} + w_2 Q_{data} + w_3 Q_{analysis} + w_4 Q_{report} \quad , \qquad (1)$$

де Q_{code} - якість програмної реалізації, Q_{data} - якість підготовки та обробки даних, $Q_{analysis}$ - рівень аналітичної обробки інформації, Q_{report} - якість представлення результатів дослідження, а w_i - вагові коефіцієнти відповідних критеріїв оцінювання.

Структуру критеріїв оцінювання практичних проєктів наведено у таблиці 1.

Таблиця 1. Критерії оцінювання практичних проєктів студентів

Критерій	**Зміст оцінювання**
Якість коду	Коректність роботи програмного рішення та використання інструментів автоматизованого збору даних
Підготовка даних	Очищення, нормалізація та інтеграція наборів даних
Аналіз даних	Використання методів інтелектуального аналізу та інтерпретація результатів
Представлення результатів	Якість візуалізацій, аналітичного звіту та презентації результатів

Використання методів інтелектуального аналізу даних у навчальному процесі дозволяє поєднати автоматизований збір інформації з практичним аналізом реальних відкритих даних. Такий підхід сприяє розвитку у студентів навичок роботи з великими масивами даних, формує системне розуміння процесів аналітичної обробки інформації та забезпечує підготовку до практичної діяльності у сфері Data Science та аналізу даних.

Практичні аспекти використання відкритих вебданих у навчальному процесі

Використання відкритих вебданих у навчальному процесі дозволяє наблизити навчальні завдання до реальних практичних задач сучасної аналітики даних. На відміну від штучно сформованих навчальних наборів даних, відкриті вебджерела характеризуються динамічністю, неоднорідністю структури та необхідністю постійного оновлення інформації. Це створює умови для формування у студентів практичних навичок роботи з реальними даними та сучасними інструментами автоматизованого аналізу інформації.

У межах навчального процесу студенти працюють із відкритими API, державними порталами даних, статистичними ресурсами та інформаційними вебплатформами. Практичні завдання охоплюють аналіз структури вебресурсів, автоматизований збір інформації, підготовку наборів даних та проведення аналітичного опрацювання результатів. Особлива увага

приділяється роботі з даними, що мають неоднорідний формат представлення та потребують попереднього очищення й нормалізації.

Практична реалізація завдань передбачає використання сучасних засобів аналізу даних та мов програмування для автоматизації процесів збору й обробки інформації. У процесі виконання проєктів студенти реалізують сценарії автоматизованого оновлення даних, інтеграції інформації з кількох джерел та побудови аналітичних візуалізацій. Такий підхід дозволяє формувати навички побудови інформаційно-аналітичних систем та роботи з великими масивами відкритих даних.

Однією з переваг використання відкритих вебджерел у навчальному процесі є можливість адаптації практичних завдань до сучасних інформаційних ресурсів та актуальних наборів даних. Це забезпечує постійне оновлення змісту навчальних проєктів і підвищує практичну цінність отриманих результатів. Крім того, робота з відкритими даними сприяє розвитку у студентів навичок самостійного пошуку інформації, аналізу структури даних та інтерпретації результатів дослідження.

Таким чином, використання відкритих вебданих у навчальному процесі забезпечує поєднання теоретичної підготовки з практичною реалізацією задач автоматизованого збору та інтелектуального аналізу інформації. Це дозволяє підвищити ефективність підготовки фахівців у сфері аналізу даних та забезпечує формування сучасних практичних компетентностей у галузі Data Science.

Висновки

У роботі досліджено особливості автоматизованого збору інформації з відкритих вебджерел та використання методів інтелектуального аналізу даних у навчальному процесі. Описано основні підходи до організації роботи з відкритими вебданими, використання технологій вебскрейпінгу, API та засобів попередньої обробки інформації.

Проаналізовано можливості використання pipeline-орієнтованого підходу у процесі підготовки фахівців з аналізу даних. Показано, що поєднання автоматизованого збору інформації, обробки даних та аналітичного аналізу дозволяє формувати у студентів комплексні практичні навички роботи з великими масивами відкритих даних.

Описано особливості використання методів інтелектуального аналізу даних у навчальному процесі, зокрема методів попередньої обробки даних, кореляційного аналізу, кластеризації та візуалізації результатів. Визначено,

що використання реальних відкритих вебджерел сприяє підвищенню практичної підготовки студентів та забезпечує наближення навчальних завдань до сучасних задач у сфері Data Science та інформаційної аналітики.

Отримані результати підтверджують доцільність використання практично орієнтованих підходів до викладання технологій автоматизованого збору та аналізу відкритих даних. Подальші дослідження можуть бути спрямовані на розширення використання сучасних методів машинного навчання та вдосконалення практичної підготовки студентів у сфері аналізу даних.

Список використаних джерел

1. Mitchell R. Web Scraping with Python: Collecting More Data from the Modern Web. 2-ге вид. Sebastopol : O'Reilly Media, 2018. 366 с.

2. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 2-ге вид. Waltham : Morgan Kaufmann, 2011. 703 с.

3. Пасічник В., Шаховська Н. Інтелектуальний аналіз даних. Львів : Видавництво Львівської політехніки, 2018. 432 с.

4. Python Language Reference. *Python.org*. *Documentation*. URL: https://www.python.org/ (дата звернення: 27.03.2026).