УДК 004.056.5

Бондар В. О.

МЕТОДИ ЗБОРУ ТА АНАЛІЗУ ІНФОРМАЦІЇ З ВЕБ-ДЖЕРЕЛ

Вінницький національний технічний університет

Анотація. *У цих тезах розглянуто сучасні методи збору та аналізу інформації з веб-джерел. Проведено аналіз основних підходів до веб-скрапінгу, роботи з API та агрегування даних з різноманітних Інтернет-ресурсів. Розглянуто переваги і недоліки використання різних технологій, зокрема Python-бібліотек (BeautifulSoup, Scrapy), сервісів новинних API та інструментів для попередньої обробки та класифікації даних. Окреслено основні виклики, такі як обмеження доступу до даних, часті зміни структури веб-ресурсів, необхідність обробки великого обсягу інформації та виявлення фейкових новин. Вказано на перспективи розвитку систем автоматизованого збору та аналізу новин для підвищення достовірності та релевантності отриманої інформації.*

Ключові слова: збір даних, аналіз інформації, веб-скрапінг, API, новинні агрегатори, машинне навчання, обробка текстів, фільтрація новин.

Abstract. *These theses consider modern methods of collecting and analyzing information from web sources. The analysis of the main approaches to web scraping, working with APIs, and aggregating data from various Internet resources is carried out. The advantages and disadvantages of using different technologies are discussed, including Python libraries (BeautifulSoup, Scrapy), news API services, and tools for preprocessing and classification of data. The main challenges are outlined, such as data access limitations, frequent changes in web resource structures, the need to process large volumes of information, and the detection of fake news. The prospects for the development of automated news collection and analysis systems to improve the reliability and relevance of the obtained information are indicated.*

Keywords: data collection, information analysis, web scraping, API, news aggregators, machine learning, text processing, news filtering.

Вступ

Сьогодні Інтернет є основним джерелом інформації для різних галузей: від медіа до бізнесу та науки. Потік новин постійно зростає, що створює потребу в автоматизованих системах збору та аналізу даних. Застосування таких систем дозволяє не лише зменшити часові та трудові витрати, але й забезпечити оперативне виявлення важливих подій, трендів, настроїв суспільства. Актуальність теми зумовлена потребою у створенні ефективних та гнучких інструментів, здатних збирати структуровану інформацію з численних гетерогенних джерел, автоматично її обробляти та оцінювати достовірність, що особливо важливо в умовах поширення фейкових новин та інформаційних маніпуляцій. Метою дослідження є аналіз існуючих методів збору та обробки новин із веб-джерел та визначення напрямків для підвищення їх ефективності.

Результати дослідження

Сучасні системи збору інформації з веб-джерел базуються на кількох основних підходах [1]:

1. Веб-скрапінг — технологія автоматизованого витягання інформації зі сторінок сайтів. Основними інструментами є Python-бібліотеки BeautifulSoup, Scrapy, Selenium. Перевагою є доступ до широкого спектра ресурсів, але існують ризики порушення політик доступу, а також нестабільність через зміну HTML-структур сайтів [2].

2. Робота з API — багато новинних порталів, соціальних мереж, агрегаторів пропонують спеціальні інтерфейси (NewsAPI, Mediastack, RSS), що забезпечують доступ до якісних структурованих даних. Недоліком може бути обмежений обсяг безкоштовних запитів або неповний набір інформації [3].

3. Агрегування даних — комбінування інформації з кількох джерел із подальшою обробкою для видалення дублікатів, нормалізації форматів та класифікації новин за тематиками, тональністю, джерелом [4].

4. Попередня обробка та фільтрація — включає методи токенізації, видалення стоп-слів, лематизації, а також класифікацію текстів за допомогою алгоритмів машинного навчання (SVM, Decision Trees, Naive Bayes) для виявлення спаму, дезінформації або тематичних відповідностей [5].

Висновки

Методи збору та аналізу даних із веб-джерел є потужним інструментом для створення інформаційних систем, здатних оперативно реагувати на зміни інформаційного простору. Подальший розвиток таких систем пов'язаний із застосуванням машинного навчання для автоматизації аналізу контенту, підвищенням ефективності збору інформації за рахунок використання API, а також із впровадженням механізмів перевірки достовірності новин для боротьби з фейками та дезінформацією.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Mitchell R. *Web Scraping with Python: Collecting More Data from the Modern Web*. — 2nd ed. — O'Reilly Media, 2018. – 290 p.
2. Janert P.K. *Data Analysis with Open Source Tools*. — O'Reilly Media, 2010. – 530 p.
3. NewsAPI Documentation. Режим доступу: https://newsapi.org/docs.
4. Russell M. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More*. — 3rd ed. — O'Reilly Media, 2019. – 423 p.
5. Bird S., Klein E., Loper E. *Natural Language Processing with Python*. — O'Reilly Media, 2009 – 504 p.

Ліщинська Людмила Броніславівна– доктор технічних наук, професор кафедри програмного забезпечення, Вінницький національний технічний університет, м. Вінниця, Україна. e-mail: llb@vntu.edu.ua.

Бондар Владислав Олександрович – студент групи 2ПІ-21б, факультет інформаційних технологій та комп'ютерної інженерії, Вінницький національний технічний університет, email: vladbondart11@gmail.com.

Lishchynska Lyudmyla Bronislavivna – PhD, Professor of the software department, Vinnytsia National Technical University, Vinnytsia, Ukraine. e-mail: llb@vntu.edu.ua.

Bondar Vladislav Oleksandrovich – student of 2PI-21b group, Faculty of Information Technologies and Computer Engineering, Vinnytsia National Technical University, Vinnytsia, Ukraine, email: vladbondart11@gmail.com.