УДК 004.415.2.043

КУВІЧКА МАКСИМ ЄВГЕНОВИЧ
ОЛІЙНИК ЮРІЙ ОЛЕКСАНДРОВИЧ

УНІФІКАЦІЯ СТРУКТУРИ НАДВЕЛИКИХ МАСИВІВ ТЕКСТОВИХ ДАНИХ ЗІБРАНИХ З РІЗНИХ ДЖЕРЕЛ

Анотація. У даній роботі розглянуто проблему стандартизації структури та форматів надвеликих масивів текстових даних, зібраних з різних джерел. Кожне джерело текстових даних має свою унікальну структуру даних, але при зборі інформації з декількох таких джерел потрібно уніфікувати такі дані для спрощення їх подальшої обробки.
КЛЮЧОВІ СЛОВА: ДЖЕРЕЛА ДАНИХ, ЗБІР ДАНИХ, ВЕЛИКІ ДАНІ, RSS, TELEGRAM.

Abstract. This article deals with the problem of standardizing the structure and formats of big text data collected from different sources. Each source of text data has its own unique data structure, but when collecting information from several such sources, it is necessary to unify such data to simplify their further processing.
KEY WORDS: DATA SOURCES, DATA COLLECTING, BIG DATA, RSS, TELEGRAM.

Вступ. З останніми технологічними досягненнями кількість даних зростає з кожним днем. Наприклад, ресурси з новинами, сенсорні та соціальні мережі генерують незліченні потоки даних. Загалом, великі дані створюються з кількох джерел у різних форматах з величезною швидкістю і тому їх важко зберігати та обробляти за допомогою традиційного програмного забезпечення. Для ефективного аналізу і обробки великих даних, їх спочатку треба зібрати, проте цей процес має певні складнощі, які необхідно вирішити. Однією з найбільших проблем є те, що кожне джерело даних має дані в своєму унікальному форматі, що і ми намагаємось вирішити в рамках цієї статті на прикладі різних видів новинних ресурсів.

Основна частина. В якості прикладу було обрано два види джерел текстових даних: Telegram-канали та RSS-стрічки новинних ресурсів. Розглянемо особливості їх структури та парсингу.

RSS (Really Simple Syndication) – це родина XML-форматів, що використовується для публікації та постачання інформації, яка часто змінюється, наприклад, записів в блозі, заголовків новин, анонсів статей, зображень, аудіо і відео матеріалів. Документ в форматі RSS, який також називають «стрічкою», «вебстрічкою» або «каналом», складається з повного або часткового тексту і метаданих.[3] Приклад вигляду RSS-стрічки з новинного ресурсу LB.ua представлено на Рис. 1. Теги першого рівня <title>, <link>, <description> - метадані джерела даних, а кожен тег <item> містить інформацію про статтю на цьому веб-ресурсі. Так, ми можемо отримати заголовок статті з тегу <title>, посилання на веб-сторінку зі статтею з тегу <link>, дату публікації статті з тегу <pubDate>, категорію новини з тегу <category>, короткий опис статті з тегу <description>, тощо. Проте RSS-стрічка не надає власне текст статті, що нам потрібно для повного аналізу. Для цього необхідно застосувати метод вебскрапінгу – перетворення у структуровані дані інформації з вебсторінок, які призначені для перегляду за допомогою браузера. Використовуючи будь-яке зручне програмне рішення для скрапінгу, можна отримати повний текст статті, адже

html-розмітка всіх веб-сторінок зі статтями має однакову, зручну для цього структуру.

Telegram-канали – це спеціальний формат обміну повідомленнями в месенджері Telegram, призначений для формату спілкування автор-підписник. Так тільки автор може публікувати і редагувати інформацію в каналі, а переглядати її може будь-який підписник. Також цей месенджер надає доступ до безкоштовного використання свого API, що робить можливим автоматизований збір даних і використання його як джерела великих текстових даних. Перш ніж отримати доступ до даних з будь-якого Telegram-каналу, необхідно авторизуватись в системі використовуючи номер телефону та пароль існуючого користувача та підписатися від імені цього користувача на цей Telegram-канал. Після виконання цих дій, ми можемо переглядати і збирати дані з обраного каналу. Приклад вигляду повідомлення з новинного Telegram-каналу «Новинач» представлено на Рис. 2. Це JSON-об'єкт, що містить два типи полів: системні та прикладні. Системні поля надають користувачу детальну інформацію про те, як повідомлення було опубліковано, і використовуються для побудови зручного користувацького інтерфейсу, але не несуть жодної користі для аналізу власне контенту або його метаданих. Наприклад, поле silent показує чи було повідомлення відправлено без сповіщення, поле from_scheduled показує чи повідомлення було відправлено як відкладене повідомлення, тощо. На противагу, прикладні поля – це саме ті, які зберігають текст повідомлення або корисні для аналізу метадані. Наприклад поле message зберігає повний текст повідомлення з його форматуванням, поле peer_id.channel_id зберігає унікальний ідентифікатор каналу, в якому це повідомлення було опубліковано, тощо. За допомогою поля peer_id.channel_id можна дізнатись метадані джерела даних такі як ім'я каналу. Також так як поле message зберігає форматування повідомлення, треба попередньо обробити та прибрати такі символи як «+», «\n», тощо.

В якості уніфікованої структури для даних, взятих з різного роду джерел, запропоновано використовувати формат JSON, бо його підтримують на вхід найбільш поширені платформи для обробки big data, як Hadoop, Apache Storm, Apache Spark Streaming, Flink, Kafka Streams. Кожен окремий запис представлений об'єктом з полями *title*, *link*, *body*, *author*, *category* та *date*. Поле *title* зберігає заголовок статті або *null* для даних з Telegram-каналів, поле *link* зберігає посилання на статтю або повідомлення, поле *body* – повний текст статті або повідомлення, поле *author* – назву веб-ресурсу або Telegram-каналу, *category* – категорія новини або значення *null* для даних з Telegram-каналів, поле *date* – дату публікації даних. Така структура зберігає всі необхідні для подальшого аналізу дані та метадані про кожне повідомлення та його джерело, є достатньо гнучкою для використання її з іншими джерелами даних та має строгу структуру, що передбачає відсутність значень для деяких полів (значення вказане як *null*), проте саме поле обов'язково має бути присутнім. Приклад даних, приведених до запропонованої структури, представлено на Рис. 3.

В роботі [4] розглядався метод обробки надвеликих масивів даних у форматі XML. Але формат XML є заважким для зберігання текстових даних, більш доцільно використати формат JSON, тому що файл у цьому форматі займає менше місця і швидше передається мережею, а його парсинг відбувається значно швидше і простіше, ніж у XML.

Запропонована структура зберігання даних характеризується:

1) Зберіганням мітки часу, що притаманно елементам потоків даних.
2) Зберіганням джерела даних, що дозволяє розрізнити звідки дані

були зібрані після їх потокової обробки.

3) Декларуванням строгої структури, яка полегшує обробку потоку таких даних за рахунок відсутності необхідності перевірки поля на існування, а лише роботи з його значенням.

```
<?xml version="1.0" encoding="utf-8"?>
<rss version="2.0" xmlns:atom="http://www.w3.org/2005/Atom">
<channel>
<atom:link href="https://lb.ua/rss/ukr/rss.xml" rel="self" type="application/rss+xml"/>
<title>LB.ua</title>
<link>https://lb.ua</link>
<description>LB.ua</description>
<language>ru</language>
<webMaster>info@lb.ua</webMaster><pubDate>Mon, 21 Nov 2022 20:59:07 +0200</pubDate><lastBuildDate>Mon, 21
<item>
        <guid>https://lb.ua/society/2022/11/21/536640_nbu_vipustiv_pamyatnu_medal_misto.html</guid>
        <pubDate>Mon, 21 Nov 2022 20:57:00 +0200</pubDate>
        <title>НБУ випустив пам’ятну медаль &amp;quot;Місто героїв – Харків&amp;quot;</title>
        <link>https://lb.ua/society/2022/11/21/536640_nbu_vipustiv_pamyatnu_medal_misto.html</link>
        <category>Суспільство</category>
        <author>info@lb.ua (LB.ua)</author>
        <enclosure url="https://i.lb.ua/018/54/637bc870c17a8.jpeg" type="image/jpeg" length="141343"/>
        <description>&lt;p&gt;Аверс медалі містить колаж із визначних пам’яток міста.&lt;/p&gt;</description>
</item>
<item>
        <guid>https://lb.ua/world/2022/11/21/536639_moldova_otrimaie_vid_frantsii_100_mln.html</guid>
        <pubDate>Mon, 21 Nov 2022 20:54:00 +0200</pubDate>
        <title>Молдова отримає від Франції 100 млн євро допомоги через війну в Україні, – Макрон</title>
        <link>https://lb.ua/world/2022/11/21/536639_moldova_otrimaie_vid_frantsii_100_mln.html</link>
        <category>Світ</category>
        <author>info@lb.ua (LB.ua)</author>
        <enclosure url="https://i.lb.ua/041/08/637bc9857b119.jpeg" type="image/jpeg" length="91446"/>
        <description>&lt;p&gt;10 мільйонів із них країна отримає до кінця цього року у вигляді грантів Верховному комісару
</item>
```

Рис. 1. Приклад вигляду RSS-стрічки новинного ресурсу LB.ua.

```
_: 'message',
flags: 11060864,
out: false,
mentioned: false,
media_unread: false,
silent: false,
post: true,
from_scheduled: false,
legacy: true,
edit_hide: true,
pinned: false,
noforwards: false,
id: 30334,
peer_id: { _: 'peerChannel', channel_id: '1134948258' },
date: 1669036487,
message: 'Парламентська асамблея НАТО визнала росію державою-терористом,
  '\n' +
  'Про це повідомив голова постійної делегації України в ПА НАТО Єгор Чер
media: {
  _: 'messageMediaPhoto',
  flags: 1,
  photo: {
    _: 'photo',
    flags: 0,
    has_stickers: false,
    id: '5465542576938533625',
    access_hash: '4377408161334908280',
    file_reference: [Uint8Array],
    date: 1669036486,
    sizes: [Array],
    dc_id: 2
  }
},
```

Рис. 2. Приклад вигляду повідомлення з Telegram-каналу «Новинач».

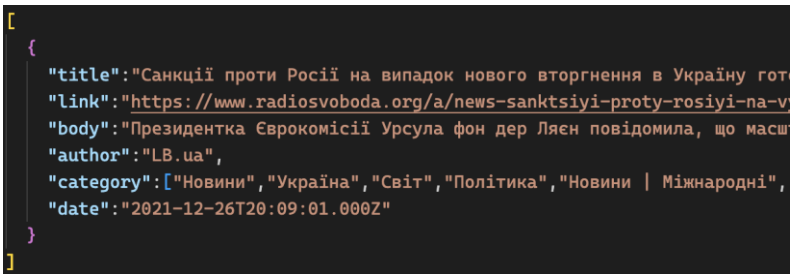

Рис. 3. Приклад вигляду даних, приведених до запропонованої структури.

Висновки.

В роботі розглянуто процес уніфікації структури та формату надвеликих масивів текстових даних на основі аналізу таких джерел як Telegram-канали та RSS-стрічки, визначено специфіку збору даних з них і особливості структур та форматів даних, які використовуються цими джерелами. Запропоновано використати формат JSON для зберігання та обробки надвеликих масивів текстових даних.

Запропоновано та описано уніфіковану структуру даних для різного роду новинних джерел великих текстових даних, що включає зберігання мітки часу та джерела даних, а також декларування строгої структури.

Список літератури

1. I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. U. Khan, The rise of ‘big data’ on cloud computing: Review and open research issues, Inform. Syst., vol. 47, pp. 98–115, 2015.
2. J. H. Yu and Z. M. Zhou, Components and development in big data system: A survey, J. Electr. Sci. Technol., vol. 17, no. 1, pp. 51–72, 2019.
3. RSS [Електронний ресурс] – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/RSS.
4. Педоренко, О. Р. Математичне та програмне забезпечення обробки надвеликих масивів даних у форматі XML : магістерська дис. : 121 Інженерія програмного забезпечення / Педоренко Олег Русланович . - Київ, 2019. - 75 с.