РЕФЕРАТ

Розмір пояснювальної записки – 107 аркушів, містить 20 ілюстрацій, 28 таблиць, 3 додатки, 21 посиланя на джерела.

Актуальність теми. З кожним роком даних стає все більше, вони можуть принести користь в будь-якій сфері нашого життя за умови правильної обробки. Тема роботи є актуальною, оскільки на сьогодні універсального засобу для збору надвеликих масивів текстових даних з різних джерел не існує.

Метою роботи є створення уніфікації структури та формату надвеликих масивів текстових даних за рахунок використання архітектурних рішень, які дозволяють користувачам розширювати його для власних цілей з мінімальними зусиллями. Для досягнення цієї мети необхідно вирішити такі **задачі**:

- порівняльний аналіз наявних рішень для збору надвеликих масивів текстових даних;
- формулювання технічних особливостей збору надвеликих масивів текстових даних;
- розробка уніфікованої структури надвеликих текстових даних, зібраних з різних джерел;
- розробка програмного забезпечення для збору надвеликих масивів текстових даних;
- реалізація модульної архітектури в програмному рішенні;
- оцінка ефективності запропонованого рішення.

Об'єктом дослідження роботи є математичне, інформаційне та програмне забезпечення збору надвеликих масивів текстових даних.

Предметом дослідження є методи збору надвеликих масивів текстових даних.

Науковою новизною роботи є створення уніфікованого структури даних для джерел великих текстових даних різної природи, що включає зберігання мітки часу та джерела даних, а також декларування строгої структури.

Практичне значення отриманих результатів полягає у можливості використання запропонованої уніфікованої структури для інтеграції між різними системами збору надвеликих масивів текстових даних.

Зв'язок роботи з науковими програмами, планами, темами: дисертаційна робота виконувалась на кафедрі інформатики та програмної інженерії Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського» в рамках теми «Методи та технології високопродуктивних обчислень та обробки надвеликих масивів даних». Державний реєстраційний номер 0117U000924.

Апробація: Основні положення роботи доповідались і обговорювались на III Всеукраїнській науково-практичній конференції молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології (Soft-Tech-2022)».

Публікації. Наукові положення дисертації опубліковані в:

1) Кувічка М.Є. Уніфікація структури надвеликих масивів текстових даних, зібраних з різних джерел / М.Є. Кувічка, Ю.О. Олійник // Матеріали III Всеукраїнської науково-практичної конференції молодих вчених та студентів «Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2022 осінь) – м. Київ: НТУУ «КПІ ім. Ігоря Сікорського», 23-25 листопада 2022 р.

Ключові слова: ВЕЛИКІ ДАНІ, ЗБІР ДАНИХ, СТРУКТУРИЗАЦІЯ ДАНИХ, ВЕБСКРАПІНГ.