РЕФЕРАТ

Розмір пояснювальної записки – 94 аркуші, містить 17 ілюстрацій, 25 таблиць, 5 додатків, 29 посилань на джерела.

Актуальність теми. У роботі розглянуто проблему в області обробки потоків текстових даних, показано основні особливості наявних платформ обробки потоків текстових даних, їх переваги та недоліки. Виявлено потребу в удосконаленні методу обробки природної мови для потоків текстових даних.

Мета дослідження. Основною метою є покращення наявних інструментів обробки природної мови для забезпечення підтримки україномовних текстів та розробка програмного забезпечення, здатного проводити аналіз потоків текстових даних в реальному часі.

Об’єкт дослідження: потоки текстових даних.

Предмет дослідження: методи та засоби створення програмного забезпечення для обробки природної мови для потоків текстових даних в режимі реального часу.

Для реалізації поставленої мети **сформульовані наступні завдання**:

- порівняльний аналіз наявних рішень для обробки потоків текстових даних в реальному часі;
- формулювання структурних та технічних особливостей джерел потоків текстових даних;
- підбір та підготовка україномовного словника;
- впровадження наявних рішень для забезпечення підтримки морфологічного аналізу;
- розробка програмного забезпечення обробки потоків текстових даних із використанням морфологічного аналізатора в режимі реального часу;

- оцінка ефективності запропонованого рішення.

Наукова новизна результатів магістерської дисертації полягає в удосконаленні методу обробки природної мови текстових даних за рахунок впровадження підтримки потокової обробки у режимі реального часу, що підвищує швидкодію обробки та дозволяє виконувати розподілені обчислення, а також покращенні рівня підтримки обробки україномовних текстів за рахунок інтеграції словника ВЕСУМ.

Практичне значення отриманих результатів полягає в тому, що запропоновано архітектуру програмного забезпечення обробки потоків текстових даних в реальному часі із використанням Apache Spark та бібліотеки потокової обробки Spark Streaming з зберіганням результатів в пошуковий сервер Elasticsearch із використанням рушія візуалізації Kibana, а також розроблено програмне забезпечення з використанням запропонованої архітектури для аналізу потоків україномовних текстових даних. Розроблене програмне забезпечення може бути використане в подальшому для обробки потоків текстових даних з україномовних джерел, а також для виконання ширшого спектру задач NLP (наприклад, сентимент-аналіз або інтелектуальних аналіз текстових даних).

Зв'язок з науковими програмами, планами, темами. 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. Наукові положення дисертації пройшли апробацію на III Всеукраїнській науково-практичній конференції молодих вчених та студентів

«Інженерія програмного забезпечення і передові інформаційні технології» (SoftTech-2022 осінь) – м. Київ.

Публікації. Наукові положення дисертації опубліковані в:

1) Федорович І.А. Моделі обробки потоків текстових даних в рушії Apache Spark Structured Streaming / І.А. Федорович, 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: ОБРОБКА ПРИРОДНОЇ МОВИ, ОБРОБКА ПОТОКІВ ТЕКТОВИХ ДАНИХ, ОБРОБКА ПОТОКІВ В РЕАЛЬНОМУ ЧАСІ, APACHE SPARK, SPARK STRUCTURED STREAMING.