УДК 004.415.5

Федорович І. – ст. гр. ІП-13мп
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ОЦІНКА ЕФЕКТИВНОСТІ РОБОТИ РОЗРОБЛЕНОГО ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ ОБРОБКИ ПРИРОДНОЇ МОВИ ДЛЯ ПОТОКІВ ТЕКСТОВИХ ДАНИХ

Науковий керівник: к.т.н., доцент Олійник Ю. О.

Fedorovych I.
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

PERFORMANCE EVALUATION OF THE DEVELOPED TEXT DATA STREAM NATURAL LANGUAGE PROCESSING SOFTWARE

Supervisor: Oliinyk Y.

Ключові слова: обробка природної мови, обробка потоків текстових даних, apache spark structured streaming.
Keywords: natural language processing, text data stream processing, apache spark structured streaming.

Вступ. Станом на 2020 рік, у світі щороку генерується понад 44 зетабайт даних [1]. З огляду на надвеликий обсяг даних, сфера обробки великих даних, або ж BigData, користується попитом і постійно розвивається, забезпечуючи більшу кількість засобів оброки надвеликих масивів даних та збільшуючи їх пропускну здатність. Одним з видів даних є текстові дані, які в сучасних реаліях генеруються щосекунди, для яких стандартні засоби обробки тексту не є ефективними. Тому існує потреба в засобах обробки потоків текстових даних, у яких текстові дані надходять безперервно та обробляються поодинці або згруповано. Для таких засобів одним з ключових показників ефективності є швидкість обробки. З огляду на те, що текстові дані містять в собі певну інформацію, яка може бути видобута, важливою задачею є обробка природної мови (Natural language processing, NLP). Особливо актуальним є питання забезпечення NLP для української мови, оскільки розвиток цієї сфери є достатньо низьким та потребує покращення. З цих причин було розроблено програмне забезпечення обробки природної мови для потоків текстових даних з використання рушія Apache Spark Structured Streaming [2], морфологічного аналізатора Pymorphy2 та Великого електронного словника української мови (ВЕСУМ) [3].

Основна частина. Для дослідження ефективності розробленого програмного забезпечення було використано ПК з ОС Windows 11, процесором AMD Ryzen 7 3700U (4x2.3-4.0 GHz), ОЗУ 16 Гб DDR4 2400 MHz та відеокартою AMD Radeon Vega 10. В якості вхідних даних було обрано перекладений фрагмент художнього твору на українську мову обсягом у 1556 слів, 64 речення та 10035 символів. Розроблене програмне забезпечення було порівняне з програмним забезпеченням, описаним у роботі [4]. Результати проведення експерименту зображені на рисунках 1-2. Порівняння результатів експерименту наведено у таблиці 1.

Таблиця 1 – Порівняльна таблиця результатів експерименту

	Робота [4]	Розроблене програмне забезпечення
Загальний час, с	281	13.93
Час структуризації, с	9.42	10.93
Час токенізації, с	128.37	0.47
Час фільтрації, с	141.25	0.46
Середній час токенізації одного слова, с	0.082	0.000302
Середній час фільтрації одного слова, с	0.237	0.0008
Середній час обробки (структуризація та токенізація) одного слова, с	0.18	0.009

Висновки. Згідно результатів експерименту, розроблене програмне забезпечення має в 20 разів менший загальний час обробки, ніж існуючий аналог, та досягає швидкості обробки у 9 мілісекунд на слово, або ж близько 111 слів за секунду. Основне прискорення було досягнуто за рахунок покращення процесу токенізації та фільтрації, а також застосування вдосконалених інструментів DataFrame та Apache Spark Structured Steaming, на противагу старішим альтернативам – RDD та Apache Spark Streaming.

Список літератури

1. F. Mostajabi, A. A. Safaei and A. Sahafi, "A Systematic Review of Data Models for the Big Data Problem," in IEEE Access, vol. 9, pp. 128889-128904, 2021, doi: 10.1109/ACCESS.2021.3112880.
2. Structured streaming: A declarative API for real-time applications in Apache Spark / [M. Armbrust, T. Das, J. Torres та ін.]. // Proc. Int. Conf. Manage. Data. – 2018. – С. 601–613.
3. Рисін А., Старко В. Великий електронний словник української мови (ВЕСУМ). Вебверсія 5.6.2. 2005-2022 [Електронний ресурс] / Андрій Рисін, Василь Старко – Режим доступу до ресурсу: https://r2u.org.ua/vesum/
4. Якимчук, О. А. Програмна бібліотека обробки текстової інформації для Apache Spark : магістерська дис. : 121 Інженерія програмного забезпечення / Якимчук Олександр Анатолійович. – Київ, 2020. – 76 с.