УДОСКОНАЛЕННЯ МЕТОДІВ ЗБЕРІГАННЯ ТЕКСТОВИХ ДАНИХ

Каланча А.Д., Угрин Д.І.
Чернівецький національний університет ім. Ю. Федьковича, Чернівці

Natural Language Processing (NLP) - це галузь штучного інтелекту, що досліджує можливості комп'ютерів у розумінні, аналізі та взаємодії з людською мовою. Вона охоплює широкий спектр технологій, включаючи попередню обробку текстів [1] для підготовки їх до подальшого аналізу та використання. У нашому випадку, NLP використовується для аналізу текстів [2] та визначення ключових слів [3], що містяться в Telegram-каналах. Стрімкий розвиток аналізу великих даних вимагає не тільки вдосконалення технік застосування, а й пошук ефективних та дешевих методів обробки. Одна з основних проблем - це неефективне зберігання великої кількості текстових даних, що надходять з джерел. Розв'язання цієї проблеми має велике значення, оскільки дозволяє цільовим системам використовувати мінімум об'єму сховища при збереженні найважливішої інформації.

Дослідження проводилися шляхом аналізу ефективності різних методів збереження та обробки текстових даних. Розглянуті методи включали збереження неочищеного тексту, збереження очищених токенів у форматі масиву та рядка, збереження у базу даних кодів оброблених токенів у форматі стрічки разом зі словником токен-код. Проаналізовано швидкість генерування вхідних даних для того щоб збалансувати роботу системи в часовому просторі. Результати досліджень показали, що збереження масивів токенів у форматі рядків, об'єднаних одним символом, дозволило зменшити розмір даних на 17% порівняно з методом збереження неопрацьованих даних.

Найкращий результат отримано за допомогою запропонованого методу кодування, який, хоча пов'язаний з ризиками для цілісності бази даних, пропонує кращу економію пам'яті для великих об'ємів даних, але гіршу для малих об'ємів. Пікова генерація повідомлень відбувається у проміжку 5:00 до 21:00. Лінійної залежності об'єму від дня тижня чи місяця не виявлено.

Весь процес попередньої обробки текстів, який відбувається після завантаження з Telegram, займає менше часу аніж попередній етап, що виключає зворотній тиск у нашу алгоритмі. Дані дослідження сприяють розвитку та масштабуванню додатків природньої обробки мови при мінімальних ресурсах.

Література:

1. Camacho-Collados J. On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *eprint arXiv.* 2018. № 1707(01780). С. 1–4. URL: https://arxiv.org/abs/1707.01780 (дата звернення: 01.05.2024).

2. Lytvyn V. Analysis of statistical methods for stable combinations determination of keywords identification. *Information technology: Eastern-European Journal of Enterprise Technologies.* 2018. № 2/2 (92). С. 23–37.

3. Lytvyn V. Development of a method for determining the keywords in the slavic language texts based on the technology of web mining. *Information technology. Industry control systems: Eastern-European Journal of Enterprise Technologies.* 2017. № 2/2 (86). С. 14–23.