УДК 004.8

Ю.О. ОЛІЙНИК
Національний технічний університет України
«Київський політехнічний інститут ім. Ігоря Сікорського»

СИСТЕМА АНАЛІЗУ ТЕКСТОВИХ ПОТОКІВ ДАНИХ

Дослідження присвячене розробці системи аналізу текстових потоків даних. В постановці завдання наведено проблему обробки потоків текстової інформації та відзначається відсутність програмних засобів одночасної обробки потоків текстових даних українською та російською мовами.

Проведено аналіз останніх досліджень та встановлено, що для обробки потоків даних необхідно застосовувати спеціалізоване програмне забезпечення обробки потокових даних. Виявлено, що існує вкрай мало засобів для обробки україномовних текстів, а також те, що не існує засобів для одночасної підтримки україномовних та російськомовних текстів.

Метою даного дослідження є розробка архітектури та реалізація програмного забезпечення системи аналізу текстових потоків даних. Наведено опис математичної моделі потоку текстових даних на основі ковзного вікна. Наведено завдання для обробки потоків текстових даних від базових перетворень тексту та попередньої обробки до інтелектуального аналізу текстових потоків даних. Сформульовано математичну постановку завдання визначення емоційного забарвлення потоків текстових даних на основі моделі ковзного вікна.

В процесі дослідження виділено 4 підсистеми: підсистему збору та транспортування повідомлень потоків даних, підсистему аналізу текстових потоків, підсистему зберігання результатів аналізу потоків даних та підсистему візуалізації. Особливістю системи є підтримка обробки україномовних текстів, для чого було спеціально розроблено програмну бібліотеку UANLP. Дана бібліотека дозволяє також оброблювати російськомовні тексти. Обробка потоків текстових даних виконується на основі компоненту Spark Streaming, що підтримує роботу з вікнами. Бібліотека Spark MLib та ML дозволяють використовувати засоби машинного навчання для аналітичної обробки потоків текстових даних, на основі яких виконується сентимент аналіз, виявлення аномалій, елементів пропаганди, дезінформації тощо.

Обґрунтовано використання програмних компонент – сервісу повідомлень Kafka, технології розподіленої обробки даних Apache Spark, бази даних Elasticsearch та сервісу візуалізації Kibana. Описано процес обробки даних від генерації потоків даних до візуалізації результатів аналізу.

Ключові слова: потоки текстових даних, онлайн обробка, text mining, Apache Spark.

Ю.А. ОЛЕЙНИК
Национальный технический университет Украины
«Киевский политехнический институт им. Игоря Сикорского»

СИСТЕМА АНАЛИЗА ТЕКСТОВЫХ ПОТОКОВ ДАННЫХ

Исследование посвящено разработке системы анализа текстовых потоков данных. В постановке задачи приведены проблемы обработки потоков текстовой информации и отмечается отсутствие программных средств одновременной обработки потоков текстовых данных на украинском и русском языках.

10.32782/2618-0340/2020.1-3.15

Проведен анализ последних исследований и установлено, что для обработки потоков данных необходимо применять специализированное программное обеспечение обработки потоковых данных. Выявлено, что существует крайне мало средств для обработки украиноязычных текстов, а также то, что не существует средств для одновременной поддержки обработки украиноязычных и русскоязычных текстов.

Целью данного исследования является разработка архитектуры и реализация программного обеспечения системы анализа текстовых потоков данных. Приведено описание математической модели потока текстовых данных на основе скользящего окна. Приведены задачи для обработки потоков текстовых данных от базовых преобразований текста и предварительной обработки до интеллектуального анализа текстовых потоков данных. Сформулирована математическая постановка задачи определения эмоциональной окраски потоков текстовых данных на основе модели скользящего окна.

В процессе исследования выделено 4 подсистемы: подсистему сбора и транспортировки сообщений потоков данных, подсистему анализа текстовых потоков данных, подсистему хранения результатов анализа потоков данных и подсистему визуализации.

Особенностью системы является поддержка обработки украиноязычных текстов, для чего была специально разработана программная библиотека UANLP. Данная библиотека позволяет также обрабатывать русскоязычные тексты. Обработка потоков текстовых данных выполняется на основе компонента Spark Streaming, поддерживающий работу с окнами. Библиотека Spark MLib и ML позволяют использовать средства машинного обучения для аналитической обработки потоков текстовых данных, на основе которых выполняется анализ тональности, выявление аномалий, элементов пропаганды, дезинформации и тому подобное.

Обосновано использование таких программных компонент как сервис сообщений Kafka, технологии распределенной обработки данных Apache Spark, базы данных Elasticsearch и сервиса визуализации Kibana. Описан процесс обработки данных от генерации потоков данных до визуализации результатов анализа.

Ключевые слова: потоки текстовых данных, онлайн обработка, text mining, Apache Spark.

Yu.O. OLIINYK
National Technical University of Ukraine
'Igor Sikorskiy Kyiv Politechnical Institute'

TEXT DATA STREAM ANALYSIS SYSTEM

The study is devoted to text data stream analysis system development. The problem statement deals with the problem of text data stream processing and the lack of software for simultaneous processing of text data streams in Ukrainian and Russian.

The analysis of the last researches is carried out and established that for data flow processing it is necessary to apply the specialized software of data stream processing. It was found that there are only few tools for processing Ukrainian-language texts are exists, as well as the fact that there are no tools for simultaneous support of procession Ukrainian-language and Russian-language texts.

The purpose of this study is to software architecture development and implementation of the data streams analysis software. A description of the mathematical model of text data flow based on a sliding window is given. The tasks for processing text data streams are defined. Tasks from basic text transformations and pre-processing to intellectual analysis of

10.32782/2618-0340/2020.1-3.15

text data streams are given. The mathematical definition of the problem of determining the emotional color of text data streams on the base of the sliding window model is formulated.

For subsystems are allocated: a collecting and transporting messages of data streams subsystem, an analysis of text streams subsystem, a storage of results of the analysis of data streams subsystem and a visualization subsystem.

A systems features are support of Ukrainian-language texts processing, for this purpose the UANLP library was specially developed. This library also supports Russian-language texts processing. Processing of text data streams is performed on the base of the Spark Streaming component, which supports work with sliding windows. The Spark MLib and ML libraries allow the use machine learning tools for analytical processing of text data streams, such as sentiment analysis, detection of anomalies, elements of propaganda, misinformation are performed.

Main software component – messaging service Kafka, distributed data processing technology Apache Spark, Elasticsearch database and Kibana visualization service. Made data processing description from data streams generation to analysis results visualization.

Keywords: text stream, online data processing, text mining, Apache Spark.

Постановка проблеми

Кожного дня продукуються потоки текстових даних [1] у вигляді новин, повідомлень в соціальних мережах, месенджерах. Інформаційні виклики сьогодення потребують оперативного аналізу та реагування в режимі онлайн на інформацію, що проходить в цих потоках [1–2]. Без сучасних підходів обробки надвеликих масивів інформації Big Data неможливо обробити такі потоки. Крім того, згідно джерел [1, 4] відсутні програмні засоби, що здатні оброблювати потоки текстових даних одночасно українською та російською мовами, що актуально для нашої країни. Тому необхідно виконати проектування та програмну реалізацію системи аналізу потоків текстових даних (надалі Системи). Для цього необхідно вирішити такі завдання:

1) виділити завдання обробки та аналізу потоків текстових даних;
2) спроектувати архітектуру програмного забезпечення Системи;
3) виконати програмну реалізацію Системи;
4) забезпечити підтримку обробки україномовних та російськомовних текстів.

Аналіз останніх досліджень і публікацій

Згідно книги [3] повний проект обробки потоків даних складається з таких етапів: знайти вихідні дані, завантажити дані, виконати аналіз, зберегти результат, показати результат. В 2008 році з’явився програмних комплекс Hadoop, де було реалізовано парадигму MapReduce. Крім того, реалізовано розподілену файлову систему HDFS. Але на даний час Hadoop є досить застарілою технологією та поступається новітнім засобам, таким як Apache Spark, Storm, Kafka, Flink, Akka та іншим [3].

В роботах [1, 4] описуються проблеми, пов’язані з підтримкою україномовних текстових даних, що обумовлено вкрай малим переліком програмних засобів та моделей. Крім того, практично відсутні засоби для одночасної підтримки україномовних та російськомовних текстів. В роботах [1, 2, 4] описуються окремі елементи системи аналізу текстових потоків даних, але не описується Система в цілому.

Для обробки потоків даних необхідно застосовувати високопродуктивне спеціальне програмне забезпечення, здатне масштабуватись, наприклад, Apache Spark Streaming [6] (рис.1), Apache Storm, Hadoop Streaming, Splunk Stream тощо.

10.32782/2618-0340/2020.1-3.15

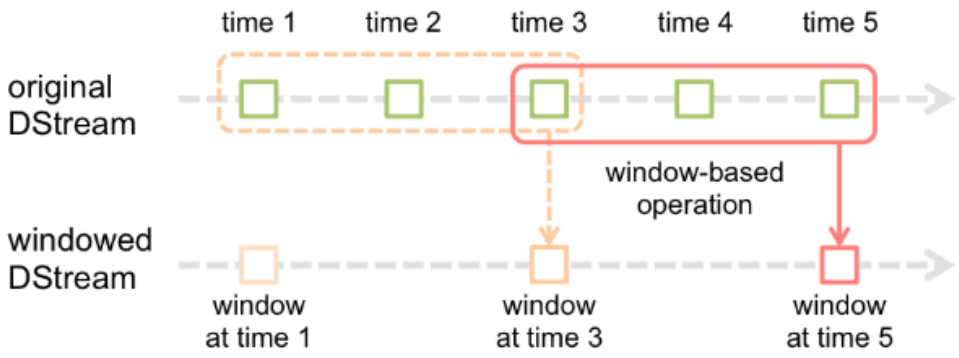

Рис. 1. Вікно обробки потоку текстових даних в Apache Spark Streaming.

Мета дослідження

Метою даного дослідження є розробка архітектури та реалізація програмного забезпечення системи аналізу текстових потоків даних.

Викладення основного матеріалу дослідження

Опис математичної моделі потоку текстових даних на основі ковзного вікна. Під текстовим потоком даних розуміємо нескінченний потік даних, які надходять з одного або декількох джерел, де пара (d_j; t_j) визначає, що текстове повідомлення d_j отримано в точці часу t_j.

На рис. 1 та рис. 2 представлено приклад віконного процесу обробки даних. «Ковзне вікно» (Sliding Window) – часове вікно з певним зсувом. Якщо часовий інтервал між часовими відмітками дорівнює 1 секунді, інтервал вікна складе 3 с., а інтервал зсуву складе 2 с.

Математична модель потоку даних наводиться на рис. 2.

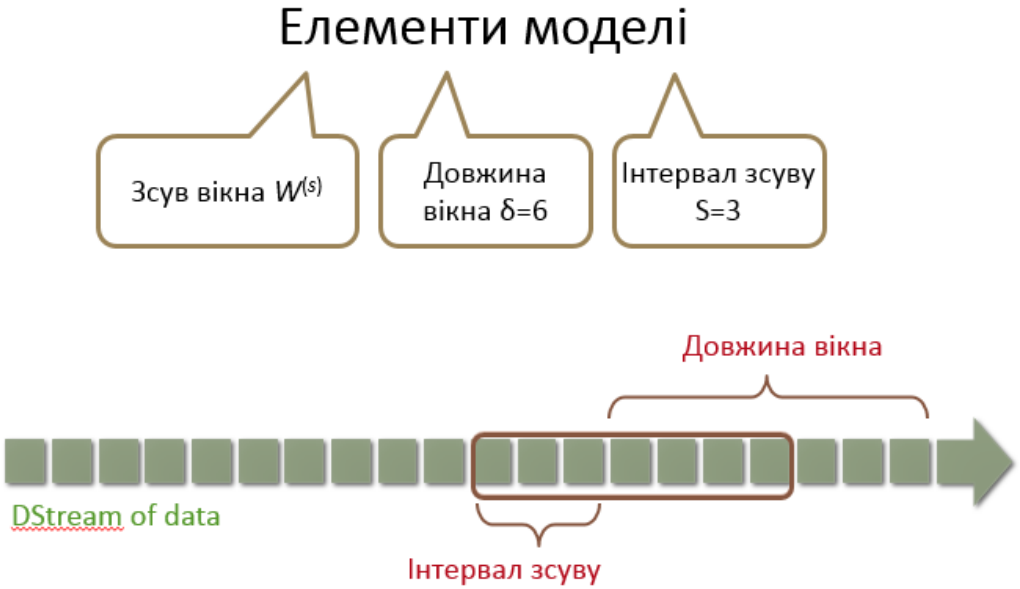

Рис. 2. Візуалізація елементів моделі потоку текстових даних.

Вікно W_i є часовим інтервалом фіксованого розміру δ, що починається в точці i. Вікно W_i містить документи потоку даних за інтервал часу $[i..i+\delta]$.

10.32782/2618-0340/2020.1-3.15

Позначимо за s інтервал зсуву (ковзання) вікна – інтервал фіксованого розміру.

Тоді $W_i^{(s)}$ – вікно фіксованого розміру δ, що починається в точці $i+s$.

Завдання обробки потоків текстових даних. Визначимо завдання, що ставляться перед обробкою текстових потоків даних.

Базові перетворення та попередня обробка кожного текстового елементу потоку даних:

- видалення стоп-слів та видалення нерелевантних символів;
- розбиття на речення;
- розбиття на токени;
- нормалізація – складання списку лем;
- визначення статистичних показників – кількість слів, токенів, речень, довжина тексту тощо;
- визначення мови текстового елементу;
- інші.

Завдання обробки текстового потоку даних в цілому:

- побудова моделі BOW [1];
- виділення ознак текстових елементів (TF-IDF та інших) [1];
- підрахунок косинуса подібності [1].

Інтелектуальний аналіз текстових потоків даних:

- класифікація даних (однозначна та багатокласова). Методами класифікації вирішується величезна кількість різних задач: сентимент аналіз [2], виявлення аномалій [1], виявлення елементів пропаганди, дезінформації та інші.
- кластеризація даних;
- реферування [1];
- перевірка правопису;
- інші.

Базові перетворення, попередня обробка текстів та завдання обробки текстового потоку даних в цілому описуються в статтях [1] та [5], тому на них зупинятись не будемо. Далі розглянемо такі задачі інтелектуального аналізу потоків текстових даних, як сентимент-аналіз та програмну реалізацію.

Визначення емоційного забарвлення потоків текстових даних. Над кожним елементом потоку даних d_j проводиться аналіз визначення емоційного забарвлення за допомогою класифікатора $\Phi(d)$. Де $\Phi(d)$ =–1, якщо негативне забарвлення, +1, якщо позитивне забарвлення, 0, якщо нейтральне забарвлення.

Визначимо емоційне забарвлення на рівні вікна $W^{(s)}$:

$$\Phi(W^{(S)}) = \frac{\sum_{k=0}^{n^{(S)}} \Phi(d_k, c)}{n^{(S)}}, \qquad (1)$$

де $n^{(s)}$ – кількість елементів в вікні $W^{(s)}$.

Якщо $\Phi(W^{(s)}) > 0$, вікно w зі зсувом s має позитивне забарвлення.

Якщо $\Phi(W^{(S)}) < 0$, вікно зі зсувом s має негативне забарвлення.

Якщо $\Phi(W^{(S)}) = 0$, вікно зі зсувом s має нейтральне забарвлення.

10.32782/2618-0340/2020.1-3.15

Позначимо $C=\{c1,c2,\ldots c|C|\}$ – множину можливих класів («негативне», «позитивне», «нейтральне» повідомлення).

На рис. 2 наведено приклад формування ковзного вікна при надходженні документів в потік зі швидкістю 1 документ/с. Так вікно W^0 містить документи від ID=1 до ID=11, а W^2, відповідно, документи від ID=3 до ID=13.

Рис. 3. Використання ковзного вікна для аналізу потоку текстових даних.

У якості класифікатора окремих текстових елементів було використано методи Gradient Boosting Tree та Random Forest, що показали найвищу точність у 79,42% та 75,74% відповідно на навчальній вибірці [7].

Виявлення аномалій в потоках текстових даних. Підхід до виявлення аномалій вже був описаний в статті [1]. Підхід будується на виконанні попередньої обробки тексту, реферуванні тексту, побудови BOW та визначення ознак TF-IDF. В той же час в Україні досить часто зустрічається ситуація, коли в межах одного потоку текстових даних (коментарі та повідомлення на сайтах, соціальних мережах, повідомлення в групах месенджерів) зустрічаються повідомлення різними мовами: найчастіше українською та російською. І тому необхідно врахувати дану ситуацію при обробці текстових потоків.

Архітектура програмного забезпечення Системи. В процесі дослідження виділено 4 підсистеми.

1. *Підсистема збору та транспортування повідомлень потоків даних*. Основний програмний компонент підсистеми – сервіс повідомлень Kafka, що використовується для обробки потоків даних, сервісів в режимі реального часу. Kafka підтримує горизонтальне масштабування, підтримує відмовостійкість.
2. *Підсистема аналізу текстових потоків*. Основний програмний компонент – Apache Spark, вибір якого було описано в [2].
3. *Підсистема зберігання результатів аналізу потоків даних*. Основний програмний компонент підсистеми - нереляційна база даних Elasticsearch, що має такі переваги, як гнучкість, масштабованість, високу продуктивність,

10.32782/2618-0340/2020.1-3.15

інтеграцію з засобами візуалізації. Альтернативно можна використати такі бази даних, як MongoDB, HBase.

4. *Підсистема візуалізації*. Основний програмний компонент підсистеми – Kibana, це служба візуалізації даних Elasticsearch, що допомагає створювати різні дашборди, налаштовувати форму візуалізації, формувати інтерактивні графіки.

На рис. 4 зображено обробку даних програмними компонентами Системи.

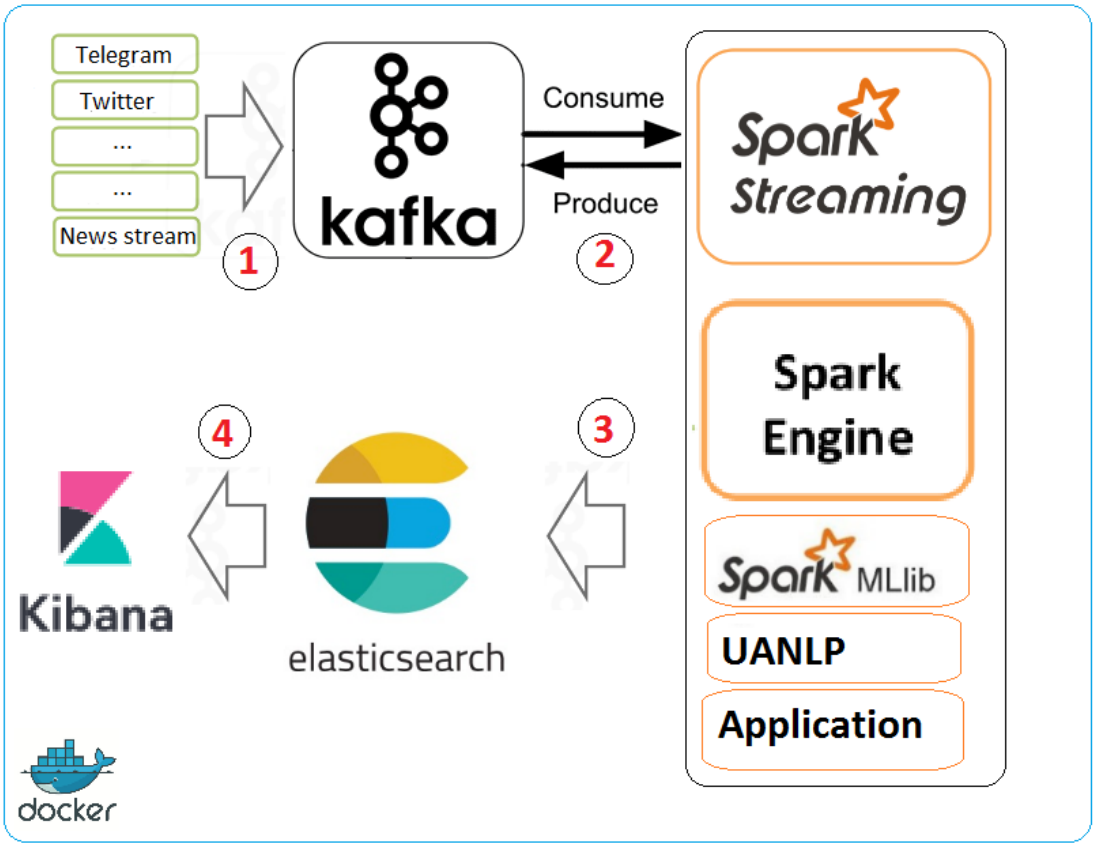

Рис. 4. Програмні компоненти системи аналізу текстових потоків даних.

Вхідні потоки текстових даних (від месенджерів, соціальних мереж, новинних сервісів) надходять на обробку в сервіс збору повідомлень Kafka (етап 1 на рис. 4), де відбувається накопичення даних для відправки на аналітичну обробку (етап 2 рис. 4).

Аналітична обробка виконується за допомогою технології швидкісної розподіленої обробки даних Apache Spark. Оскільки в даній технології відсутня NLP підтримка україномовних текстів, було додатково розроблено програмно бібліотеку UANLP[8] для обробки україномовних та російськомовних текстів на основі морфоаналізатора PyMorphy2 [9]. Бібліотека UANLP виконує всі необхідні базові перетворення та попередню обробку кожного текстового елементу потоку даних, а також побудову моделі BOW, виділення ознак текстових елементів TF-IDF, підрахунок косинуса подібності, автоматичне реферування, визначення мови тексту. Обробка потоків текстових даних виконується на основі компоненту Spark Streaming, що підтримує роботу з вікнами. Бібліотека Spark MLib та ML дозволяють використовувати засоби машинного навчання для аналітичної обробки потоків текстових даних, на основі яких виконується сентимент аналіз, виявлення аномалій, елементів пропаганди, дезінформації тощо.

10.32782/2618-0340/2020.1-3.15

Після аналітичної обробки дані завантажуються в БД Elasticsearch (етап 3, рис. 4), після чого автоматично потрапляють у сервіс візуалізації Kibana (етап, 3 рис. 4), де автоматично кожні *n*-секунд дані відправляються для візуалізації потоку в режимі реального часу [10] (рис. 5).

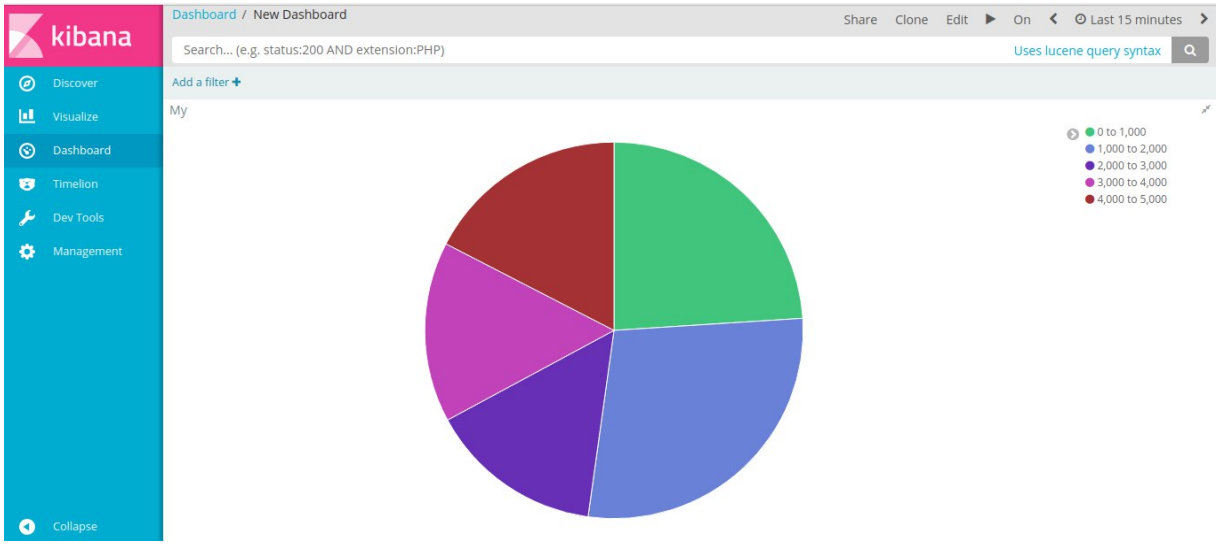

Рис. 5. Приклад візуалізації розподілу повідомлень потоку даних за довжиною текстового повідомлення.

Система використовується в навчальному процесі кафедри АСОІУ в рамках впровадження міжнародного проекту Erasmus+ project 'Establishing Modern Master-level Studies in Information Systems' (MASTIS) [11].

Така програмна архітектура має ряд переваг: гнучкість, підтримка різних джерел потоків даних, масштабованість, високопродуктивні обчислення, аналітичну обробку, візуалізацію обробки потоків даних в режимі онлайн.

Висновки

У досліджені наведено структуру та компоненти програмного забезпечення системи аналізу текстових потоків даних. Перевагою системи є підтримка обробки україномовних та російськомовних текстів, для чого було спеціально розроблено бібліотеку UANLP. Система дозволяє вирішувати наступні завдання обробки текстових потоків даних: виконувати базові перетворення та попередню обробку кожного текстового елементу потоку даних, побудову моделі BOW, виділення ознак текстових елементів TF-IDF, підрахунок косинуса подібності, автоматичне реферування, визначення мови тексту, інтелектуальний аналіз текстових потоків даних. Система, на відміну від аналогів, підтримує розроблену модель ковзного вікна для таких задач інтелектуального аналізу даних, як визначення емоційного забарвлення та виявлення аномалій. Крім того, система має наступні переваги: масштабованість, відмовостійкість, високопродуктивні обчислення, аналітичну обробку, візуалізацію обробки потоків даних в режимі онлайн.

Список використаної літератури

1. Олійник Ю. О., Афанасьєва О. Є., Аршакян Г. Д. Підхід до виявлення аномалій в потоках текстових даних. *Системні технології*. 2020. № 2(127). С. 126–139. DOI: https://doi.org/10.34185/1562-9945-2-127-2020-10

10.32782/2618-0340/2020.1-3.15

2. Tomashevskii V. M., Oliynik Y. O., Yaskov V. V., Romanchuk V. M. Realtime Text Stream Anomalies Analysis System. *Вісник Херсонського національного технічного університету*. 2018. № 3 (1). P. 361–365.
3. Oram A. Streaming Data. USA, Newton: O'Reilly Media, Inc., 2019. 28 p.
4. Степанюк Є. Ю., Олійник Ю. О. Дослідження методів аналізу тональності тексту. *Інформаційні системи та технології управління – ІСТУ-2019:* матеріали Всеукраїнської науково-практичної конференції молодих вчених та студентів. (м. Київ, 26 листопада 2019 р.), Київ: НТУУ «КПІ ім. Ігоря Сікорського», 2019. С. 32–39.
5. Гавриленко О. В., Олійник Ю. О., Ханько Г. В. Огляд та аналіз алгоритмів TEXT MINING. *Управління проектами, системний аналіз і логістика*. 2017. № 19. С. 15–23
6. Apache Spark Streaming. URL: http://spark.apache.org/docs/latest/streaming-programming-guide.html
7. Набір даних URL: https://github.com/dmytro-verner/sentiment-analysis-ukrainian-tweets
8. Ukrainian NLP Library for Apache Spark. URL: https://github.com/oliyura/UANLP/ [Назва з екрана].
9. Морфологійчний аналізатор pymorphy2. URL: https://pymorphy2.readthedocs.io/ [Назва з екрана].
10. Kibana. Your window into the Elastic Stack. URL: https://www.elastic.co/kibana [Назва з екрана].
11. Establishing Modern Master-level Studies in Information Systems URL: https://mastis.pro/[Назва з екрана]

References

1. Oliinyk, Yu. O., Afanasieva, O. Ye., & Arshakian, H. D. (2020). Pidkhid do vyiavlennia anomalii v potokakh tekstovykh danykh. *Systemni tekhnolohii*. **2**(127). С. 126–139. DOI: https://doi.org/10.34185/1562-9945-2-127-2020-10
2. Tomashevskii, V. M., Oliynik, Yu. O., Yaskov, V. V., & Romanchuk, V. M. (2018). Realtime Text Stream Anomalies Analysis System. *Visnyk Khersonskoho natsionalnoho tekhnichnoho universytetu*. **3**, Part 1, 361–365.
3. Oram A. Streaming Data. (2019). USA, Newton: O'Reilly Media, Inc.
4. Stepaniuk Ye. Yu., Oliinyk Yu. O. (2019). Doslidzhennia metodiv analizu tonalnosti tekstu. *Informatsiini systemy ta tekhnolohii upravlinnia – ISTU-2019: materialy Vseukrainskoi naukovo-praktychnoi konferentsii molodykh vchenykh ta studentiv*. (Kyiv, November 26, 2019 r), Kyiv: NTUU «KPI im. Ihoria Sikorskoho», pp. 32–39.
5. Havrylenko, O. V., Oliiny,k Yu. O., & Khanko, H. V. (2017). Ohliad ta analiz alhorytmiv TEXT MINING. *Upravlinnia proektamy, systemnyi analiz i lohistyka*. **19**, 15–23.
6. Apache Spark Streaming. URL: http://spark.apache.org/docs/latest/streaming-programming-guide.html
7. Nabir danykh Retrieved from: https://github.com/dmytro-verner/sentiment-analysis-ukrainian-tweets
8. Ukrainian NLP Library for Apache Spark. Retrieved from: https://github.com/oliyura/UANLP/ [Title from the screen].
9. Morfolohiichnyi analizator pymorphy2. Retrieved from: https://pymorphy2.readthedocs.io/ [Title from the screen].
10. Kibana. Your window into the Elastic Stack. Retrieved from: https://www.elastic.co/kibana [Title from the screen].

10.32782/2618-0340/2020.1-3.15

11. Establishing Modern Master-level Studies in Information Systems Retrieved from: https://mastis.pro/ [Title from the screen].

Олійник Юрій Олександрович – старший викладач кафедри автоматизованих систем обробки інформації і управління Національного технічного університету України «КПІ ім. Ігоря Сікорського», e-mail: oliyura@gmail.com, ORCID: 0000-0002-7408-4927.

10.32782/2618-0340/2020.1-3.15