РЕФЕРАТ

Магістерська дисертація: 64 с., 17 рис, 33 таб., 2 додатки, 38 джерел.

Актуальність теми: Сьогодні важливість обробки текстових даних стрімко збільшується. Це пов'язано з великою кількістю текстової інформації, доступної через Інтернет. Оскільки мільйони символів вмісту формуються щодня, людина не має фізичної здатності обробляти всю інформацію.

На українському ринку поки відсутні застосунки для виявлення аномалій. Українські медіа, наукова сфера та бізнес все ще не мають інструменту для виявлення аномальних даних в текстах рідною мовою, що робить ці сфери менш розвинутими ніж такі ж сфери, що працюють у англомовному середовищі.

Мета дослідження: покращення аналізу україномовних потокових текстових даних та виявлення в них аномалій в режимі реального часу

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

- обґрунтувати вибір методу виявлення аномалій;
- створити математичну модель вибраного методу виявлення аномалій;
- виконати програмну реалізацію методу виявлення аномалій;
- дослідити ефективність методу виявлення аномалій.

Об’єкт дослідження: потоки україномовних текстових даних.

Предмет дослідження: виявлення аномалій в потокових текстових даних.

Методи дослідження: методи text mining, методи інтелектуального аналізу даних.

Наукова новизна: Найбільш суттєвими науковими результатами магістерської дисертації є:

- розробка адаптованого методу Isolation Forest виявлення аномалій в потоках текстових даних з підтримкою української мови.

Практичне значення отриманих результатів визначається тим, що запропонований модифікований алгоритм Isolation Forest, який підтримує виявлення аномалій в україномовних даних.

Зв’язок роботи з науковими програмами, планами, темами: робота виконувалась на кафедрі автоматизованих систем обробки інформації та

управління 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 «Інформаційні системи та технології управління» (ІСТУ-2019)

Публікації: Наукові положення дисертації опубліковані в Афанасьєва О.Є Виявлення аномалій в потоках текстових даних/ О.Є. Афанасьєва, Ю.О. Олійник // Матеріали III всеукраїнської науково-практичної конференції молодих вчених та студентів «Інформаційні системи та технології управління» (ІСТУ-2019) – м. Київ: НТУУ «КПІ ім. Ігоря Сікорського», 20-22 листопада 2019 р.

Ключові слова: ПОТОКИ ДАНИХ, ВИЯВЛЕННЯ АНОМАЛІЙ, МЕТОД ІЗОЛЯЦІЙНОГО ЛІСУ, УКРАЇНОМОВНІ ДАНІ, ТЕКСТОВІ ДАНІ, ОБРОБКА ТЕКСТОВИХ ДАНИХ.