# ДОСЛІДЖЕННЯ МЕТОДІВ ОПТИМІЗАЦІЇ ПОШУКУ СХОЖИХ ТЕКСТІВ ІЗ ЗАСТОСУВАННЯМ ELASTICSEARCH

Мезенцева Д. В., Бабій А. С.
Харківський національний університет радіоелектроніки, Харків, Україна

У задачах інтелектуального аналізу тексту текстове уявлення має бути не лише ефективним, а й інтерпретованим, оскільки це дозволяє зрозуміти операційну логіку, що лежить в основі моделей інтелектуального аналізу даних. Традиційні методи векторизації тексту, такі як TF-IDF і Bag-of-words, ефективні та мають інтуїтивно зрозумілу інтерпретованість, але страждають від «прокляття розмірності» і не можуть розуміти сенсу слів. З іншого боку, сучасні розподілені методи ефективно визначають приховану семантику, але вимагають великих обчислювальних ресурсів та часу, а також не вистачає інтерпретованості.

Щоб застосувати різні методи машинного навчання та інтелектуального аналізу даних, необроблені документи необхідно перетворити на формат зрозумілий машині [1].

Першим кроком до того, щоб зробити текстові документи машиночитними, є векторизація, яка визначається як перетворення текстового документа на цифровий вектор, і є процесом вилучення ознак з тексту для виконання будь-яких завдань інтелектуального аналізу тексту та математичного вирішення проблем [2].

**Метою доповіді** є розробка методу оптимізації пошуку схожих текстів із застосуванням Elasticsearch, який представляє документ відповідно до інформації про концепти, що міститься в ньому. Запропонований метод створює концепти за допомогою кластеризації векторів слів (тобто вбудовування слів), і використовує частоти цих кластерів концептів для представлення векторів документів.

Щоб збагатити підсумкове подання документа, пропонується нова модифікована вагова функція для зважування концептів на основі статистики, отриманої з інформації вкладень слів.

Вектори, згенеровані за допомогою запропонованого методу, характеризуються інтерпретованістю, низькою розмірністю, високою точністю, а також низькими обчислювальними витратами при використанні у задачах кластеризації.

**Список літератури**

1. Нога Р. Аналітичний огляд методів та засобів опрацювання текстової інформації /Р.Нога, Н.Б.Шаховська //Вісник національного університету «Львівська політехніка». — 2011. — С. 323—332.

2. Tomas Mikolov. Efficient estimation of word representations in vector space /Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. — arXiv preprint arXiv:1301.3781. ICLR Workshop, 2013. — P. 1—12.