УДК 004.42

АЛГОРИТМІЧНІ МЕТОДИ ОБРОБКИ ТЕКСТОВИХ ДАНИХ ДЛЯ ОЦІНКИ ТА ПІДВИЩЕННЯ РЕЛЕВАНТНОСТІ ВЕБДОКУМЕНТІВ З ПИТАНЬ ЕНЕРГОЕФЕКТИВНОСТІ

Піменко Д.Ю. – гр. МгІТ-1-24, магістр, primenkoden@gmail.com
Астістова Т.І. – к.т.н., доцент, astistova@ukr.net
Київський національний університет технологій та дизайну

Метою роботи є дослідження алгоритмічних методів обробки текстових даних для підвищення релевантності вебдокументів, пов'язаних з енергоефективністю та відновлюваними джерелами енергії. Розроблені методи спрямовані на вдосконалення пошуку інформації в енергетичних базах даних, наукових публікаціях і звітах, що допоможе фахівцям швидше знаходити інноваційні рішення у сфері енергозбереження, енергоаудиту та оптимізації споживання ресурсів.

Обробка текстових даних (Text Data Processing) є ключовою складовою інтелектуальних систем підтримки рішень [1]. В енергетичній сфері це дозволяє аналізувати великі обсяги технічної документації, стандартів, державних програм з енергоефективності, досліджень та звітів.

Класичні методи, такі як TF-IDF, дозволяють визначити частотну вагу термінів (*«теплова ізоляція», «споживання електроенергії», «енергоаудит»*), тоді як сучасні алгоритми на основі глибинного навчання (Word2Vec, BERT, MUM) здатні розуміти контекст і семантичні зв'язки між поняттями (*«енергоефективні технології» – «зменшення втрат тепла» – «сталий розвиток»*) [2–3].

Використання таких методів дає змогу створювати системи, що підвищують точність тематичного пошуку у базах знань, присвячених енергозбереженню. Основою дослідження є патенти Google, зокрема «Context vectors for improving search results» [4] та «Information retrieval based on word relationships» [5], які описують принципи формування контекстних векторів для пошукових систем.

Для оцінки ефективності методів було сформовано добірку текстів із тематики енергоефективності: наукові статті, нормативні документи, енергетичні звіти та вебресурси про технології відновлюваної енергетики.

Було застосовано три підходи для аналізу релевантності запитів типу *«енергозбереження у промисловості», «оптимізація теплових систем», «енергоефективне будівництво»*:

1. TF-IDF – базова статистична оцінка частоти термінів.

2. Word2Vec – виявлення семантичних зв’язків між поняттями.

3. BERT embeddings – контекстна оцінка відповідності запиту та документа.

Результати показали, що моделі на основі нейронних мереж значно підвищують точність тематичного пошуку документів, особливо при роботі з термінами, які мають широкий контекст (наприклад, «енергетичний менеджмент» або «сталий розвиток»).

Висновки:

1. Розглянуті алгоритмічні методи обробки текстових даних можуть бути ефективно застосовані для створення інтелектуальних систем інформаційного пошуку в галузі енергоефективності.

2. Системи такого типу дозволяють автоматизувати аналіз великої кількості джерел, виявляти тренди у впровадженні енергозберігаючих технологій та підвищувати доступність знань у сфері сталого енергоспоживання.

3. Подальший розвиток дослідження може бути спрямований на створення моделі «енергетичного семантичного пошуку», яка поєднає машинне навчання та лінгвістичні методи для підтримки наукових і управлінських рішень у сфері енергетики та енергоефективності.

Список використаних джерел:

1. Manning, C. D., Raghavan, P., Schütze, H. *Introduction to Information Retrieval.* Cambridge University Press, 2022.
2. Mikolov, T. *Efficient Estimation of Word Representations in Vector Space.* arXiv:1301.3781.
3. Devlin, J. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* arXiv:1810.04805.
4. Google Patents: Context vectors for improving search results (US20170365495A1).
5. Google Patents: Information retrieval based on word relationships (US8661029B1).