УДК 004.85

ПІДВИЩЕННЯ ЯКОСТІ СЕМАНТИЧНОГО ПОШУКУ В БАЗАХ ЗНАНЬ ІЗ ВИКОРИСТАННЯМ ВЕКТОРНИХ ПОДАНЬ

Нестеренко В. В.
e-mail: vitalii.nesterenko@nure.ua
Науковий керівник – к.т.н., доцент Русакова Н.Є.
Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

This paper presents an applied evaluation of semantic search quality in a knowledge-base retrieval system. The study uses a reproducible pipeline that combines preprocessing, chunking, vector indexing, querying, and benchmark-based assessment across two domains. The system's performance was evaluated using a specially prepared set of queries and widely used metrics for assessing search effectiveness. Results demonstrate that contextual embeddings achieve superior ranking quality, while lexical methods ensure comprehensive coverage. The proposed approach supports iterative quality control and further development of semantic and hybrid retrieval systems for enterprise knowledge management.

Для сучасних інформаційних систем швидкий доступ до релевантних знань є критичним; у роботі [1] обґрунтовано проблему «лексичного розриву» в корпоративному пошуку та виконано теоретичне порівняння підходів до векторних подань (FastText, BERT, SBERT, OpenAI Embeddings). Лексичний пошук часто втрачає ефективність через синонімію та варіативність формулювань [2]. Мета роботи – підвищити якість семантичного пошуку шляхом порівняння моделей векторних подань у єдиному середовищі з однаковими умовами оцінювання.

У цій роботі акцент зроблено на практичній реалізації семантичного пошуку для корпоративної бази знань із фокусом на фактичній пошуковій видачі. Застосунок об’єднує підготовку даних, чанкінг, побудову векторних індексів і бенчмарк-оцінювання в одному відтворюваному контурі, що дає змогу швидко перевіряти вплив змін моделі або параметрів індексації на кінцеву якість видачі.

Дані нормалізуються, розбиваються на смислові фрагменти (чанки), далі векторизуються та індексуються у структурі найближчих сусідів [3]. Пошук виконується по чанках, що покращує потрапляння в локально релевантний контекст. Для кожної моделі зберігаються індекс, вектори та метадані.

Інтерфейс дозволяє оцінювати не лише наявність релевантного фрагмента, а й його позицію у топ-видачі, що критично для практичного використання. Це дає змогу поєднати кількісні метрики з якісним аналізом конкретних відповідей для реальних запитів користувача. На рисунку 1 показано приклад пошукової видачі для тематичного запиту.

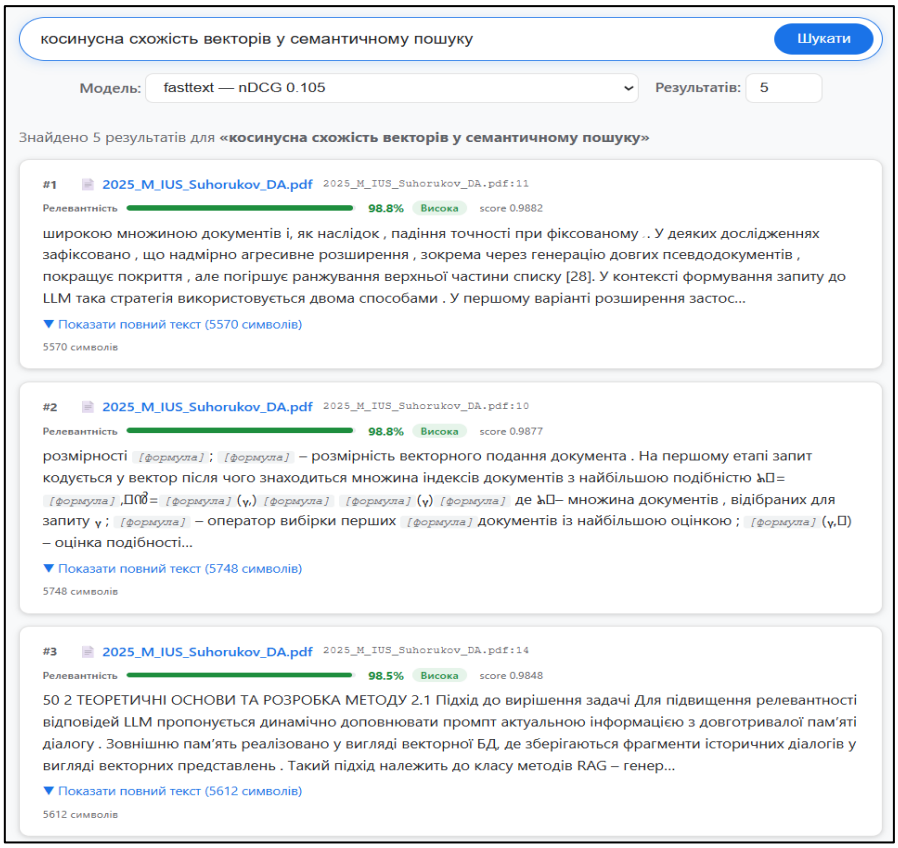

Рисунок 1 – Результати пошуку для запиту «косинусна схожість векторів у семантичному пошуку» (модель FastText, топ-5)

Для AI-запиту «косинусна схожість векторів у семантичному пошуку» у верхніх позиціях отримано 98.8%, 98.8% і 98.5% локальної релевантності; для запиту з іншої предметної області («кінематика COREXY у системах FDM-друку») – 63.0%, 57.5%, 47.9% у перших трьох позиціях. На рисунку 2 подано зведені результати бенчмарк-оцінювання.

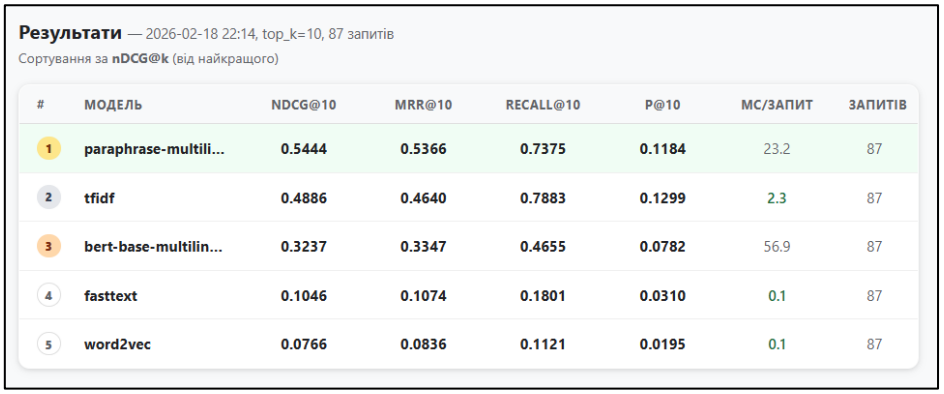

Рисунок 2 – Сторінка бенчмарку з відображенням метрик для кількох моделей

Якість оцінено на benchmark-наборі (87 запитів, top_k=10) за метриками nDCG@10, MRR@10, Recall@10 і P@10. Найкращий

nDCG@10 має paraphrase-multilingual-MiniLM-L12-v2 (0.5444), далі TF-IDF (0.4886) і BERT (0.3237); fastText і word2vec показали суттєво нижчі результати – 0.1046 та 0.0766 відповідно. За Recall@10 лідирує TF-IDF (0.7883), що свідчить про здатність лексичного підходу знаходити релевантні документи навіть за відсутності семантичного розуміння. За MRR@10 перевагу має SBERT (0.5366) проти TF-IDF (0.4640), що вказує на кращу позицію найрелевантнішого документа [4]. Результати BERT виявилися проміжними між SBERT та класичними ембедингами, що пов'язано з відсутністю спеціальної адаптації до задач пошуку через Siamese-архітектуру. Слабкі результати fastText і word2vec пояснюються обмеженим контекстом: ці моделі не враховують порядок слів і довгі залежності, тому втрачають семантичні нюанси складних технічних запитів. Метрика Precision@10 також підтверджує перевагу SBERT у точності верхніх позицій завдяки кращому розумінню синонімії та перефразувань. Отже, контекстна модель краще ранжує верхні позиції завдяки семантичному розумінню, а TF-IDF забезпечує високу повноту через точне лексичне співставлення. Для сценаріїв, де користувач очікує швидку правильну відповідь у топ-3, оптимальним є SBERT; для повного аналізу всіх можливих збігів – доцільніший TF-IDF. Результати вказують на доцільність гібридного підходу, який поєднує переваги контекстних і лексичних методів для максимізації як точності, так і повноти пошуку.

Запропонований підхід забезпечує відтворюване порівняння моделей і практичний контроль якості пошуку в єдиному контурі «дані – індекс – запит – оцінювання». Отримані результати можуть бути використані для подальшого розвитку семантичного і гібридного пошуку.

Список використаних джерел:

1. Нестеренко В. В., Русакова Н. Є. Порівняльний аналіз ефективності моделей векторних ембеддінгів для задач семантичного пошуку в корпоративних базах знань // Innovative Research in Science and Economy : матеріали II Міжнар. наук.-практ. конф. С. 127.

2. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. Cambridge : Cambridge University Press, 2008. 482 p.

3. Johnson J., Douze M., Jégou H. Billion-scale similarity search with GPUs // IEEE Transactions on Big Data. 2019. Vol. 7, No. 3. P. 535–547.

4. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP). 2019. P. 3982–3992.