УДК 004.89:004.63

ЗАСТОСУВАННЯ МОДЕЛЕЙ NLP ДЛЯ ОПТИМІЗАЦІЇ ПОШУКУ В ФАЙЛОВИХ СХОВИЩАХ

Коваль Г.К., Гриньова О.Є.
e-mail: hlib.koval@nure.ua, olena.hrynova@nure.ua
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна

This research implements AI-powered search for file storage platforms using vector search and NLP models. The system analyzes and indexes documents, enabling semantic searches based on content and context. It uses Sentence Transformers for vector representations and Pinecone for efficient storage and retrieval. This innovation overcomes traditional search limitations, enhancing document management efficiency across various domains. The project aims to improve user productivity by providing more relevant search results and streamlining file retrieval processes in large-scale document repositories.

Основною проблемою сучасного цифрового середовища є управління величезними обсягами даних, які постійно зростають. Це вимагає спеціальної інфраструктури та технології для ефективного зберігання та обробки. Крім того, документи у сховища надходять у різних форматах, що ускладнює їх структуризацію та аналіз. Швидкість доступу до документів, а саме пошук «голки» в хаотичній «копиці» файлів та тек, також створює виклики. Традиційні методи пошуку, такі як пошук за назвою файлу, мають серйозні обмеження: результат пошуку залежить від точності введення запиту, чутливості до регістру, проблем із пробілами та спеціальними символами. Крім того, старі методи не дозволяють робити семантичний пошук.

Технології ШІ мають потенціал для застосування у критичних галузях, таких як інформаційна безпека та правоохоронна діяльність. Наприклад, системи ШІ можуть ефективно відслідковувати загрози та аналізувати великі обсяги даних для виявлення потенційно небезпечних осіб, шляхом виявлення документів з недозволеним контентом.

Отже, актуальність цієї роботи обґрунтовується необхідністю розробки ефективних методів управління зростаючими обсягами даних, потенціалом технологій штучного інтелекту для вирішення складних завдань у різних сферах, а також потребою в підвищенні продуктивності та інноваційності в науковій та практичній діяльності [1].

Мета роботи полягає в аналізі та впровадженні сучасних методів інтелектуального пошуку файлів з використанням ШІ для підвищення ефективності управління документами.

Проєкт спрямований на розробку вебзастосунку для зберігання, редагування та доступу до документів за пошуковим запитом з використанням ШІ, що підвищить релевантність результатів пошуку.

У дослідженні розглядаються два основні підходи до інтелектуального пошуку файлів: семантичний аналіз з використанням моделей обробки природної мови (NLP) та векторний пошук з перетворенням текстового вмісту у векторні представлення.

В роботі розглянуто два основні підходи до інтелектуального пошуку файлів. Один із методів передбачає пошук за ключовими словами та семантичний аналіз. Це можна зробити завдяки використання моделей обробки природної мови (NLP), таких як BERT або GPT, для аналізу текстового вмісту файлів і визначення їхньої релевантності до запиту. Це дозволяє знаходити файли навіть за неточними або контекстно пов’язаними запитами. Друга стратегія передбачає векторний пошук, що дозволяють знаходити документи не за точним збігом ключових слів, а за їхнім значенням у контексті. Для цього текстовий вміст файлів перетворюється у векторні представлення за допомогою моделей, таких як Word2Vec, FastText або Sentence Transformers. Ці вектори зберігаються в оптимізованих базах для швидкого пошуку, наприклад, FAISS або Pinecone, що дає змогу ефективно знаходити документи з близьким вмістом.

Технології та інструменти.

BERT (Bidirectional Encoder Representations from Transformers) – це модель штучного інтелекту (МШІ), розроблена Google, яка забезпечує двонаправлений аналіз контексту слів у реченні. Завдяки цьому МШІ ефективно виконує завдання семантичного аналізу та класифікації тексту. GPT (Generative Pre-trained Transformer) – це модель від OpenAI, яка працює в односпрямованому режимі (зліва направо) і використовується для генерації текстів, автоматичного доповнення запитів та інших завдань.

Запропоновано використати бібліотеку Sentence Transformers, яка призначена для створення компактних багатовимірних векторів тексту (sentence embeddings), для зберігання змісту текстового документа [2]. Це оптимальне рішення для семантичного пошуку.

Для зберігання та швидкого пошуку векторних представлень була обрана хмарна платформа Pinecone. Pinecone забезпечує масштабованість і високу швидкість роботи з великими наборами даних, що робить її ефективним рішенням для задач семантичного пошуку.

Для програмної реалізації інтелектуального пошуку файлів була застосована друга стратегія, а саме векторний пошук. Початковим етапом є вилучення контексту з документів за допомогою бібліотек python-pptx, python-docx, pdfplumber та стандартного контекстного менеджера мови Python. Розбиття великих документів на менші фрагменти. Після вилучення контексту здійснюється перевірка його довжини. У разі, якщо документ містить велику кількість сторінок, його текстовий вміст розбивається на менші фрагменти (чанки). Конвертація фрагментів у векторні представлення. Ці фрагменти конвертуються у векторні

представлення за допомогою моделі all-MiniLM-L6-v2 із бібліотеки Sentence Transformers. Зберігання векторів у базі даних Pinecone. Отримані векторні представлення зберігаються у векторній базі даних Pinecone, що забезпечує ефективний та швидкий пошук. Обробка користувацьких запитів та порівняння з наявними векторами. Фінальний етап передбачає обробку користувацького запиту, який також перетворюється у векторне представлення та надсилається до Pinecone для порівняння з наявними векторами. У результаті система повертає від 7 до 10 документів із найвищим рівнем відповідності або, за відсутності релевантних збігів, не знаходить жодного схожого документа.

Якщо порівнювати ці два підходи, то вони мають суттєві відмінності в принципах роботи та ефективності (таблиця 1). Перший підхід базується на класичному пошуковому індексуванні, де документи знаходяться за точним або частковим збігом слів, що забезпечує швидкість, але не гарантує розуміння контексту. Векторний пошук, навпаки, використовує нейромережеві моделі для перетворення тексту в багатовимірні вектори, що дозволяє знаходити документи за змістовною схожістю, навіть якщо ключові слова відсутні. Однак він вимагає більше обчислювальних ресурсів і складніший у впровадженні. Тому для задач, де важлива швидкість і точний збіг, підходить перший метод, а коли потрібно враховувати семантику та контекст, краще використовувати векторний пошук.

Таблиця 1 – Порівняльний аналіз класичного індексування та векторного пошуку

Критерій	Класичне індексування	Векторний пошук
Розуміння контексту	Низьке	Високе
Точність	Середня	Дуже висока
Швидкість	Висока	Середня
Ресурсоємність	Низька	Висока
Простота впровадження	Висока	Низька

Використання моделей BERT, GPT або Sentence Transformers у поєднанні з платформами, на кшталт Pinecone, дозволяє знаходити документи навіть за нечіткими запитами або за змістовною схожістю. Ця робота має потенціал для значного покращення ефективності роботи з документами у різних галузях.

Список використаних джерел:

1. М.П. Дудник, С.Г. Удовенко, Л.Е. Чала, М.М. Соколовська. Нейромережева технологія багатомовної класифікації електронних текстів // Біоніка інтелекту. – 2021. – Вип. 2 (97). – С. 3-12

2. SentenceTransformers Documentation | Sentence Transformers documentation. SentenceTransformers Documentation: URL : https://sbert.net/ (дата звернення 25.02.2025).