Лабораторна робота №3

Гібридний пошук та робота з метаданими

Мета роботи: навчитись поєднувати семантичний пошук із традиційними методами фільтрації для створення точних та релевантних пошукових систем.

Науковий аспект.

Доведення переваг гібридного підходу для запитів з унікальними назвами або артикулами.

Стек технологій:

LangChain - *фреймворк для побудови ланцюжків ШІ.*

ChromaDB - *для зберігання векторів та метаданих.*

Sentence-Transformers - *для генерації ембедингів.*

Зміст роботи

Завдання 1. *Дослідити зміну конфігурація метаданих. Визначити структуруи додаткових полів (рік, жанр, рейтинг).*

Реалізація гібридних запитів. Використання операторів фільтрації (eq, gt, in) разом із векторним пошуком.

Дослідження реранкінгу. Аналіз того, як зміна порядку видачі впливає на користувацький досвід.

Аналіз ефективності. Порівняння швидкості виконання запитів з фільтрами та без них.

Завдання 2. ***Дослідження та апробація прикладного кейсу семантичного аналізу.***

Формалізація параметрів моделі:

– Обґрунтувати вибір полів тексту для векторизації (High-dimensional feature space).

– Визначити структуру метаданих для реалізації механізмів жорсткої фільтрації (Post-filtering) відповідно до обраної наукової задачі.

Проектування та конфігурація пошукового інтерфейсу:

– Розробити стратегію семантичних запитів.

– Реалізувати алгоритм фільтрації на рівні векторної БД для оптимізації релевантності видачі.

Експериментальне тестування:

– Провести серію контрольних запитів (Inference) для верифікації роботи моделі.

Комплексний аналіз та валідація результатів:

– *Оцінка релевантності.* Виконати аналіз числових значень косинусної відстані (L2 distance / Cosine similarity) між запитом та результатами.

– *Компаративний аналіз.* Зіставити емпіричні результати з теоретично очікуваними (Expected vs Observed). Оцінити наявність семантичного шуму.

Висновки. Сформулювати науково обґрунтований висновок щодо ефективності використання щільних векторних представлень (Dense Embeddings) для розв'язання поставленої прикладної задачі.

Індивідуальні завдання для дослідження

№	Назва науково-прикладного завдання	Об'єкт дослідження	Простір векторизації
1	Семантичний аналіз атмосфери та сеттингу	Дослідження здатності моделі ідентифікувати абстрактні концепти та стилістику (напр. «похмура антиутопія», «нуар»).	description
2	Багатофакторна фільтрація у векторному просторі	Оцінка точності семантичного пошуку при накладанні часових обмежень (release_year) через метадані.	description + metadata
3	Ідентифікація акторських амплуа	Аналіз семантичної близькості професійних профілів акторів на основі їхнього попереднього досвіду (кастингу).	cast
4	Семантична бінарна класифікація	Експериментальне визначення близькості неструктурованого контенту до полярних категорій («Educational» vs «Violent»).	description
5	Контекстуальна рекомендаційна система	Побудова моделі подібності «item-to-item» на основі семантичного ядра сюжетної лінії без урахування тегів.	description

6	Статистично-семантичний аналіз хронометражу	Кореляція між статистичними показниками (df.describe()) та семантичними маркерами короткої тривалості.	duration
7	Крос-культурний семантичний аналіз	Дослідження відображення національного менталітету та культурних особливостей у векторних представленнях контенту.	country + description
8	Валідація безпеки контенту (Safety AI)	Проектування фільтра безпеки для дитячого контенту на основі семантики та вікових рейтингів.	rating + description
9	Діахронічний аналіз семантики (Еволюція)	Дослідження трансформації семантичного значення концептів (напр. «герой») у медіа-дискурсі різних десятиліть.	release_year + description
10	Пошук семантичних колізій (Дублікати)	Виявлення прихованих дублікатів та франшиз через знаходження максимальних значень косинусної схожості в масиві даних.	description

Контрольні запитання

1. Чим відрізняється *Pre-filtering* від *Post-filtering* у векторних базах даних?
2. Як *Cosine Distance* корелює з релевантністю при гібридному пошуку?
3. Які переваги дає використання *Cross-Encoders* для реранкінгу порівняно з *Bi-Encoders*?
4. Що таке *Boolean filtering* у контексті *ChromaDB* та які оператори (and, or, gt) він підтримує?
5. Як метадані допомагають вирішити проблему семантичного шуму в дуже великих колекціях?
6. Поясніть концепцію *Reciprocal Rank Fusion (RRF)* для об'єднання результатів різних типів пошуку.
7. Чому важливо індексувати метадані окремо від векторних представлень?
8. Як вибір полів для фільтрації впливає на швидкість (*latency*) запиту?
9. У яких випадках семантичний пошук може програвати звичайному фільтру за метаданими?
10. Як реалізувати багаторівневий пошук (наприклад: фільтрація -> семантичний пошук -> реранкінг)?

Самостійна робота до лабораторної роботи №3

Завдання: Реалізуйте логіку, яка автоматично підвищує ранг фільмів з високим рейтингом при однакових показниках семантичної схожості.

Порівняйте точність пошуку за запитом «Sci-fi movies» при використанні чистої семантики та фільтрації за полем genre.