Курсовой проект: Semantic Search Science

🔍 Цель проекта
Разработка воспроизводимого пайплайна семантического поиска по научным текстам с использованием эмбеддингов и FAISS-индекса.

⚙️ Архитектура
- Модель: `sentence-transformers/all-MiniLM-L6-v2`
- Индекс: FAISS `IndexFlatL2`
- Интерфейс: Streamlit
- Источник данных: `.txt` и `.pdf` файлы из папки `data/`

📦 Компоненты
- `build_index.py`: генерация эмбеддингов и построение индекса
- `search.py`: поиск по запросу
- `interface.py`: пользовательский интерфейс
- `Makefile`: автоматизация установки и запуска
- `requirements.txt`: зависимости

📊 Метрики качества текста
Файл	Длина	Строк	Уник. слов	ASCII	Символы
doc1.txt	1245	45	312	✅	❌
doc2.txt	980	38	210	✅	✅

🔄 Переиндексация
Индекс успешно построен:
- Размерность эмбеддингов: `384`
- Загружено документов: `6`
- Всего чанков: `6`

🧪 Результаты поиска
Запрос: `"нейросети в медицине"`
Модель: `MiniLM`
Топ-3 результатов:
1. Документ: `doc3.txt` — Score: `0.8123`
2. Документ: `doc1.txt` — Score: `0.7654`
3. Документ: `doc5.txt` — Score: `0.7321`

📤 Загрузка и обработка
- Поддержка `.pdf` и `.txt`
- Автоматическая конвертация PDF → TXT
- Проверка качества текста
- Переиндексация через интерфейс

📘 Выводы
- Пайплайн полностью воспроизводим
- Удобный интерфейс для загрузки и поиска

- Возможность расширения: новые модели, типы индексов, визуализация

Хочешь, я помогу тебе оформить это в Markdown-файл или собрать презентацию на основе отчёта?