**Тема:** Система семантического поиска научных публикаций по большим базам научных публикаций на основе информации о внешнем цитировании с использованием нейросетевых моделей.

**Общее представление:** Система, которая опирается не только на текстовое содержание публикаций, но и на взаимосвязь между статьями на основе цитирований, а также контекстное значение запроса, что позволит пользователю находить наиболее релевантные публикации в интересующей его области.

**Предмет реализации:** Web-приложение.

**Цель:** Создать web-приложение для семантического поиска научных публикаций, которое будет учитывать не только текстовое содержание, но и взаимосвязи между статьями через внешние цитирования. Исследовать качество нейросетевой модели.

**Задачи:**

1. Анализ предметной области.
   1. Изучение существующих систем.
   2. Изучение моделей для семантического поиска (модели на базе BERT).
   3. Исследование использования данных о цитировании для улучшения релевантности поиска.
   4. Оценка возможностей нейросетевых моделей для обработки текстов и анализа взаимосвязей через цитирования.
2. Определение требований к системе.
   1. Определение основных функциональных требований к веб-приложению.
   2. Выбор технологий и стека разработки приложения.
   3. Выбор нейросетевой модели для векторизации с сохранением семантики.
3. Проектирование архитектуры.
   1. Проектирование архитектуры приложения (взаимодействие между ключевыми частями)
   2. Проектирование структуры клиентской базы данных.
   3. Проектирование структуры базы данных для хранения текстов публикаций?
   4. Разработка API для взаимодействия с внешними системами публикаций?
4. Сбор и разметка данных.
   1. Сбор текстов публикаций и информации о цитировании.
   2. Объединение данных из различных источников и их стандартизация.
   3. Предобработка данных перед обучением модели. Разметка и предобработка текстов (Например токенизация, векторизация, нормализация).
5. Обучение модели. (Начинаем что-то не абстрактное:D)
   1. Разработка или адаптация существующей нейросетевой модели для обработки текстов публикаций и анализа данных о цитировании.
   2. Обучение и тестирование модели на различных датасетах.
   3. Оценка качества получившейся модели.
   4. Дообучение модели.
   5. Сравнение качества различных моделей?
6. Разработка веб-приложения.
   1. Разработка серверной части (Python, Go) с использованием фреймворков (Django, Flask, FastAPI, GIN).
   2. Создание пользовательского интерфейса.
   3. Интеграции обученной нейронной сети и обеспечение доступа при помощи API.
7. Тестирование и отладка.
   1. Провести тестирование работы приложения на различных наборах данных.
   2. Исправить ошибки, выявленные на этапе тестирования, и улучшить производительность системы.
8. Оптимизация и финальная доработка.
9. Документация?
   1. Для разработчиков
   2. Для пользователей

**План реализации:**

1. Подготовительный этап:

* Анализ существующих решений.
* Определение требований к системе, выбор инструментов и технологий для разработки.

1. Проектирование системы:

* Разработка общей архитектуры приложения.
* Определение структуры базы данных и API для взаимодействия с внешними базами публикаций.

1. Разработка и обучение нейросетевой модели:

* Сбор и разметка данных.
* Обучение и тестирование модели на наборе данных научных публикаций.

1. Реализация web-приложения

* Разработка бэкенда (серверная часть с использованием фреймворков, таких как Django, GIN).
* Создание фронтенда
* Интеграция с нейросетевой моделью и базами данных научных публикаций.

1. Тестирование системы

* Оценка работы модели по метрикам точности, полноты и релевантности поиска.
* Исправление ошибок, доработка.