ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

**ПЕРВЫЙ МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ**

**УНИВЕРСИТЕТ ИМЕНИ И.М.СЕЧЕНОВА**

**МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**(СЕЧЕНОВСКИЙ УНИВЕРСИТЕТ)**

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**

на тему: Чат-бот для автоматизации выдачи научных статей

наименование темы

Направление подготовки 09.03.02 Информационные системы и технологии

**«Допущен к защите»** **Исполнитель:**

Исполнитель: Протокол № от

Протокол № от Рубанов Матвей Валентинович

Ф.И.О, подпись

(гр.636-01, очная форма подготовки)

**Заведующий кафедрой:**  **Научный руководитель:**

Лебедев Георгий Станиславович Савина Екатерина Александровна

к.т.н., профессор, профессор

подпись Ф.И.О., уч. степень, уч. звание, подпись

**«Прошел защиту»**

Оценка

**Москва, 2025**

**Актуальность:** Рост количества научных публикаций создает проблему информационного шума. Традиционные системы поиска, основанные на ключевых словах, не всегда эффективно понимают смысл запроса исследователя, что приводит к потере времени и пропуску релевантных работ.

**Цель работы:** Разработка прототипа чат-бота для семантического поиска научных статей, повышающего релевантность выдачи за счет использования моделей машинного обучения.

**Задачи работы:**

1. Провести анализ предметной области и существующих решений для поиска академической информации.
2. Сформулировать требования к прототипу чат-бота и выдвинуть гипотезу о повышении релевантности с помощью ML.
3. Разработать и реализовать математическую модель семантического поиска на основе векторных представлений текста.
4. Провести сравнительные испытания разработанной модели с базовым методом (TF-IDF) и оценить ее эффективность.

**Объект исследования:** Процесс информационного поиска в массиве научных текстов.

**Предмет исследования:** Методы и алгоритмы семантического поиска на основе машинного обучения.

**Методы исследования:** Системный анализ, сравнительный анализ, методы машинного обучения (включая векторное представление текста и анализ косинусного сходства), объектно-ориентированное программирование на Python.

**Научная новизна:** Адаптация и сравнительный анализ современных моделей векторизации текста (Sentence Transformers) для задачи семантического поиска в научных базах данных в рамках чат-бота.

**Практическая значимость:** Результаты работы могут быть использованы для создания плагинов к существующим научным порталам или как основа для standalone-сервиса, помогающего исследователям в их работе.

**Положения, выносимые на защиту:**

Семантический поиск на основе Sentence Transformers превосходит по релевантности поиск на основе TF-IDF для задачи подбора научных статей.

Прототип чат-бота, реализующий данную модель, является работоспособным и демонстрирует применимость выбранного подхода.

Выбранный стек технологий (Python, scikit-learn, sentence-transformers) является эффективным для быстрого прототипирования подобных систем.

**4. Основная часть**

**Глава 1. Аналитический обзор и постановка задачи**

1.1. Анализ предметной области: Проблемы современных исследователей.

1.2. Анализ существующих решений: Сравнение Google Scholar, PubMed, Semantic Scholar. (Таблица: Функции, плюсы, минусы).

1.3. Формулировка требований к системе:

* Функциональные: Прием текстового запроса, семантический поиск по базе, вывод списка статей.
* Нефункциональные: Время ответа < 2 сек., точность (релевантность).

1.4. Выводы по главе и постановка задачи проектирования.

**Глава 2. Проектирование модели семантического поиска**

\*2.1. Разработка теоретической модели: Выбор архитектуры (Кодировщик -> Векторы -> Косинусное сходство).\*

\*2.2. Выбор и обоснование технологического стека: Python, pandas, scikit-learn, sentence-transformers, numpy.\*

\*2.3. Проектирование процесса работы чат-бота (UML Activity Diagram).\*

2.4. Выводы по главе.

**Глава 3. Реализация и испытания**

3.1. Реализация ключевых модулей: (Как раз код, который мы написали выше, будет здесь с пояснениями).

3.2. План испытаний: Сравнение двух методов на наборе тестовых запросов.

3.3. Анализ результатов испытаний: (Привести результаты выполнения нашего кода, таблицы, графики схожести).

\*3.4. Выводы по главе: Гипотеза подтвердилась, модель готова к интеграции в чат-бота.\*

**5. Заключение**  
(Кратко повторить задачи и как они были решены, подтвердить достижение цели и положений на защиту).

**6. Список использованных источников**

1. Devlin, J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019.
2. Reimers, N., Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. – 2019.
3. Агравал, А. Информационный поиск в Web. – М.: Диалектика, 2020. – 352 с.
4. Manning, C. D., Raghavan, P., Schütze, H. Introduction to Information Retrieval. – Cambridge University Press, 2008.
5. Официальная документация библиотеки scikit-learn [Электронный ресурс]. – URL: <https://scikit-learn.org/stable/>

**7. Приложения**

* Приложение А: Исходный код прототипа.
* Приложение Б: Скриншоты работы программы.