НПОУ «ЯКУТСКИЙ КОЛЛЕДЖ ИННОВАЦИОННЫХ ТЕХНОЛОГИЙ» (НПОУ «ЯКИТ»)

Отделение информационных технологий и туризма

Специальность 09.02.07 «Информационные системы и программирования»

КУРСОВАЯ РАБОТА

по дисциплине: Языки программирования

по теме:

Создание десктопного приложения для семантического анализа русского языка с использованием библиотеки Hugging Face Transformers на Python

Исполнитель студент гр. КИСП-23(1)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Л.П. Лебедева

подпись, дата

Руководитель,

преподаватель\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д.А.Фёдоров

подпись, дата

Якутск, 2025

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc194747538)

[1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 5](#_Toc194747539)

[1.1 Терминология 5](#_Toc194747540)

[1.2 Распределение ролей и работы 6](#_Toc194747541)

[1.3 Стек технологий 7](#_Toc194747542)

[2.ПРАКТИЧЕСКАЯ ЧАСТЬ 8](#_Toc194747543)

[2.1. Архитектура приложения 8](#_Toc194747544)

[2.2.Разработка проекта по ролям 9](#_Toc194747545)

[2.3.Контроль выполнения плана 10](#_Toc194747546)

[ЗАКЛЮЧЕНИЕ 11](#_Toc194747547)

# ВВЕДЕНИЕ

Актуальность создания десктопного приложения для семантического анализа текста обусловлена несколькими факторами:

1. Рост объемов текстовой информации: В условиях цифровизации и доступности информации на русском языке возрастает потребность в инструментах, позволяющих эффективно обрабатывать и анализировать текстовые данные.
2. Необходимость в автоматизированных решениях: Бизнес и научные исследования требуют автоматизации процессов анализа текстов, что позволяет сократить время и ресурсы, затрачиваемые на ручной анализ.
3. Развитие технологий обработки естественного языка: Современные модели машинного обучения, такие как трансформеры, значительно увеличили точность и скорость семантического анализа, что делает их применение актуальным и необходимым.

Основная цель проекта заключается в разработке десктопного приложения для семантического анализа русского языка, использующего библиотеку Hugging Face Transformers. Приложение должно обеспечивать пользователям возможность быстро и эффективно анализировать текстовые данные, выявляя их смысл и настроение.

Задачи проекта

1. Исследование теоретических основ семантического анализа: Изучить ключевые аспекты семантического анализа, включая методы и подходы, применяемые в обработке естественного языка.
2. Выбор технологий и инструментов: Определить оптимальный стек технологий для реализации приложения, включая язык программирования, библиотеки и инструменты для создания интерфейса.
3. Разработка интерфейса пользователя: Создать интуитивно понятный графический интерфейс, который позволит пользователям вводить текст и получать результаты анализа.
4. Имплементация модуля обработки текста: Реализовать функциональность семантического анализа с использованием предобученных моделей из библиотеки Hugging Face Transformers.
5. Тестирование и отладка приложения: Провести тестирование приложения, выявить и исправить возможные ошибки, а также убедиться в корректности работы всех функций.
6. Документация и представление результатов: Подготовить документацию по использованию приложения и представить результаты работы проекта заинтересованным сторонам.

Объектом исследования является процесс семантического анализа текстов на русском языке, а также применяемые в этом процессе алгоритмы и технологии обработки естественного языка.

Предметом исследования выступает десктопное приложение, реализующее семантический анализ текста с использованием библиотеки Hugging Face Transformers.

1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ
   1. Терминология

В данной подглаве будут подробно рассмотрены ключевые термины и концепции, связанные с семантическим анализом и обработкой естественного языка (NLP), которые будут полезны для понимания основ нашего проекта.

Семантический анализ — это процесс извлечения значений и смыслов из текстовых данных. Он включает в себя понимание контекста, смысловых связей между словами и предложениями. Семантический анализ может быть использован для решения различных задач, таких как:

Анализ тональности: Определение эмоциональной окраски текста (положительная, отрицательная или нейтральная).

Извлечение сущностей: Определение имён, организаций, мест и других значимых элементов в тексте.

Понимание контекста: Учет контекста, в котором используются слова, для более точного анализа значений.

Обработка естественного языка (NLP) — это область искусственного интеллекта, занимающаяся взаимодействием между компьютерами и человеческим языком. Цели NLP включают:

Понимание текста: Создание моделей, способных интерпретировать и анализировать текст на естественном языке.

Генерация текста: Создание новых текстов на основе заданных параметров или примеров.

Перевод текста: Автоматический перевод текста с одного языка на другой.

NLP охватывает множество методов и подходов, включая машинное обучение и глубокое обучение.

Токенизация — это процесс разделения текста на отдельные элементы, называемые токенами. Эти токены могут быть словами, фразами или символами. Токенизация является одним из первых шагов в обработке текста, поскольку она позволяет модели работать с отдельными частями текста.

Модели трансформеров — это современные архитектуры нейронных сетей, которые обеспечивают высокую точность в задачах NLP. Они основаны на механизме внимания, который позволяет моделям учитывать контекст слов в предложении, независимо от их расположения. Основные модели трансформеров:

BERT (Bidirectional Encoder Representations from Transformers): Модель, разработанная Google, которая позволяет учитывать контекст слов слева и справа от целевого слова, что значительно улучшает точность семантического анализа.

GPT (Generative Pre-trained Transformer): Модель от OpenAI, которая может генерировать текст и выполнять задачи, такие как ответ на вопросы и создание описаний, на основе заданного контекста.

Эти модели были предобучены на больших объемах текстовых данных и могут быть дообучены на специфичных для задачи данных, что делает их очень гибкими и мощными инструментами для семантического анализа.

Контекстуальные представления — это способ представления слов в виде векторов, где каждое слово имеет свое значение в зависимости от контекста, в котором оно используется. Это позволяет моделям более точно понимать смысл слов и их взаимосвязи. В отличие от статических векторных представлений (например, Word2Vec), контекстуальные представления учитывают порядок слов и их окружение в предложении.

Семантический анализ находит применение в различных областях, включая:

Маркетинг: Анализ отзывов и комментариев клиентов для определения их мнения о продукте или услуге.

Социальные сети: Мониторинг и анализ упоминаний брендов или событий.

Образование: Автоматическая оценка эссе и тестов на основе анализа содержания и структуры текстов.

* 1. Распределение ролей и работы

Лебедева Лира Петровна занималась написанием отчета, отвечала за исправление и написание отчета.

План по выполнению работы по части написания отчета:

1. Написание структуры отчета.
2. Написание отчета.
3. Просмотр отчета на наличие ошибок.
4. Исправление ошибок.

Динганорбоев Эрдэм Булатович занимался поиском информации, отвечал за теорию и сбор материала.

План по выполнению работы по части поиска информации:

1. Написание плана работы.
2. Поиск информации.
3. Выделение важных терминов и определений.

Иннокентьев Влад Александрович занимался программированием, отвечал за репозиторий.

План по выполнению работы по части программирования:

1. Создание основы.
2. Создание UI.
3. Добавление Hugging Face библиотеки.
4. Написание кода.

Николаев Аслан Анатольевич занимался программированием, отвечал за корректность кода.

План по выполнению работы по части исправления кода:

1. Просмотр кода.
2. Debug кода.
3. Push кода в GitHub.
   1. Стек технологий

Для реализации десктопного приложения были выбраны следующие технологии:

1. Языки программирования:
   * Python: Основной язык для разработки приложения, обладающий богатой экосистемой библиотек для обработки текста.
2. Библиотеки:
   * Hugging Face Transformers: Библиотека, предоставляющая доступ к предобученным моделям для семантического анализа.
   * Tkinter: Библиотека для создания графического пользовательского интерфейса (GUI).
3. API:

2.ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1. Архитектура приложения

2.2.Разработка проекта по ролям

2.3.Контроль выполнения плана

ЗАКЛЮЧЕНИЕ

Создание десктопного приложения для семантического анализа текста на русском языке с использованием библиотеки Hugging Face Transformers представляет собой эффективный способ автоматизации анализа текстовой информации. Использование современных методов обработки естественного языка позволяет значительно улучшить качество анализа и расширить возможности взаимодействия с пользователями. Разработка такого приложения может быть полезной в различных областях, включая маркетинг, социологические исследования и автоматизацию клиентского обслуживания.