Создание десктонпного приложения для семантического анализа английского языка текста с использованием библиотеки huggigface Transformers

2025

[ВВЕДЕНИЕ 2](#_Toc199545492)

[1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 4](#_Toc199545493)

[1.1 Терминология по проекту и глоссарий 4](#_Toc199545494)

[1.2 Распределение ролей и работы. 6](#_Toc199545495)

[1.3 Стек технологий. 6](#_Toc199545496)

# ВВЕДЕНИЕ

Чтобы лучше разобраться в том, что такое семантический анализ, сначала следует выяснить значение слова «семантика». Семантика – это дисциплина, которая изучает связь слов между собой и человеческой реальностью; определяет зависимость значения слова от контекста фразы. Семантическая модель включает слово, его определение, сочетания с другими словами, составление из него фраз и предложений.

Семантический анализ — это анализ значения, смысла слова или группы слов, учитывающий функции слов, их связь между собой и окружающей действительностью.

В контексте SEO семантический анализ — это анализ поисковых фраз, запрашиваемых пользователями в поисковых системах. Он позволяет определить самые важные ключевые слова и фразы, отображающие основы деятельности организации, что помогает грамотно сформировать семантическое ядро и привлечь целевую аудиторию.

В ходе анализа текст проходит через несколько этапов обработки: токенизация для идентификации словоформ, морфологический и синтаксический анализ. Последним идёт вторичный семантический анализ, в ходе которого устанавливаются взаимосвязи между сущностями, происходит извлечение мнений и анализ тональности текста.

Huggingface Transformers — это мощная библиотека для работы с трансформерами, которая предоставляет доступ к множеству предобученных моделей для обработки естественного языка. Основные моменты, которые стоит знать:

Модели: Библиотека предлагает модели, такие как BERT, DistilBERT, RoBERTa и другие, которые могут использоваться для различных задач (классификация, генерация текста, перевод и др.).

- Токенизация: Перед передачей текста модели, его необходимо токенизировать. Это обычно включает преобразование текста в формат, понятный модели (например, преобразование в числовые идентификаторы токенов).

- Обработка данных: Huggingface упрощает работу с данными, позволяя легко загружать и извлекать данные из популярных наборов данных.

Акуальность

Цель и задачи проекта

Предмет и объект исследования

# 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

## 1.1 Терминология по проекту и глоссарий

Проект представляет собой десктопное приложение, которое использует модели глубокого обучения для семантического анализа текста на английском языке. Основная задача приложения заключается в интерпретации и извлечении значений из текстов, что может включать в себя определение тональности, извлечение ключевых фраз, анализ намерений и другие аспекты семантического понимания.

Семантический анализ — это процесс извлечения значений и смыслов из текстов, который включает в себя понимание контекста, намерений и эмоций, выраженных в языке. К традиционным методам семантического анализа относятся правила, основанные на лексиконах, а также статистические методы, такие как векторное представление слов (word embeddings).

Hugging Face Transformers: Это библиотека, предоставляющая доступ к предобученным моделям трансформеров, таким как BERT, GPT, RoBERTa и другие. Эти модели позволяют эффективно выполнять задачи обработки естественного языка (NLP), включая семантический анализ. Архитектура трансформеров использует механизм внимания (attention), что позволяет моделям учитывать контекст слов в предложении, что критично для семантического анализа.

Так же у семантического анализа, множество разных библиотек:

NLTK (Natural Language Toolkit) - Это библиотека для работы с текстом на Python, которая предоставляет инструменты для обработки и анализа естественного языка.

* Включает в себя множество функций, таких как токенизация, стемминг, лемматизация и разбор синтаксиса.
* Содержит обширные корпуса текстов и лексиконы.
* Хорошо подходит для образовательных целей и прототипирования.

SpaCy - Это библиотека для обработки естественного языка, ориентированная на производительность и удобство использования.

* Быстрая и эффективная реализация алгоритмов NLP.
* Поддержка различных языков и возможность работы с большими объемами данных.
* Интеграция с моделями глубокого обучения и возможность создания пользовательских моделей.

Gensim - Это библиотека, предназначенная для тематического моделирования и работы с векторными представлениями слов.

* Поддержка алгоритмов для создания векторных представлений (Word2Vec, FastText и др.).
* Удобные инструменты для обработки текстов и построения тематических моделей.
* Хорошо подходит для анализа больших текстовых коллекций.

Scikit-learn - Это библиотека для машинного обучения на Python, которая предоставляет множество инструментов для классификации, регрессии и кластеризации.

* Простота использования и интеграция с другими библиотеками, такими как NumPy и Pandas.
* Подходит для предварительной обработки данных и построения простых моделей машинного обучения.
* Идеальна для задач, связанных с анализом данных и построением предсказательных моделей.

В процесс разработки приложения могут потребоваться такие аннотированные наборы данных, содержащие тексты и соответствующие метки (например, тональность, темы, ключевые фразы). Так же есть предобработка данных, она Включает в себя очистку текста, токенизацию и преобразование текста в формат, подходящий для модели.

В процессе так же присутствует обучение модели, то есть использование предобученных моделей трансформеров, которые можно дообучить на специфических данных для улучшения точности. Интерфейс пользователя: Разработка удобного интерфейса для взаимодействия с пользователем, позволяющего вводить текст для анализа и отображать результаты.

Инструменты и технологии в данном проекте происходят в формате, Python - основной язык для работы с библиотекой Hugging Face Transformers. Так же для графического интерфейса, используются фреймворки, таких как Tkinter или PyQt, для разработки десктопного приложения.

Для оценки качества семантического анализа могут использоваться, метрики производительности, такие как точность (accuracy), полнота (recall), F1-меры. Так же происходит пользовательское тестирование, то есть проверка удобства и эффективности интерфейса приложения на реальных пользователях.

1.2 Распределение ролей и работы.

Ковалёв Никита, занимался отчет, отвечал за отчет, поиск информации по отчету, поиск информации по коду.

Асмаков Александр, занимался кодом, отвечал…, поиск информации по отчету, поиск информации по коду.

Местников Артём, занимался кодом, отвечал…, поиск информации по отчету, поиск информации по коду.

1.3 Стек технологий.

Язык программирования

- Python

Библиотеки

- Huggingface Transformers для работы с предобученными моделями

- NLTK или SpaCy для предварительной обработки текста

- Tkinter или PyQt для создания графического интерфейса

Менеджер пакетов

- pip или conda для установки зависимостей

Среда разработки

- PyCharm или Visual Studio Code

База данных (по желанию)

- SQLite или PostgreSQL для хранения результатов анализа

Интеграция с API (по желанию)

- Flask или FastAPI для создания RESTful API

Дополнительные инструменты

- Git для управления версиями

- Docker для контейнеризации приложения (по желанию)