Министерство науки и высшего образования Российской Федерации

Университет информационных технологий

ДИПЛОМНАЯ РАБОТА

на тему:

«Разработка интеллектуальной системы анализа текстов на естественном языке»

Выполнил: студент группы ИС-41
Иванов Иван Иванович

Научный руководитель:
к.т.н., доцент Петров Пётр Петрович

Москва — 2025 г.

Содержание

Аннотация

Дипломная работа посвящена разработке интеллектуальной системы анализа текстов на естественном языке. В работе рассматриваются методы машинного обучения и обработки естественного языка (NLP) для автоматической классификации и извлечения смысловых единиц из текстовых данных. Разработана программная реализация на языке Python с использованием библиотеки `scikit-learn`. Практическая часть включает анализ русскоязычных текстов и построение модели классификации по тематике.

Введение

Современные технологии обработки текстовой информации играют ключевую роль в развитии искусственного интеллекта. В связи с ростом объёмов текстовых данных возрастает необходимость в автоматизированных системах, способных анализировать и понимать естественный язык. Цель данной работы — исследование и разработка интеллектуальной системы анализа текстов, обеспечивающей классификацию и тематическую группировку документов.

Для достижения цели были поставлены следующие задачи:

- анализ существующих методов обработки естественного языка;
- проектирование архитектуры системы;
- реализация алгоритмов машинного обучения для классификации текстов;
- экспериментальная оценка качества работы модели.

1. Аналитический обзор

В последние годы значительное развитие получили алгоритмы глубокого обучения, применяемые для обработки естественного языка. Модели семейства трансформеров, такие как BERT, GPT и RoBERTa, демонстрируют высокие результаты в задачах анализа тональности, классификации и генерации текста.

Однако для русскоязычных данных до сих пор актуальны классические подходы, основанные на векторизации слов (TF-IDF, Word2Vec) и методах машинного обучения, таких как логистическая регрессия и SVM. Комбинирование современных и традиционных подходов позволяет достичь баланса между точностью и вычислительной эффективностью.

1.1. Сравнение подходов

Методы глубокого обучения требуют больших вычислительных ресурсов и больших объёмов размеченных данных. Классические модели менее требовательны, но иногда уступают по качеству. В данной работе выбран гибридный подход — использование TF-IDF для векторизации и логистической регрессии как классификатора.

2. Проектирование и реализация системы

Разрабатываемая система включает три основных модуля:

1. модуль предобработки текстов;
2. модуль обучения модели классификации;
3. модуль анализа новых данных.

В качестве языка программирования выбран Python, а для реализации модели — библиотека `scikit-learn`. Архитектура системы предусматривает возможность расширения функционала за счёт подключения дополнительных алгоритмов и источников данных.

2.1. Описание работы программы

Программа принимает на вход текстовые документы, выполняет их токенизацию, удаление стоп-слов, лемматизацию и векторизацию. Затем обученная модель классификации присваивает каждому документу тематическую метку. Результаты анализа визуализируются в виде таблицы или диаграммы.

Заключение

В результате выполнения дипломной работы была разработана интеллектуальная система анализа текстов на естественном языке. Проведён анализ существующих подходов, реализована программная часть и выполнено тестирование на реальных данных. Полученные результаты подтверждают эффективность предложенного решения.

Дальнейшее развитие проекта связано с интеграцией методов глубокого обучения и расширением функционала системы.