УДК 004.912:006.72

КЛАССИФИКАЦИЯ СТРУКТУРНЫХ ЭЛЕМЕНТОВ ВКР

Козырева А. И. (Университет ИТМО)
Научный руководитель – А. В. Бережков (Университет ИТМО)

В данном докладе проводится сравнение результатов различных подходов в решении задачи классификации структурных элементов текстовых документов. Также описывается задача создания алгоритма классификации структурных элементов ВКР, а в дальнейшем системы нормоконтроля документов с возможностью автоматизированного исправления ошибок оформления на основе разработанной классификации элементов документов и их последовательности.

Введение. Качество оформления студенческих работ на всех уровнях (ВКР, курсовых работ, учебной документации и. т. д.) не соответствует текущим ГОСТам и нормативным локальным актам Университета ИТМО. Создание алгоритма решит такие проблемы, как отсутствие у студентов умения и мотивации правильно оформлять документы, неоправданно большое количество времени, которое уходит у преподавателей на документы и проверку работ и формулировка компетенций в общем виде.

Основная часть. Цель работы состоит в решении задачи классификации структурных элементов текстовых документов. Задача классификации предполагает использование машинного обучения. Набор данных для работы получен из выпускных квалификационных работ выпускников Университета ИТМО путем парсинга XML-структуры. В ходе работы предполагается определить наиболее предпочтительный метод классификации для набора данных, а также параметры модели, позволяющие получить оптимальное значение заданному функционалу качества. Были рассмотрены и проанализированы такие методы, как анализ основных компонентов (PCA), стеккинг, беггинг, бустинг. Также рассмотрены такие библиотеки, как LightGBM, XGBoost и CatBoost. Далее проводится инженерия фич признака для повышения качества модели и минимизации ошибок первого и второго порядков. Также разрабатывается метод проверки структурных элементов на соответствие ГОСТ.

Выводы. В рамках данной работы были рассмотрены существующие подходы к классификации структурных элементов текстовых документов. Также были определены наиболее подходящие методы классификации и параметры модели. В основном была проверена библиотека CatBoost, однако в дальнейшем необходима оптимизация параметров и проверка других классификаторов.

Козырева А. И. (автор) Подпись

Бережков А.В. (научный руководитель) Подпись