БИБЛИОТЕКИ ПОДПРОГРАММ (SDK)
ДЛЯ ТЕКСТОПОНИМАНИЯ И ТЕКСТОГЕНЕРАЦИИ
НА ОСНОВЕ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ

ОБЩЕЕ ОПИСАНИЕ СИСТЕМЫ

Самара, 2022

БИБЛИОТЕКИ ПОДПРОГРАММ (SDK) ДЛЯ ТЕКСТОПОНИМАНИЯ
И ТЕКСТОГЕНЕРАЦИИ НА ОСНОВЕ ТЕХНОЛОГИЙ МАШИННОГО ОБУЧЕНИЯ

предназначен для структурированного и автоматизированного извлечения единиц информации из корпоративной документации на основе онтологии и нейросетей для решения задач анализа, экспертизы и генерации больших по объему данных в документации.

В целом Система представляет собой набор библиотек подпрограмм(подсистем) (SDK) для текстопонимания и текстогенерации на основе машинного обучения для использования в системах управления цифровым контентом нового поколения и предназначено для автоматизированного анализа неструктурированных документов, экспертизы и генерации проектной документации при создании новых проектов в различных предметных областях.

Для этого в **состав** Системы реализуемых библиотек подпрограмм (SDK) для текстопонимания и текстогенерации на основе машинного обучения входят следующие основные подсистемы:

- «Конфигуратор извлекаемых данных», обеспечивающий возможность создания и управления предметно-ориентированной онтологией;
- «Модуль обработки корпоративной документации», обеспечивающий возможность структурированного извлечения единиц информации из корпоративной документации на основе онтологии;
- «Классификатор корпоративной документации», позволяющий классифицировать различные виды корпоративных документов с возможностью обучения моделей машинного обучения: подбор, обучение и стекинг моделей для классификации поступающих в систему видов корпоративной документации.

Входящие в Систему Подсистемы имеют следующие функциональные характеристики:

- «Распознавание текста в документе». Подсистема производит распознавание при различных возможных дефектах с изображений документа, таких как: следы от жидкостей, неровная линия строки, перекос страницы, посторонний «шум». Подсистема автоматически обрабатывает (корректирует) оцифрованное документ, выполняет сегментацию документа и последующее распознавание текста (оптическое распознавание символов) в оцифрованном документе.
- «Понимание смысла распознанного текста в документе». Подсистема выполняет семантический поиск фрагментов текста в документе и производит привязку

фрагментов текста документа в некоторую базу данных (тезаурус по предметной области), а также выполняет сопоставление распознанного текста в документе с текстом в других типах документов.

- «Поиск общего и частного». Подсистема выполняет поиск и выделение в распознанном документе разделов, описывающих одинаковые аспекты, а также выполняет поиск однотипных разделов определенной предметной области среди множества документов системы.
- «Верификация документов». Подсистема выполняет проверку документов на наличие ошибок, таких как: наличие грамматических и синтаксических ошибок, разделов отличающихся от остальных в документе. Это необходимо для вычленения возможных ошибок в документах: поиск документа, в котором отсутствует определённый раздел, или поиск раздела не соответствующий требованиям нормативной документации.
- «Генерация шаблонов документов». Подсистема даёт возможность в составлении шаблона для распознавания определенного вида документов, по определенному набору определенных параметров.
- «Классификация документов». Подсистема классифицирует поступающие в систему виды корпоративной документации и выполняет определение набора классификационных признаков по тексту в документе.