УДК 004.048

ОБЗОР ИНСТРУМЕНТОВ ДЛЯ РАЗМЕТКИ ТЕКСТОВЫХ ДАННЫХ

Яшихина Е.В. (Университет ИТМО)
Научный руководитель – к.т.н., доцент практики ИДУ Вайгандт Н.Ю.
(Университет ИТМО)

Введение. Разметка данных — это неотъемлемая часть машинного обучения и искусственного интеллекта (ИИ). Она требуется для обучения модели машинного обучения и относится к предварительной обработке данных. Различают несколько типов размеченных данных: текст, изображение, видео и звук. Разметка текстовых данных является одной из задач обработки естественного языка (NLP) [1].

Основная часть. На сегодняшний день существует широкий выбор готовых инструментов разметки текстовых данных. Можно рассматривать в зависимости от открытости инструмента, существуют продукты open-source (бесплатные) и закрытые (требуют оплаты). Также инструменты предлагают различные методики разметки данных, можно разметку выполнять вручную (собственной командой, специалистами в предметной области, сторонними разметчиками без профессиональной подготовки) или автоматизировано. Выбор инструмента зависит от задач исследования.

Tagtog — инструмент для разметки текстовых данных на естественном языке [2]. Процесс разметки оптимизирован для текстовых данных для создания специализированных наборов данных для текстового ИИ. С помощью этого инструмента вы можете автоматически получать соответствующие сведения из текста: он помогает выявлять закономерности, проблемы и находить решения. Пользователь может выбрать один из трёх вариантов: разметка текста вручную, приглашение команды для разметки или использование моделей машинного обучения для автоматизированной разметки. Инструмент работает с несколькими форматами: PDF, TXT, HTML, CSV и др. Также у tagtog есть возможность использовать для работы локально или в облаке/ работа по API. Существует открытая версия инструмента, платная версия имеет расширенный функционал. Кроме того, инструмент позволяет связать отношения, прикрепить атрибуты к сущностям или классифицировать весь документ.

Label Studio — инструмент с открытым исходным кодом для разметки данных. Поддерживает типы данных, такие как аудио, текст, изображения, видео и временные ряды, с помощью простого и понятного пользовательского интерфейса с возможность экспортировать их в различные форматы моделей. Инструмент можно использовать для подготовки необработанных данных или улучшения существующих обучающих данных для получения более точных моделей машинного обучения.

Doccano — это инструмент с открытым исходным кодом, который предоставляет функции аннотации для классификации текста, маркировки последовательности [3]. Можно создавать помеченные данные для анализа настроений, распознавания именованных сущностей, обобщения текста и так далее. Инструмент поддерживает ручную, полуавтоматическую (модель машинного обучения размечает данные с возможностью пользователю внести изменения) и автоматическую разметку.

LightTag — это еще один инструмент для маркировки текста, предназначенный для создания точных наборов данных для NLP. Имеет бесплатный пакет доступа, но только для одного разметчика данных, для совместной работы команды следует оплатить продукт. Инструмент предлагает упрощенный пользовательский интерфейс для управления рабочей силой и упрощения аннотаций. Инструмент также предоставляет высококачественные функции контроля для точной маркировки и создания оптимизированных наборов данных [3].

Labelbox — популярный инструмент для разметки данных. Оптимизирован для различных видов данных: изображения, видео, текст, PDF-документы, геопространственные, медицинские изображения и аудиоданные. Обеспечивает среду совместной работы для

команд машинного обучения. Но имеет ограничение бесплатной версии только в 10 тысяч единиц данных Labelbox, это достаточно для небольших команд или для задач на ранней стадии проекта.

Выводы. Таким образом, мы рассмотрели несколько инструментов для разметки текстовых данных. Следует делать выбор инструмента разметки от целей и задач проекта.
Работа выполнена в рамках проекта НИР Университета ИТМО № 622264 «Разработка сервиса выявления объектов городской среды общественной активности и ситуаций повышенного риска на основе текстовых сообщений горожан».

Список использованных источников:

1. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных // учеб. пособие / — М.: Изд-во НИУ ВШЭ. – 2017. — С 104–109.
2. Cejuela J.M, McQuilton P., Ponting L., Marygold S. J., Stefancsik R., Millburn G. H. Tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles // Database the Journal of Biological Databases and Curation. 2014.
3. Perry T. LightTag: Text Annotation Platform // 2021.