Проектирование сервиса для разметки данных

Лабораторная работа #4
Работу выполнили Атрушкевич А., Буценко С., Радионова Е.

Описание проекта

Цель проекта - создание внутренней системы компании для разметки данных. Это позволит компании самостоятельно получать данные для обучения нейросетей. Проект экономически выгоден компании, ведь ей не придется покупать датасеты для обучения.

Требования клиента

В результате интервью с клиентом были выявлены следующие требования:

- В сервис может войти любой работник компании
- Небольшой бюджет
- Большая скорость разметки
- Простой интерфейс
- Разметка текста и картинок по классам и тэгам
- Тренировочные тесты / примеры
- Определение качества разметки
- Рейтинг разметчика
- Админ для контроля проектов

Анализ конкурентов

- SuperAnnotate - самый популярный сервис на данный момент. Предоставляет мощные инструменты, например, семантическую сегментацию, ключевые точки и рисование линий, которые позволят сэкономить время и повысить точность. Однако интерфейс сервиса не кажется интуитивным.
- Supervisely - имеет более восьмидесяти вариантов аннотирования, упорядоченных в категории, обеспечивает высокую гибкость, предоставляет собственные датасеты.
- Hive Data — имеет удобные функции наподобие возможностей облачного масштабирования.

Все из вышеперечисленных сервисов обеспечивают высокую скорость разметки, имеют огромный функционал, и не только в сфере аннотирования, но на всех этапах машинного обучения.

Описание разрабатываемой системы

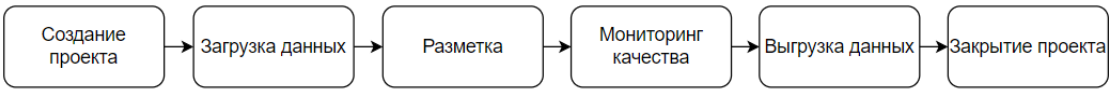
Предполагаются следующие участники будущей системы:

- Заказчик – пользователь, который загрузил хотя бы один проект. Заказчиком может быть любой пользователь системы. Цель заказчика – получить качественно размеченные данные.
- Разметчик – пользователь, который работает над проектами (размечает датасеты). Разметчиком также может быть любой пользователь.
- Администратор, в отличие от остальных пользователь имеет доступ к спискам всех проектов и пользователей и может их модерировать.

Проект - сущность внутри бизнес процесса. Цель проекта – получить файл меток в формате json, относящийся к входному датасету. Метод преобразования информации – ручная разметка.

Жизненный цикл:

Создание проекта -> Загрузка данных -> разметка/мониторинг -> выгрузка данных -> архивирование/удаление проекта



1. Создание проекта

   Создать проект может любой пользователь. При создании проекта обязательными полями являются название, инструкция, коэффициент качества, тип данных, тип разметки, теги/классы. Также пользователь обязан загрузить датасет в проект, прежде чем публиковать его. Необязательные поля - описание, дедлайн, пример разметки, тренировочные и проверочные тесты. Проект можно не публиковать, а сохранить как черновик. Для этого необязательно заполнять обязательные поля. Черновики отображаются на вкладке Мои проекты вместе с опубликованными проектами.

2. Загрузка данных

   Для того чтобы загрузить данные, пользователь выбирает папку на своем устройстве, которая помещается в архив и загружается в сервис. Из архива автоматически выбираются файлы того типа, который заказчик указал (для текста - txt, для изображений png, jpeg).

3. Разметка

Размечать проект может любой пользователь. Исключение составляют те пользователи, которых заказчик самостоятельно удалил из проекта.

Поддерживаемые форматы разметки:

- Изображение (Тег)
  Разметчик выделяет прямоугольную область на начальной картинке, которая определяет границы объекта. Далее для каждого объекта на картинке можно выбрать тег из предложенных.
- Изображение (Класс):
  Картинка не предусматривает фрагментирования т.е. одна картинка целиком относится к одному классу. При разметке пользователь только выбирает класс из набора предоставленных.
- Текст (Тег)
  Разметчик из списка доступных тегов выбирает нужный и выделяет слова в тексте, которые относятся к этому тегу.
- Текст (Класс)
  Разметчик определяет к какому классу отнести текст. Текст можно отнести только к одному классу.

4. Мониторинг качества

Для того чтобы добиться желаемого качества разметки есть несколько инструментов. Во-первых, коэффициент качества, который обозначает количество разметчикам которым будет выслано каждое задание, прежде чем считаться выполненным. Во-вторых, рейтинг разметчиков, который уникален для каждого проекта. Для подсчета рейтинга используются проверочные тесты, загружаемые заказчиком, или производятся расчеты на основе консенсуса (Ответ, который выбрало большинство разметчиков считается правильным. Если разметка выполнялась тегами на картинке, то пиксели, которые попадают в выбранный тег чаще всего считаются правильными). Заказчик имеет возможность просмотреть рейтинг разметчиков на своем проекте и удалить неугодных. Все ответы удаленных разметчиков стираются, а сами разметчики больше не имеют доступа к проекту.

5. Выгрузка данных

Заказчик имеет возможность в любой момент скачать размеченные данные из своего проекта. Данные загружаются в json файл в формате название_Файла-метка. При скачивании есть выбор - скачать как новый или догрузить данные в уже существующий файл. При догрузке данных в уже существующий файл скачиваются только те данные, которые не совпадают с данными из файла. Если метка для одного и того же файла

отличается, она заменяется на более новую. При загрузке в новый файл скачиваются все данные.

6. Закрытие проекта

   Заказчик может закрыть свой проект, после чего тот перестанет быть доступным другим пользователям. Закрытые проекты отображаются на вкладке Мои проекты иначе, чем текущие. С закрытого проекта можно выгрузить данные и опубликовать его снова.
   Если все загруженные данные были размечены, проект автоматически становится закрытым, и опубликовать снова его нельзя.
   Закрытый проект можно удалить, после чего он станет недоступен и самому заказчику, и админу.

Пользовательские истории

С пользовательскими историями и критериями приемки можно ознакомиться на доске Kanban Flow.

Пользовательское взаимодействие и дизайн

Проекты в draw.io и balsamiq приложены.