Проект по подготовке данных для обучения и оценки диалоговых эмбеддеров

Автор: Дарья Леднева

Введение

Что такое диалоговые эмбеддеры?

- ➔ Специальные модели, обученные на диалоговых данных
- ➔ Создают эмбеддинги, которые улавливают контекст, намерения и смысл реплик

Почему важны диалоговые эмбеддинги?

- ➔ Обеспечивают высокую точность в задачах обработки диалогов
- ➔ Применимы в различных сценариях: чат-боты, голосовые помощники, анализ пользовательских отзывов

Введение

Задача проекта: создание платформы для обработки диалоговых данных, чтобы в дальнейшем их использовать для разработки и оценки диалоговых эмбеддеров

Три направления:

- ➔ Формирования диалогового корпуса для предобучения моделей
- ➔ Подготовка данных для контрастивного обучения
- ➔ Подготовка данных для оценки эмбеддингов на downstream-задачи

Цели

Цели проекта:

- Обеспечить высокое качество диалоговых данных для обучения моделей
- Создать удобные инструменты для анализа данных
- Автоматизировать процессы обработки данных и формирования диалогового корпуса
- Предоставить возможность оценки качества эмбеддингов через downstream-задачи путем формирования соответствующего набора данных

Данные для предобучения

Корпус диалогов для BERT-like задач:

- **Зачем:** обучение моделей с использованием полных диалогов
- Примеры задач обучения: предсказание автора реплики, определение порядка реплик, маскирование с учетом контекста

Корпус пар реплик для контрастивного обучения:

- **Зачем:** обучение моделей на различение встречающихся рядом в диалоге реплик и случайных при помощи контрастивного обучения

Всего было использовано 9 самых популярных диалоговых наборов данных, ориентированных на решение задач

Данные для оценивания

- **Зачем:** оценка качества эмбеддингов на downstream-задачах
- Сбор данных из **мультиязычных** датасетов для классификации интентов: Banking77, CLINC150, MASSIVE, MINDS14, SNIPS
- Приведение данных к единому формату с колонками: текст реплики, метка класса, сплит, язык данных, название исходного датасета

База данных и дашборд

База данных:

- Реляционная база данных (SQLite)
- Содержит данные для предобучения эмбеддеров и downstream-задач

Дашборд:

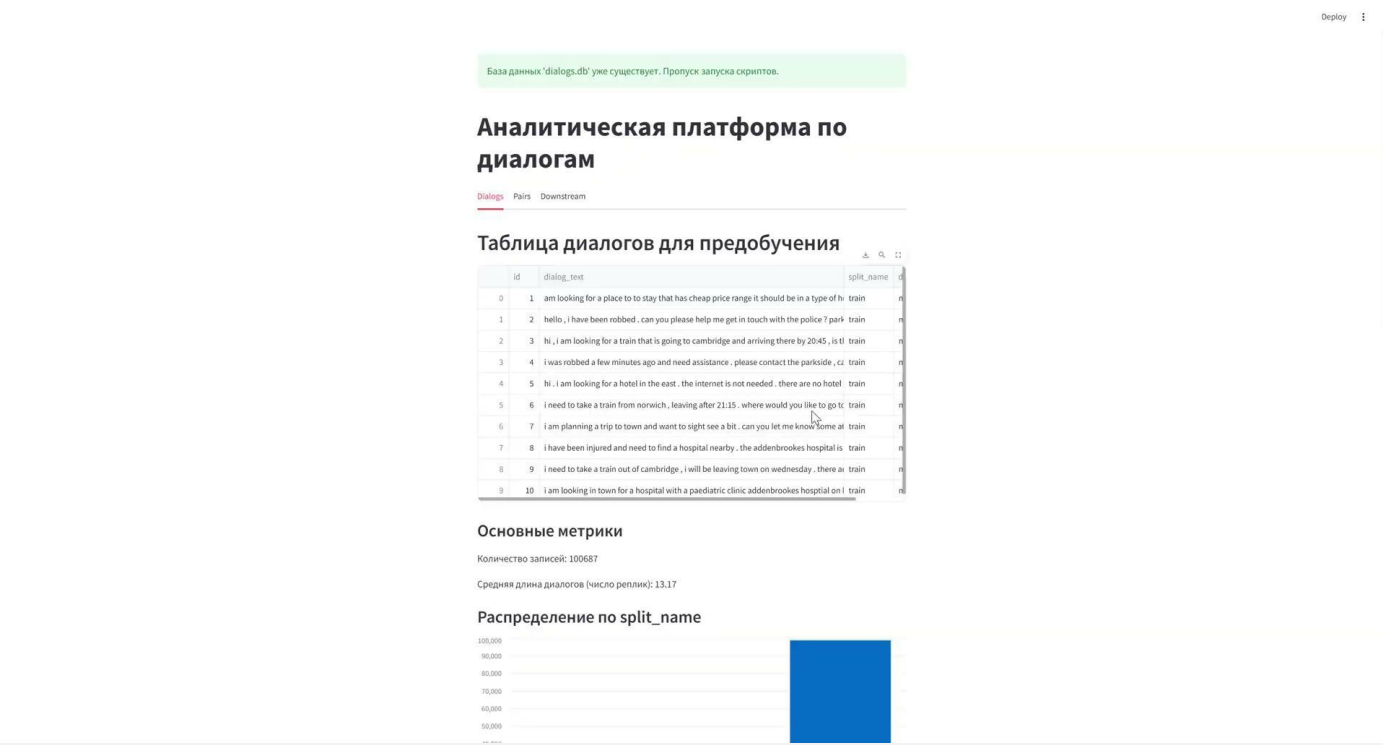
- Платформа: Streamlit.
- Содержит метрики данных — общее количество записей, статистика по репликам диалогам, распределение реплик и меток, распределение данных по языкам

Метрики качества данных

Для обеспечения качества данных используются три ключевые таблицы базы данных и их метрики:

- **Полные диалоги для предобучения:** число записей, длина диалогов и реплик, распределение данных по split и датасету
- **Пары реплик для контрастивного обучения:** количество пар реплик, длина реплик, соотношение длин первой и второй реплик в паре
- **Данные для downstream-задач:** количество записей, языков, меток, распределения данные по сплитам, меткам и языкам

Демо дашборда



QR-код репозитория

Спасибо за внимание!