Алгоритм решения кейса МТС-линк

Получение данных

Перевод базы данных ответов форм МТС-линк в текстовый файл для дальнейшей работы

- Нужно сделать файл csv/json, каждый элемент которого представляет собой ответ сотрудника на поставленный вопрос. *(Лучше сделать json, так как в ответах могут быть запятые, что попортит csv файл)*
- Нужно понять в каком формате существует массив ответов с МТС-линк форм, чтобы потом было легко внедрить продукт в компанию.

Вход: массив ответов МТС-линк форм

Выход: json-файл вида:
{«id ответа»: «ответ», ...}

Выполняет: Кирилл

Создание эмбеддингов

Применение мощной модели, предобученной на большом массиве данных создает численное представление слов.

Для представления коротких фраз: USE (universal sentence encoder). Круто что понимает на разных языках, обучен под конкретную задачу.

Вход: json-файл ответов

Выход: Массив эмбеддингов
shape = (N, 512)
N — количество ответов
512 — кол-во элементов эмбеддинга

Выполняет: Миша

Кластеризация

Кластеризация полученного датасета при помощи подходящего алгоритма.

Кандидаты:
- kmeans
- dbscan
- birch
- diameter-clustering *

Провести исследование, какой алгоритм отрабатывает лучше на сгенерированном датасете. Выбрать лучший.

Вход: Массив эмбеддингов

Выход: Массив меток кластеризации shape = (N, 1)

Выполняет: Алиса

Обобщающая кластерная фраза

Определение главной мысли кластера, подсчет фраз, относящихся к этому кластеру

Нужно найти модель, качественно делающую summary набору ответов в виде короткого предложения.

Вход: Массив эмбеддингов,
Массив меток кластеризации

Выход: Массив главных фраз
shape = (k, 1)
k -количество кластеров

Выполняет: Миша, Дима

Понижение размерности

Понижение размерности эмбеддингов длля дальнейшей визуализации.

Кандидаты:
- t-SNE
- UMAP

Визуальная оценка, выбор лучшего отображения.

Вход: Массив эмбеддингов

Выход: Сжатый массив
shape = (N, 2)

Выполняет: Миша, Алиса

Красивая визуализация

Интерактивный график, строящий k облаков (кластеров), кластер называется обобщающей фразой. При наведении курсора на кластер, отображается подробная информация о нем: количество фраз, относящихся к теме, топ представителей, наиболее близких к центру кластера

- plotly или seaborn

Вход: Сжатый после понижения размерности массив (N, 2), массив гланых фраз

Выполняет: Дима

Создание интерфейса PyQt

Интерфейс, в котором есть:

- Кнопка выбора файла ответов, из раздела «Получение данных»

- Кнопка загрузить модель ИИ (из пункта «Создание эмбеддингов») (опционально: чтобы было несколько моделей), контекстное меню если выбор из нескольких

- Кнопка построения красивой визуализации из пункта «Красивая визуализация»

Выполняет: Кирилл, Миша