HACKS-AI
DIGITAL.KURSK

Классификация обращений граждан

2022. Зубоченко Антон Константинович

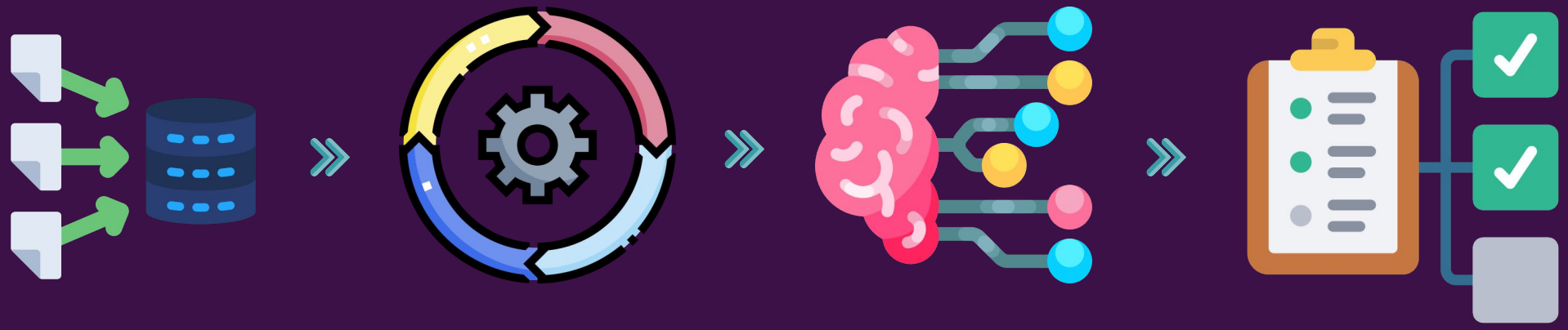
Задача

Разработать классификатор для автоматического определения категории запроса по тексту сообщения, оставленному на сайте Администрации Курской области

Цель

Сократить время ответа на обращения жителей

Pipeline



Обращения

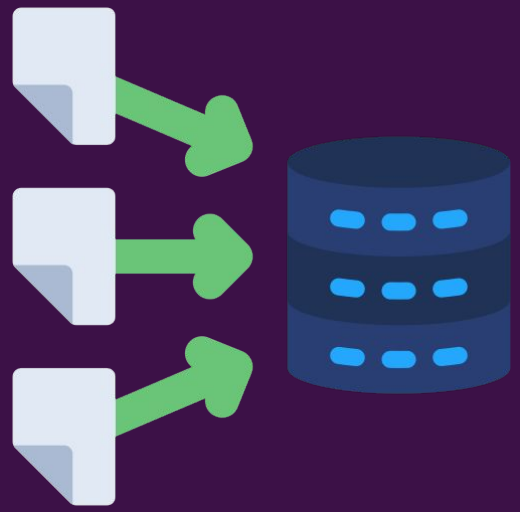
Препроцессинг и feature engineering

Модель

Результат

Pipeline

Обращения представлены в виде текста с дополнительной информацией об ответственном лице, тематике и лейблом категории



Pipeline

Препроцессинг

- Очистка корпуса от html тегов попавших в текст при сборе данных
- Очистка текста от стоп-слов
- Токенизация и удаление биграм встречающихся менее 5 раз

Pipeline

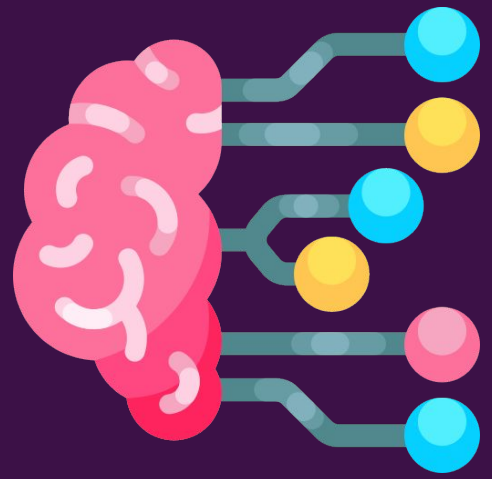
Feature engineering

- Создание словаря весов TF/IDF для всего корпуса (train+test)
- Upsampling обучающей выборки разделением текстов на куски не более 256 символов
- Кластеризация обучающей выборки в соответствии с лейблами тематики
- На основании кластеризации предсказание тематики в тестовом наборе данных через матрицу весов TF/IDF

Pipeline

Модель

- Объединение текста с предсказанной тематикой на основе кластеризации
- Bert finetuning на классификацию текстов
- В качестве базовой модели использована rubert-base-cased-conversational от DeepPavlov



Pipeline

Результ
ат

- Multi AUC-ROC на публичном лидерборде 0.998829

Спасибо за внимание!