



HACKS-AI DIGITAL.KURSK

Классификация обращений граждан

Задача

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

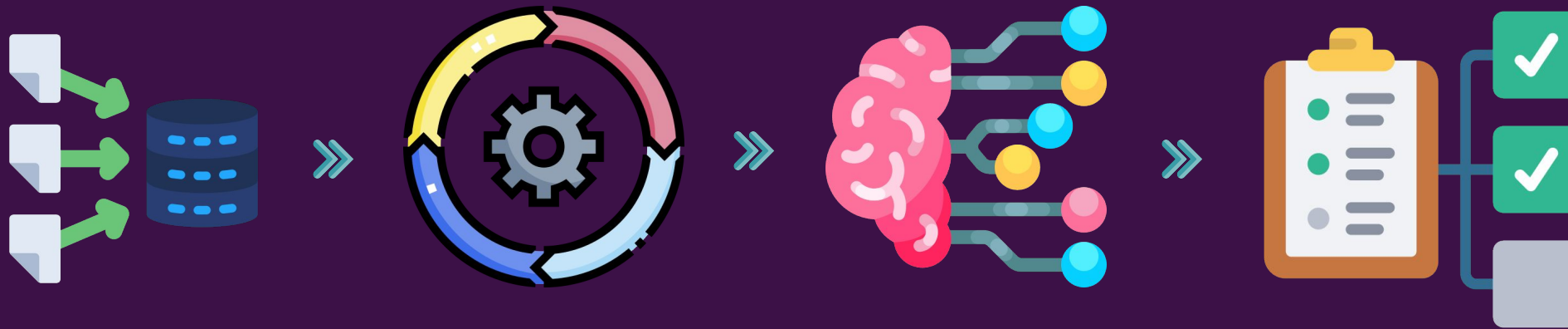


Цель

Сократить время ответа на обращения жителей



Pipeline



Обращения
я

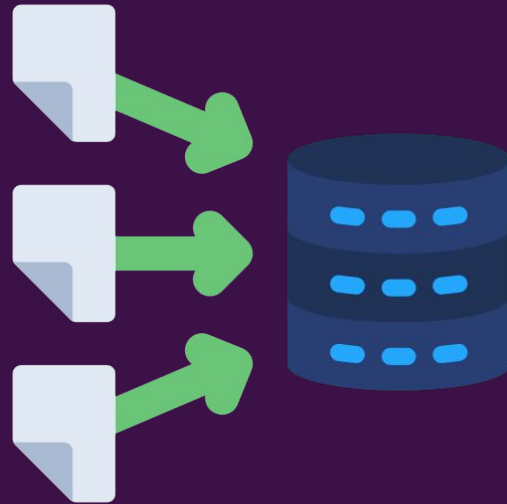
Препроцессинг
и feature engineering

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 биграмм 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

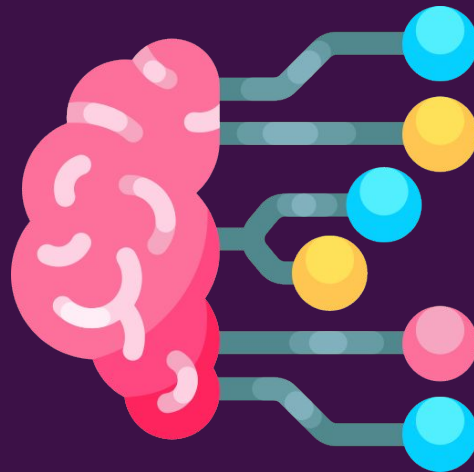


Pipeline

Модел

ь

- 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

Результат
ат

- Multi AUC-ROC на публичном лидерборде 0.998829



Спасибо за внимание!

