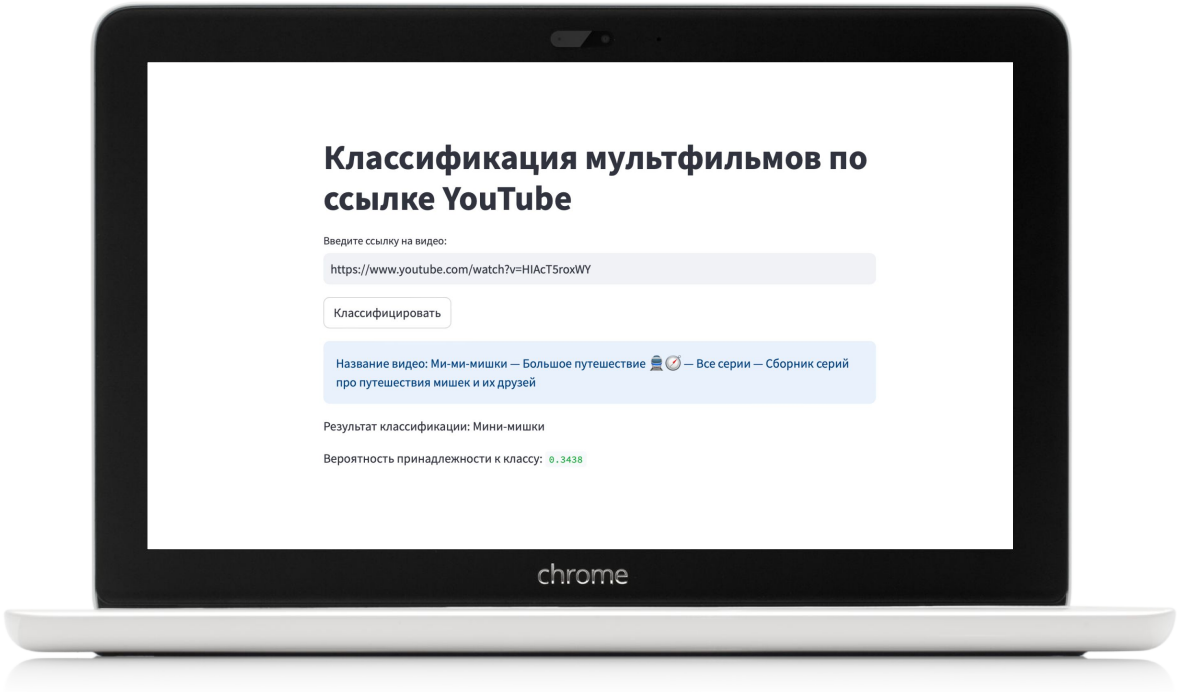
Классификация мультфильмов Youtube



Алла Бобрикова

@alla_bobrikova

https://github.com/RollingInDeepLearning/SMF_cartoon_classification/

Features

Сработало

- reel_name (очищенные от ссылок) + id канала

Не сработало

- text + id канала/название канала
- text + id канала
- reel_name + ближайшее по косинусному расстоянию название проекта
- reel_name + выделенное из text описание проекта (regexp)

Решение

Выделение классов с недостаточным количеством экземпляров и парсинг данных по данным классам (около 600 записей по 25 классам)

Векторизация TFIDF

(reel_name + id канала)

Обучение модели Random Forest и подбор наилучших параметров

F1 macro

0.86142

Дальнейшие шаги

Работа с текстовыми эмбеддингами (W2Vec, FastText)

Расширение данных (аугментация, парсинг)

Работа с features (исследование числовых параметров)

Тестирование других моделей (XGBoost или LightGBM)

Работа с трансформерами (BERT)