Определение авторства текстов

Проект первого года студентов магистратуры НИУ ВШЭ «Машинное обучение и высоконагруженные системы»

Промежуточная защита проекта (январь 2024)

Описание задачи

Создать ML-сервис для определения авторства текста по фрагменту из 5-10 предложений

Состав команды, куратор и распределение ролей

Состав команды и распределение ролей

- Дарья Мишина (предобработка данных, EDA, ML, DL)
- Кирилл Рубашевский (сбор данных, EDA, ML и деплой)
- Дмитрий Шильцов (EDA, ML)

Куратор: Елена Вольф

Данные

Для проекта отобраны 10 классических русских писателей 19 века. По каждому автору собрано 10+ прозаических произведений

Данные (тексты и метадата) собраны на сайте интернет-библиотеки Алексея Комарова. Для сбора данных написан класс IlibParser (библиотеки selenium и bs4), который поддерживает:

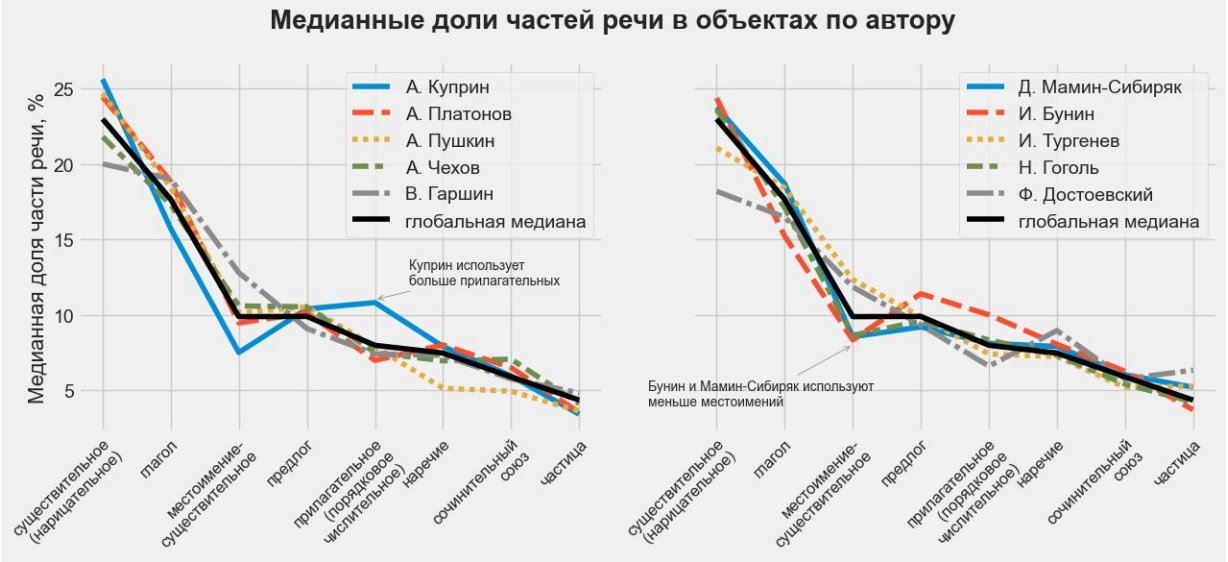
- парсинг текстов на нескольких страницах
- продолжение ранее начатого парсинга

Собранные данные размещены на S3 (Yandex Object Storage) и доступны по внешней ссылке

EDA: part-of-speech и тематическое моделирование

авторы различаются по частоте использования частей речи, но методы понижения размерности (PCA, TSNE) на этих статистиках не позволили кластеризовать авторов

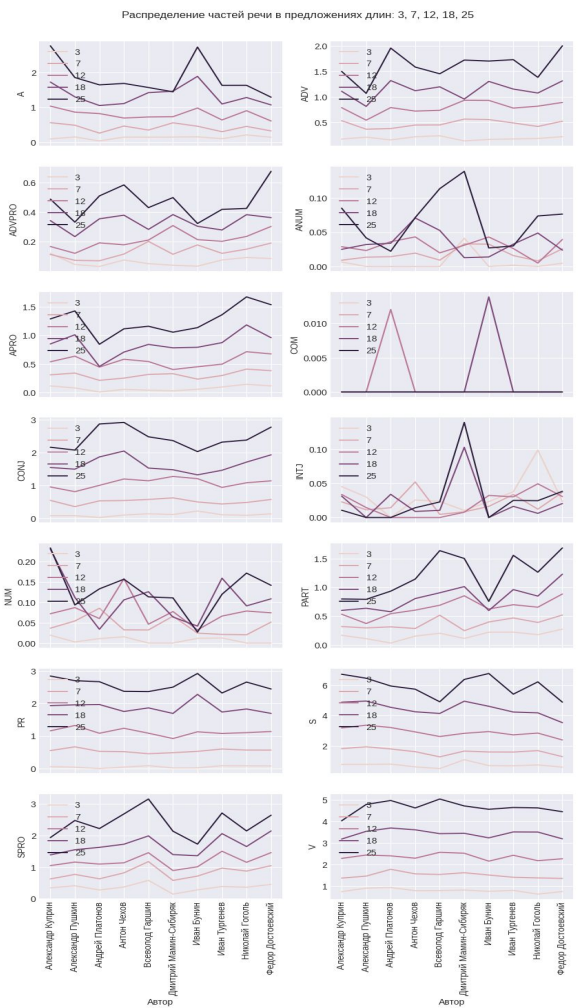
тематическое моделирование не привело к существенным результатам

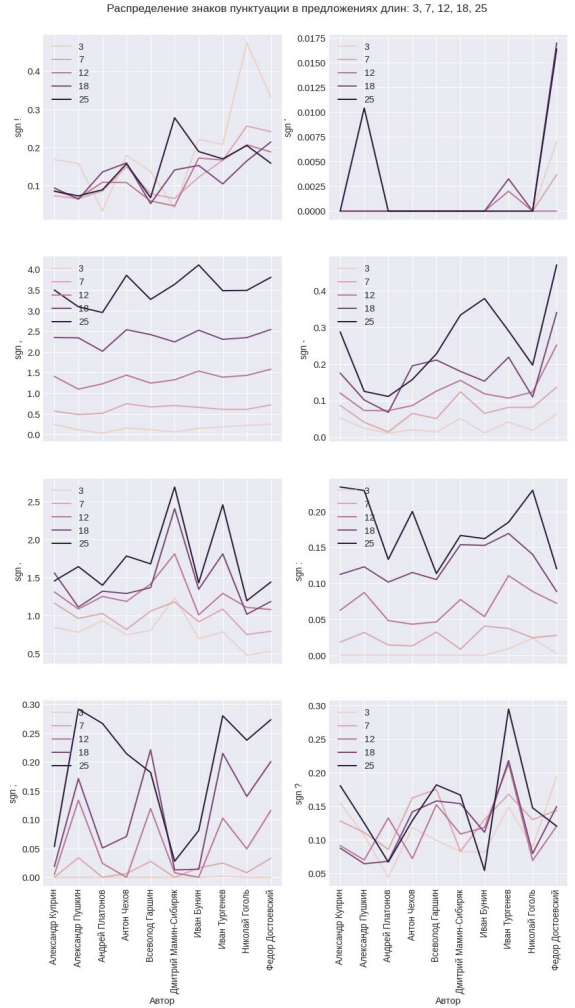


EDA: part-of-speech

авторы по-разному употребляют части речи в зависимости от длины предложения

употребление знаков препинания в зависимости от длин предложения у авторов также различается

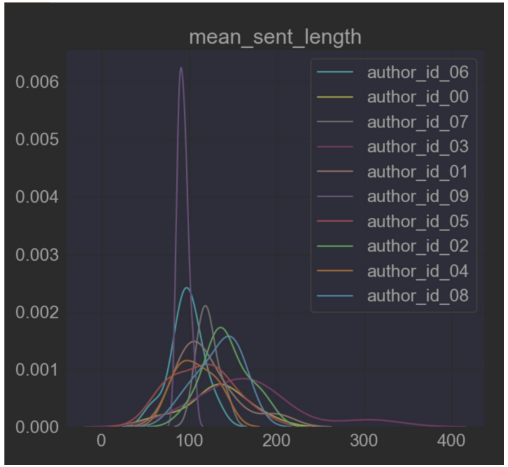


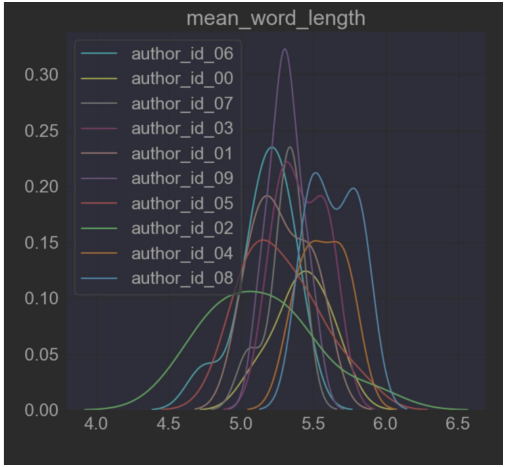


EDA: текстовые статистики

длина слов и предложений у авторов различаются, поэтому сделали углубленный анализ статистик с помощью пакета ruts, изучив количество:

- предложений
- слов
- уникальных слов
- длинных слов
- сложных слов
- простых слов
- односложных слов
- многосложных слов
- символов
- букв
- пробелов
- слогов
- знаков препинания





после анализа самых часто встречающихся слов был расширен список стоп-слов

ML - general

- целевая метрика: f1 macro
- трекинг экспериментов: wandb
- «честный» сплит (объекты одного произведения попадают в одну выборку) показал переобучение моделей на трейне, **поэтому дальше нужно использовать именно его**

ML Дмитрий

- базовая модель на основе частей речи и знаков препинания (они стали «словами») с/без падежей-склонений
- BOW / TF-IDF + линейные модели (лучший результат BOW на ненормализованных частях речи)

	precision	recall	f1-score	support
author_id_00	0.71	0.68	0.70	419
author_id_01	0.69	0.70	0.70	110
author_id_02	0.73	0.65	0.69	906
author_id_03	0.44	0.57	0.50	192
author_id_04	0.60	0.60	0.60	373
author_id_05	0.56	0.62	0.59	192
author_id_06	0.52	0.53	0.52	308
author_id_07	0.50	0.65	0.57	9
author_id_08	0.31	0.54	0.40	115
author_id_09	0.75	0.68	0.71	801
accuracy			0.64	3515
macro avg	0.58	0.62	0.60	3515
weighted avg	0.66	0.64	0.64	3515

ML Кирилл - предобработка, дисбаланс классов, feature engineering

↑ NLP-предобработка (лемматизация, удаление стоп-слов)

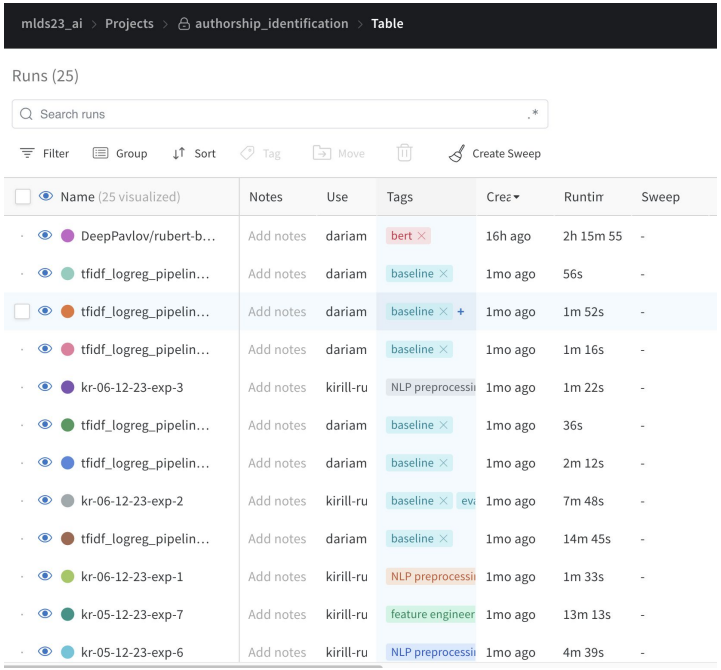
≈ разные стратегии борьбы дисбалансом классов (class weights, downsampling)

≈ статистики по частями речи

↓ N-граммы

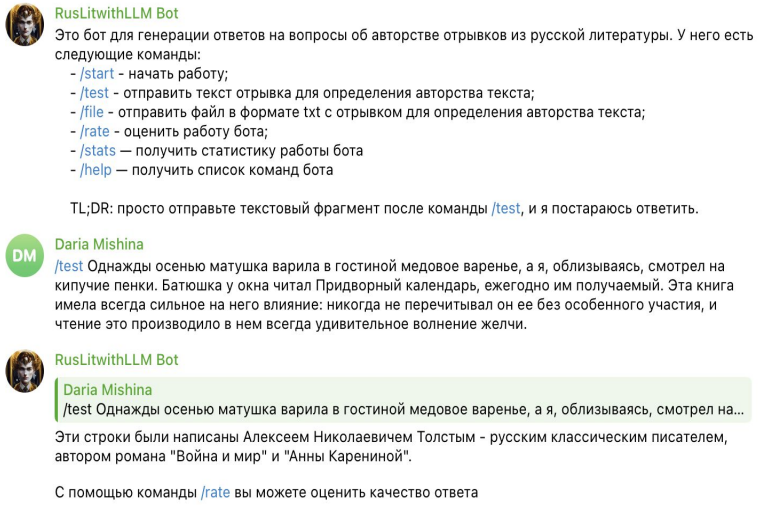
ML Дарья

- baseline: 4 варианта предобработки + TF-IDF + логистическая регрессия
- текст как табличные данные: ruts + логистическая регрессия



DL Дарья

- дообучены на нашем датасете трансформеры cointegrated/rubert-tiny2 и DeepPavlov/rubert-base-cased
- протестированы несколько опенсорсных LLM (пока выбран openchat)



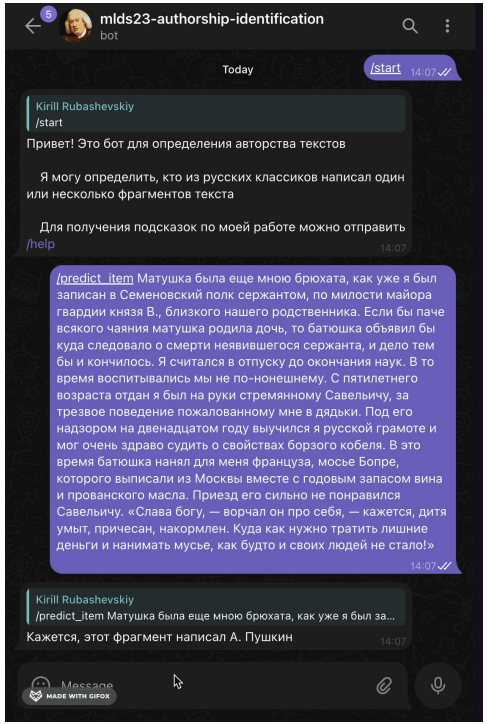
Deployment

MVP реализован в формате тг-бота, развернут на render.com и доступен по ссылке

Модель для работы бота подгружается из S3

Бот:

- может предсказывать авторство одного (сообщение) или нескольких (csv-файл) фрагментов
- возвращает confidence prediction: если вероятность всех авторов ниже порога, бот сообщает о невозможности сделать уверенное предсказание



To Do

ML + DL

- попробовать больше семейств моделей (DL) и эмбеддингов (Word2Vec, GloVe, BERT)
- оптимизировать гиперпараметры через optuna/hyperopt
- дообучить LLM

MLOps + DevOps + deployment

- вынести предсказания из бота в отдельный веб-сервис
- обернуть все в докер
- оформить проект в Streamlit app