

# 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:

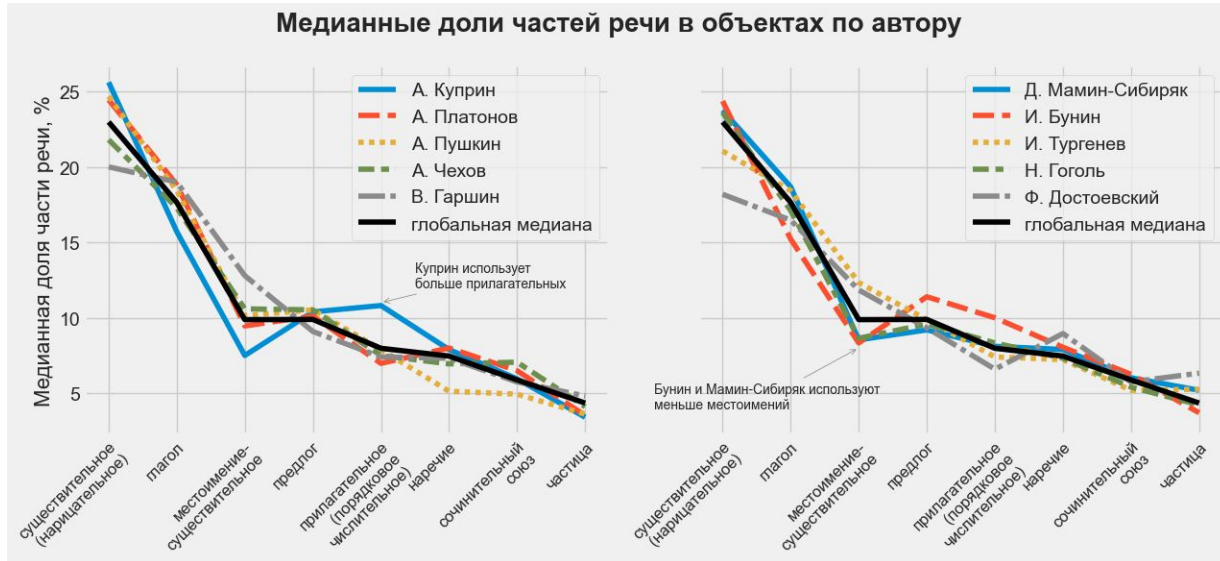
- парсинг текстов на нескольких страницах
- продолжение ранее начатого парсинга

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

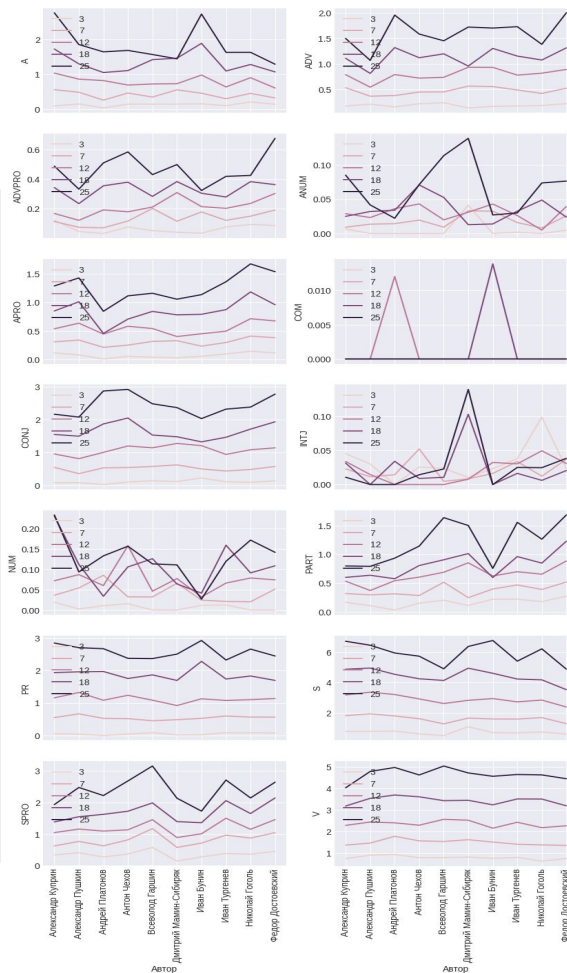


# EDA: part-of-speech

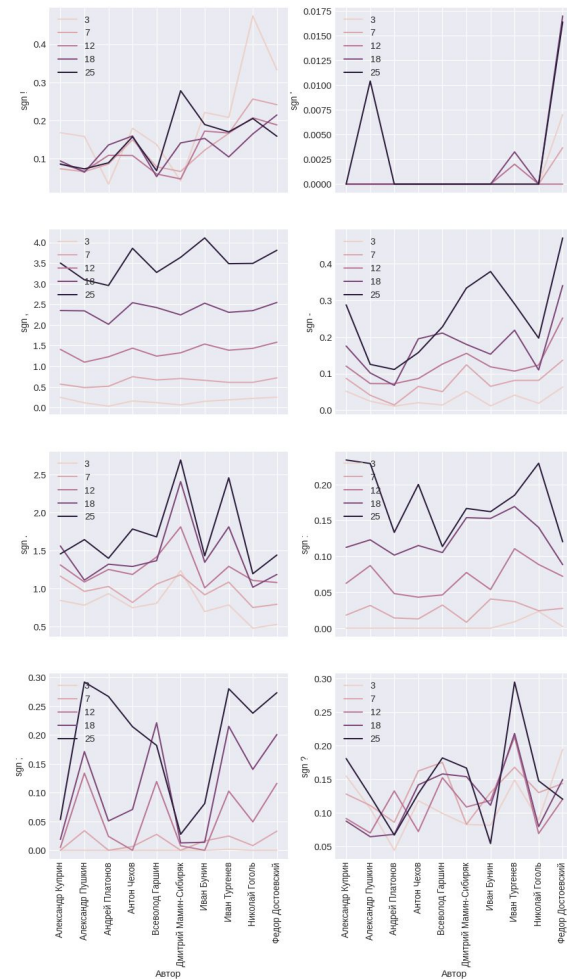
авторы по-разному  
употребляют части речи в  
зависимости от длины  
предложения

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

Распределение частей речи в предложениях длин: 3, 7, 12, 18, 25



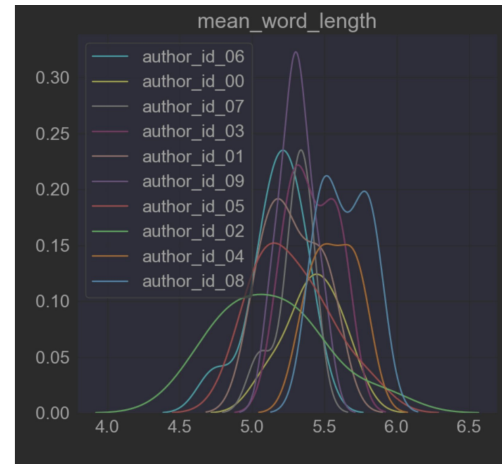
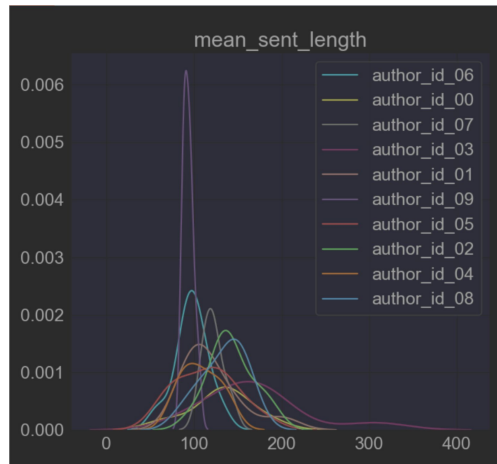
Распределение знаков пунктуации в предложениях длин: 3, 7, 12, 18, 25



# 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	91
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 частям речи

↓ N-граммы

# ML Дарья

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

mlds23\_ai > Projects > authorship\_identification > Table

Runs (25)

Search runs

Filter Group Sort Tag Move Create Sweep

<input type="checkbox"/>	Name (25 visualized)	Notes	Use	Tags	Created	Runtime	Sweep
<input type="checkbox"/>	DeepPavlov/rubert-b...	Add notes	dariam	bert ✕	16h ago	2h 15m 55	-
<input type="checkbox"/>	tfidf_logreg_pipelin...	Add notes	dariam	baseline ✕	1mo ago	56s	-
<input type="checkbox"/>	tfidf_logreg_pipelin...	Add notes	dariam	baseline ✕ +	1mo ago	1m 52s	-
<input type="checkbox"/>	tfidf_logreg_pipelin...	Add notes	dariam	baseline ✕	1mo ago	1m 16s	-
<input type="checkbox"/>	kr-06-12-23-exp-3	Add notes	kirill-ru	NLP preprocessi	1mo ago	1m 22s	-
<input type="checkbox"/>	tfidf_logreg_pipelin...	Add notes	dariam	baseline ✕	1mo ago	36s	-
<input type="checkbox"/>	tfidf_logreg_pipelin...	Add notes	dariam	baseline ✕	1mo ago	2m 12s	-
<input type="checkbox"/>	kr-06-12-23-exp-2	Add notes	kirill-ru	baseline ✕ evi	1mo ago	7m 48s	-
<input type="checkbox"/>	tfidf_logreg_pipelin...	Add notes	dariam	baseline ✕	1mo ago	14m 45s	-
<input type="checkbox"/>	kr-06-12-23-exp-1	Add notes	kirill-ru	NLP preprocessi	1mo ago	1m 33s	-
<input type="checkbox"/>	kr-05-12-23-exp-7	Add notes	kirill-ru	feature engineer	1mo ago	13m 13s	-
<input type="checkbox"/>	kr-05-12-23-exp-6	Add notes	kirill-ru	NLP preprocessi	1mo ago	4m 39s	-

# DL Дарья

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



RusLitwithLLM Bot

Это бот для генерации ответов на вопросы об авторстве отрывков из русской литературы. У него есть следующие команды:

- `/start` - начать работу;
- `/test` - отправить текст отрывка для определения авторства текста;
- `/file` - отправить файл в формате txt с отрывком для определения авторства текста;
- `/rate` - оценить работу бота;
- `/stats` — получить статистику работы бота
- `/help` — получить список команд бота

TL;DR: просто отправьте текстовый фрагмент после команды `/test`, и я постараюсь ответить.



Daria Mishina

`/test` Однажды осенью матушка варила в гостиной медовое варенье, а я, облизываясь, смотрел на кипучие пенки. Батюшка у окна читал Придворный календарь, ежегодно им получаемый. Эта книга имела всегда сильное на него влияние: никогда не перечитывал он ее без особенного участия, и чтение это производило в нем всегда удивительное волнение желчи.



RusLitwithLLM Bot

Daria Mishina

`/test` Однажды осенью матушка варила в гостиной медовое варенье, а я, облизываясь, смотрел на...

Эти строки были написаны Алексеем Николаевичем Толстым - русским классическим писателем, автором романа "Война и мир" и "Анны Карениной".

С помощью команды `/rate` вы можете оценить качество ответа

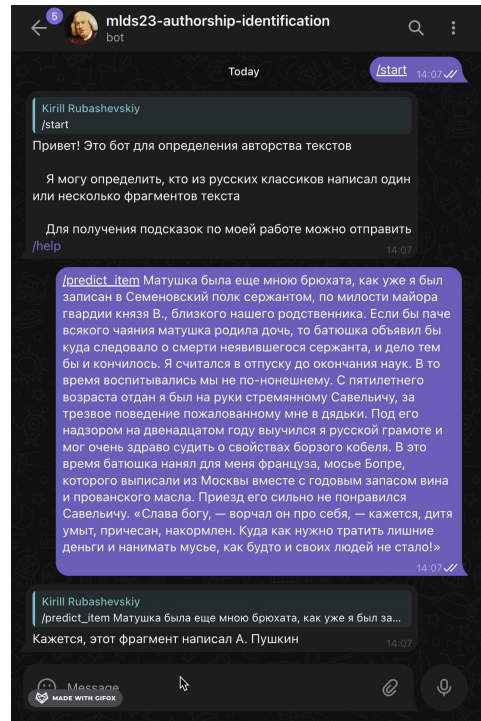
# Deployment

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