Первый этап

(разведочный анализ данных и первичная аналитика данных)

1. Изучить существующие подходы для решения подобных задач с помощью ML.
2. Выбрать 10–12 авторов среди русскоязычных писателей-классиков.
3. С помощью кода автоматически собрать датасет произведений (проза) выбранных авторов. Желательно, чтобы стили некоторых писателей были похожи между собой. Планируемый источник данных: https://www.culture.ru/literature/books
4. Разработать приложение для первичного анализа текстов, вычисляющее различные статистики: частоту употребления слов, словосочетаний (униграмм, биграмм, N-грамм), различных частей речи, служебных слов, среднюю длину слов, предложений — и визуализирующее результаты в виде таблиц, графиков и т.п. Провести анализ выбранных текстов.

Второй этап (ML)

1. Многоклассовый классификатор автора на основе простых текстовых эмбеддингов.
2. Визуализация близости авторов (лейблов) из корпусов (Толстой к Достоевскому статистически значимо ближе, чем Ильф и Петров)
3. Определение для каждого текста степени его похожести на тех или иных авторов (например, выводим топ-3 самых похожих авторов с %, насколько похожи)

Улучшение прогноза:

- Используем композицию разноплановых моделей; возможно с разной предобработкой исходных текстов

Третий этап (DL)

- Построение нейросетевого эмбеддинга автора текста (своеобразный TextSample2AuthorVec)