АНАЛИЗ ЧАСТОТНОСТИ, ЛЕКСИЧЕСКОГО РАЗНООБРАЗИЯ И ЧИТАБЕЛЬНОСТИ УГОЛОВНОГО КОДЕКСА

Проект Копыловой Любови

Описание

В проекте исследуются частотность слов, лексическое разнообразие (TTR и MATTR) и читабельность (по Флэшу, Оборневой и Мацковскому) Уголовного кодекса советского и российского периодов.

Материалы

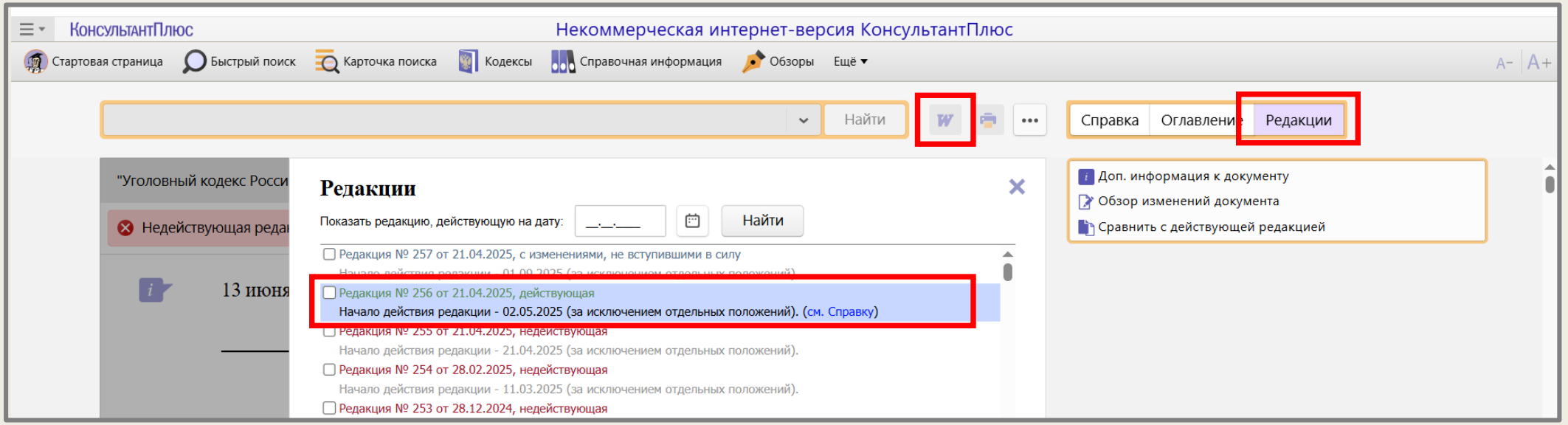
- **Уголовный кодекс РСФСР 1922 года** — первый советский кодекс, издавался в 5 редакциях.
- **Уголовный кодекс РСФСР 1926 года** — введен в действие в 1927 году, издавался в 14 редакциях.
- **Уголовный кодекс РСФСР 1960 года** — главный уголовный закон СССР и России, издавался в 4 редакциях.
- **Уголовный кодекс РФ от 1996 № 63-ФЗ** — действующий уголовный закон России, издавался в 257 редакциях.

Итого: **280** редакций Уголовного кодекса.

Задачи

Парсинг сайта «КонсультантПлюс»	selenium, tqdm
Работа с файлами и директориями	os, shutil
Предобработка текстов: предварительная очистка, удаление стоп-слов (NLTK), лемматизация (pymorphy3).	python-docx re nltk pymorphy3
Составление списка частотных слов	FreqDist из nltk
Анализ лексического разнообразия (метрики TTR и MATTR)	–
Оценка удобочитаемости, или Readability по формулам Флэша (адаптированная версия), И.В. Оборневой и М.С. Мацковского	–
Создание датафрейма и визуализация данных	pandas, matplotlib.pyplot, wordcloud

Сложности



Показаны основные клики кода при парсинге.

- Парсинг с помощью **selenium**, т.к. сайт динамический. Если смотреть код страницы, текста там нет. Поэтому проще скачать документ **в формате word.**
- Периодически код не выполнял клики — страницы не успевали погрузиться. Поэтому предусмотрена задержка по кликам на 10 секунд:

  wait = WebDriverWait(driver, 10)

- «КонсультантПлюс» ограничивает доступ к документам. Парсить можно только по будням после 20:00 и до 08:00 утра или в выходные дни.

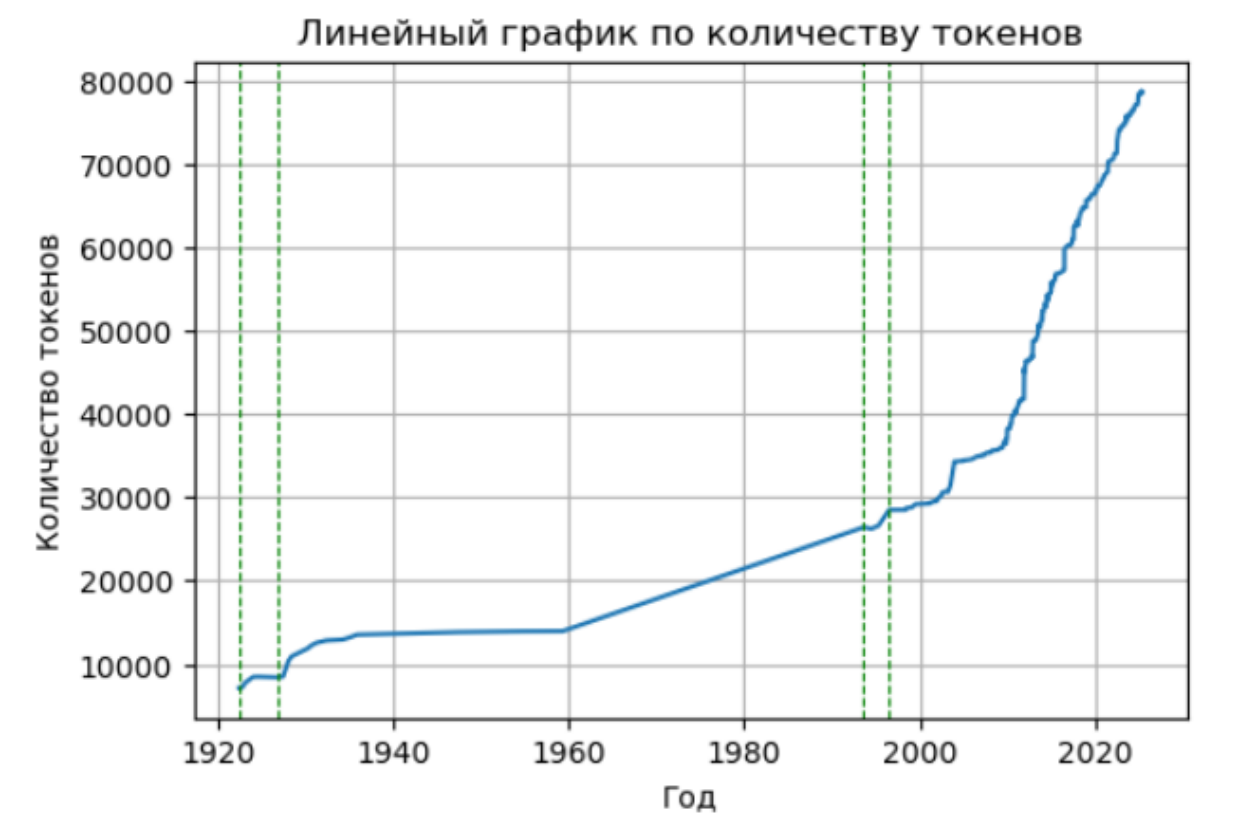
Сложности

- В процессе парсинга скачиваются 280 документов. Для удобства создается папка corpus, в которую перемещаются все нужные файлы (поиск по названию файла с помощью регулярных выражений).
- После лемматизации потребовалась дополнительная очистка от стоп-слов.
- Полное название и дату редации проще найти в метаданных файла docx (оказалось, что название записано в свойствах документа).
- Функция по предобработке на всем корпусе работает долго (около 2–3 часов). Поэтому, если она используется, данные в датафрейм добавляются долго.

Выводы

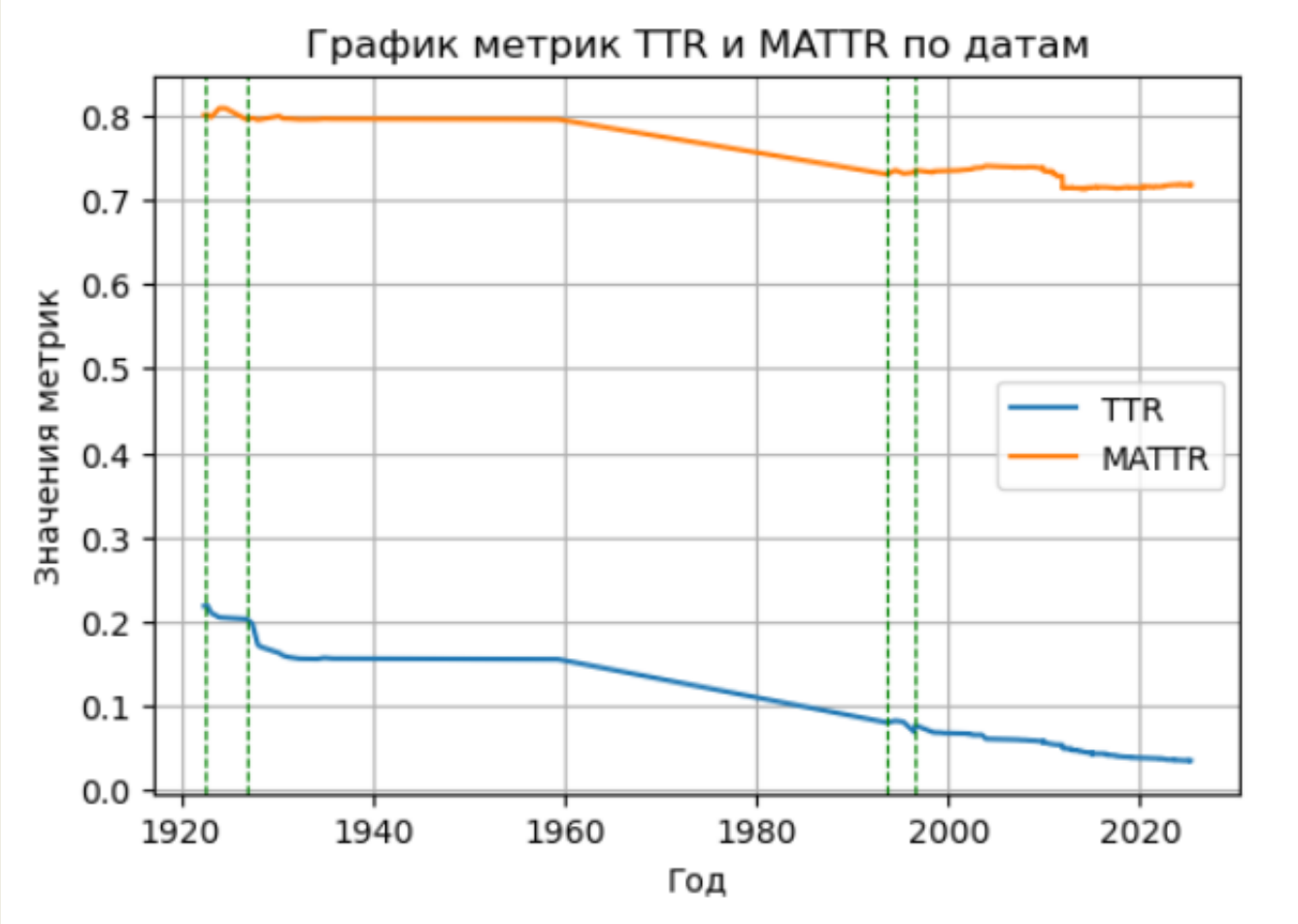
- Увеличение количества слов *(ожидаемо)*

Если сравнивать первый и последний уголовный кодекс — в 11 раз.



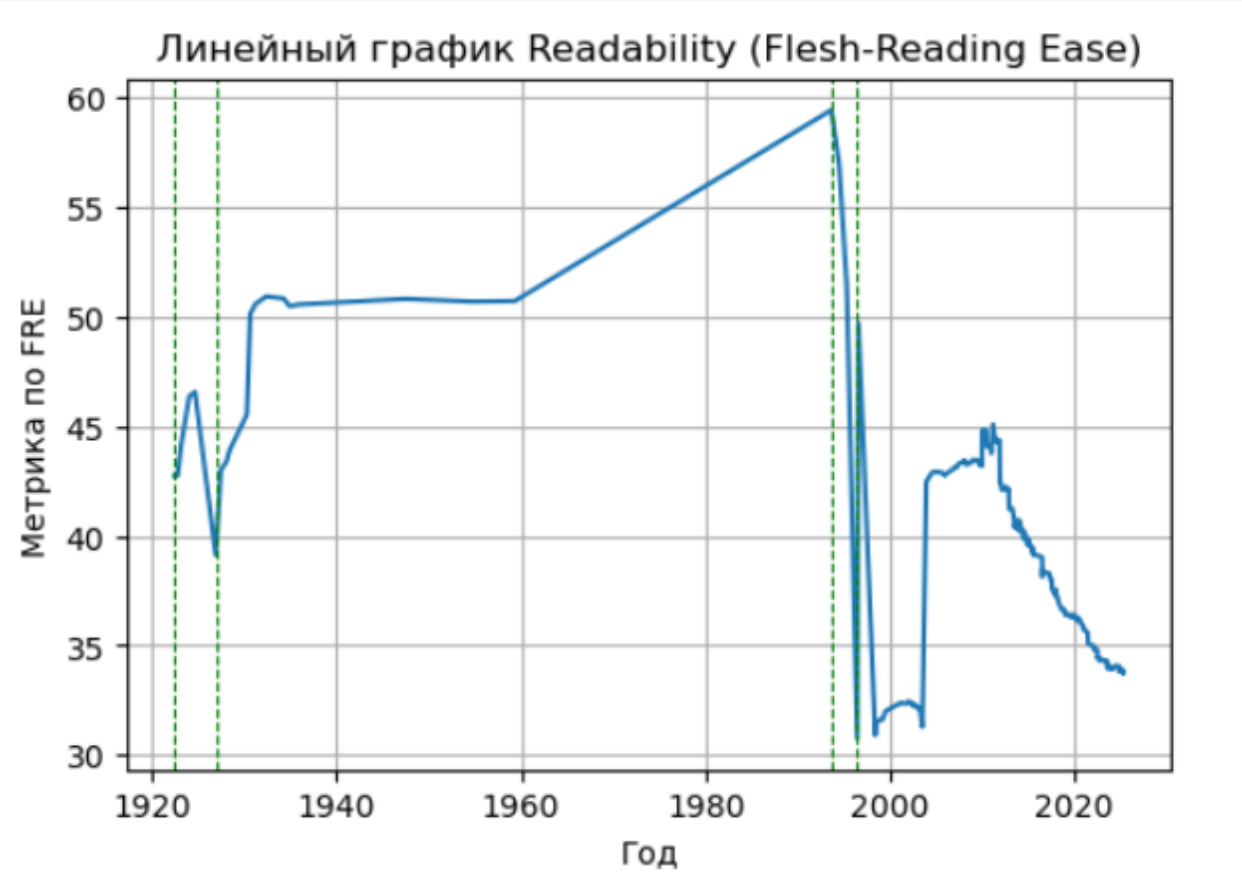
Выводы

- Снижение лексического разнообразия в текстах



Выводы

- Оценка удобочитаемости



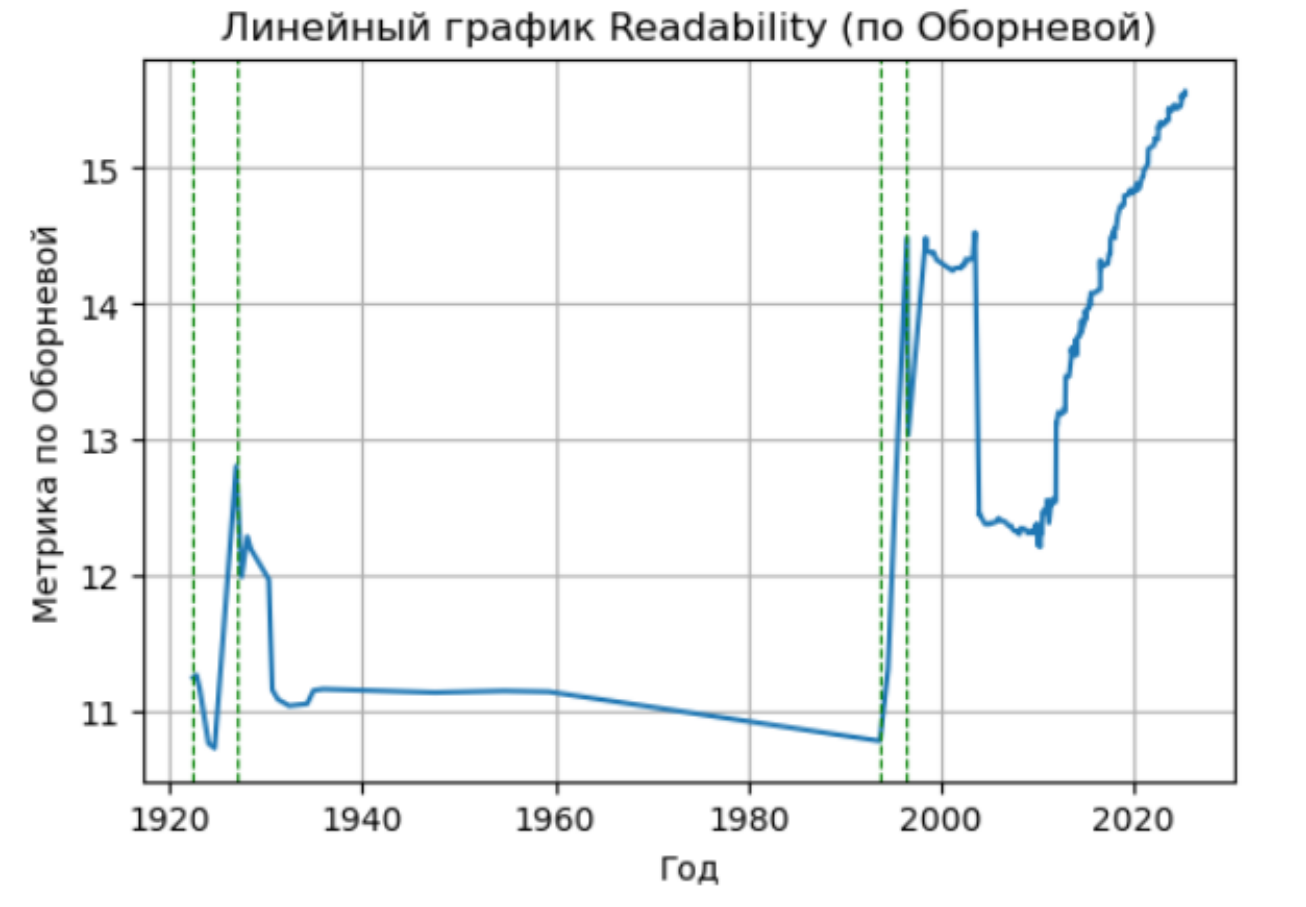
Тенденции

кодекс 1922 года (5 ред.):
упрощение → усложнение
кодекс 1926 года (14 ред.): упрощение
кодекс 1960 года (4 ред.): усложнение
кодекс РФ от 1996:
усложнение → упрощение → усложнение

50–59	Текст средней сложности, требует внимания и усилий при чтении	Для профессиональной аудитории
30–49	Трудный для понимания текст, требует хорошего владения языком	Для специалистов или академической аудитории
0–29	Очень сложный текст, предназначен для экспертов или научных публикаций	Для профессиональных ученых и экспертов

Выводы

- Оценка удобочитаемости

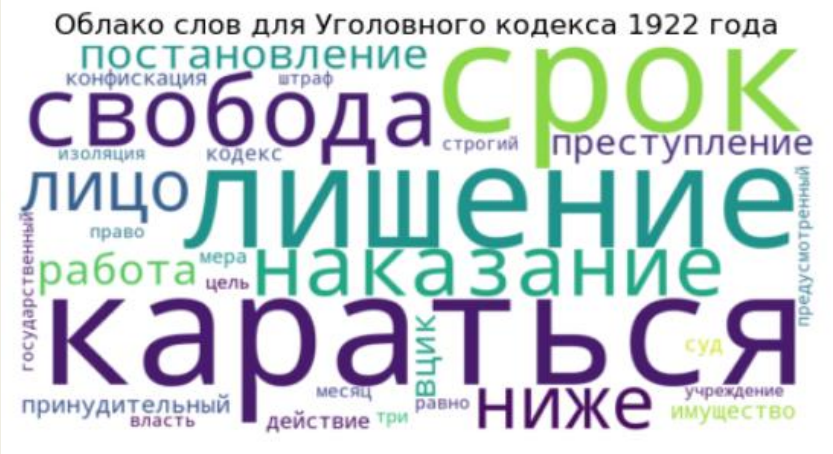


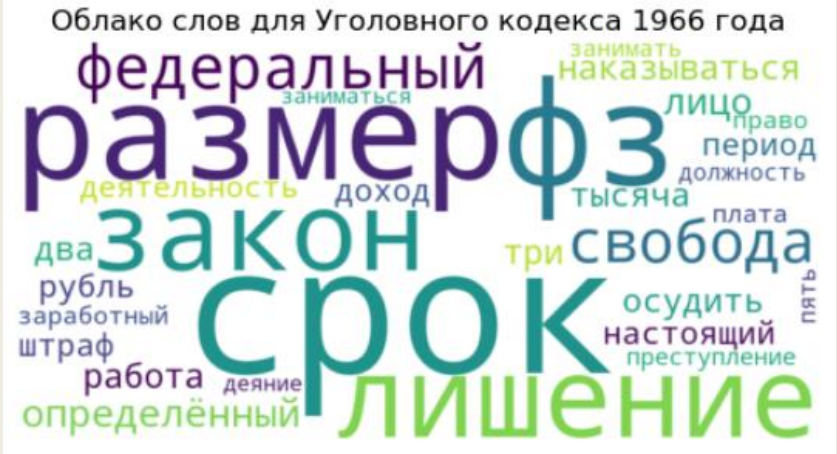
Тенденции зеркальны (?)
кодекс 1922 года (5 ред.):
усложнение → упрощение
кодекс 1926 года (14 ред.): усложнение
кодекс 1960 года (4 ред.): упрощение
кодекс РФ от 1996:
Упрощение → усложнение → упрощение

0–19	Очень сложный текст, предназначен для экспертов	Для ученых и профессионалов

Выводы

- Частотные списки (самые частые слова)





Выводы

- Частотные списки (малоупотребимые слова)