Анализ либретто мюзиклов на русском языке

Анастасия Бутакова
ДПО «Компьютерная лингвистика»

старт:

Насколько тексты похожи друг на друга, какие можно выделить тематики?

В ИТОГЕ:

...лексическое разнообразие?

Сбор корпуса - тексты либретто

1. Из открытых источников — если текст не файлом, то парсинг данных (Beautiful Soup), чтобы сразу было нужное форматирование
   a. Официальные тексты, выложенные в сеть
   b. Коллективные расшифровки, сделанные фанатами

2. Запросы к авторам, артистам, критикам, «старожилам» фандома
   a. Профи игнорируют или не готовы делиться материалами
   b. Авторы и театры перебрасывают друг на друга ответственность
   c. У любителей редко что-то есть

3. «Слитые» в сеть материалы

Все тексты оформлены совершенно по-разному и часто содержат доп.информацию, не нужную для анализа

72

текста собрано

30

использовано в проекте

Использованные библиотеки

os, re, docx2python, pdfminer.six

Превращение непонятно как оформленного текста непонятно с чем внутри в рабочий материал

Nltk, pymorphy3

Токенизация, лемматизация удаление стоп-слов; определение частей речи и проч

sklearn

Tf-idf
Косинусная близость

Matplotlib, Seaborn

Визуализация

Pandas, numpy

Работа с датафреймами и не только

Stanza, spacy

Попытки работать с именованными сущностями (тщетные)

Этапы работы

1. Сбор корпуса
2. Приведение текстов в единообразный вид (форматирование и метки) и сохранение в txt
3. Отсеивание ненужной информации из текста (названия песен, сценическая информация, повествовательные элементы, имена персонажей, которые говорят, технические пометки)
4. Остается только **разговорный текст** (песни, диалоги) => предобработка
5. Различные манипуляции: TF-IDF, косинусная близость, оценка лексического разнообразия

```
# Объединяем редактирование ВСЕХ возможных форматов файлов. Вариант с ОПЦИЯМИ
def v2_complete_formatting_woptions(filename, set_info, speakers, to_remove):
    '''Единая функция для форматирования текстов в формате text, docx/doc и pdf для частных случаев. С возможностью ввода специфических паттернов,
       которые позволят обработать конкретно этот текст.
       Настройка параметров:
       1. set_info - информация о сценографии: о действиях персонажей, технических моментах, характеристиках реплик типа "с подозрением", "тихо сме
          - 'none' - явного паттерна нет или все стандартно по нашей схеме
          - ...any custom value here... - паттерн, если есть единый стиль, которым обозначена эта информация
       2. speakers - говорящие/поющие:
          - 'none' - все в порядке, доп. действия здесь не требуется
          - 'absent' - если они не указаны в тексте вообще (поможет избежать ненужных исправлений)
          - ...any custom value here... - паттерн, если у спикеров есть явный паттерн, но не тот, что мы взяли по умолчанию
       3. to_remove - специфические пометки, которые захламляют текст и не нужны (например, номера страниц и колонтитулы в пдф, примечания в тексте
          - 'none' - пометки для удаления отсутствуют
          - ...any custom value here... - паттерн(ы) для удаления, задаются в виде списка
```

TF-IDF - провал?

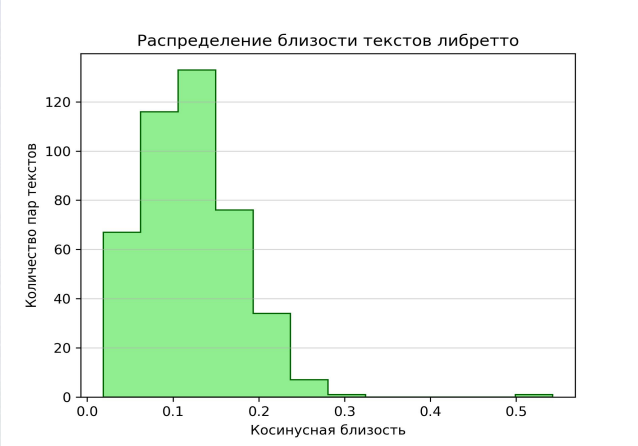
- Проблема именованных сущностей - они становятся ключевыми словами, но их сложно удалить
- Специфика стихотворного текста, которая жестко усугубляет проблему с NER
- Колебания длины текстов (есть немного гораздо более коротких, чем общая масса)

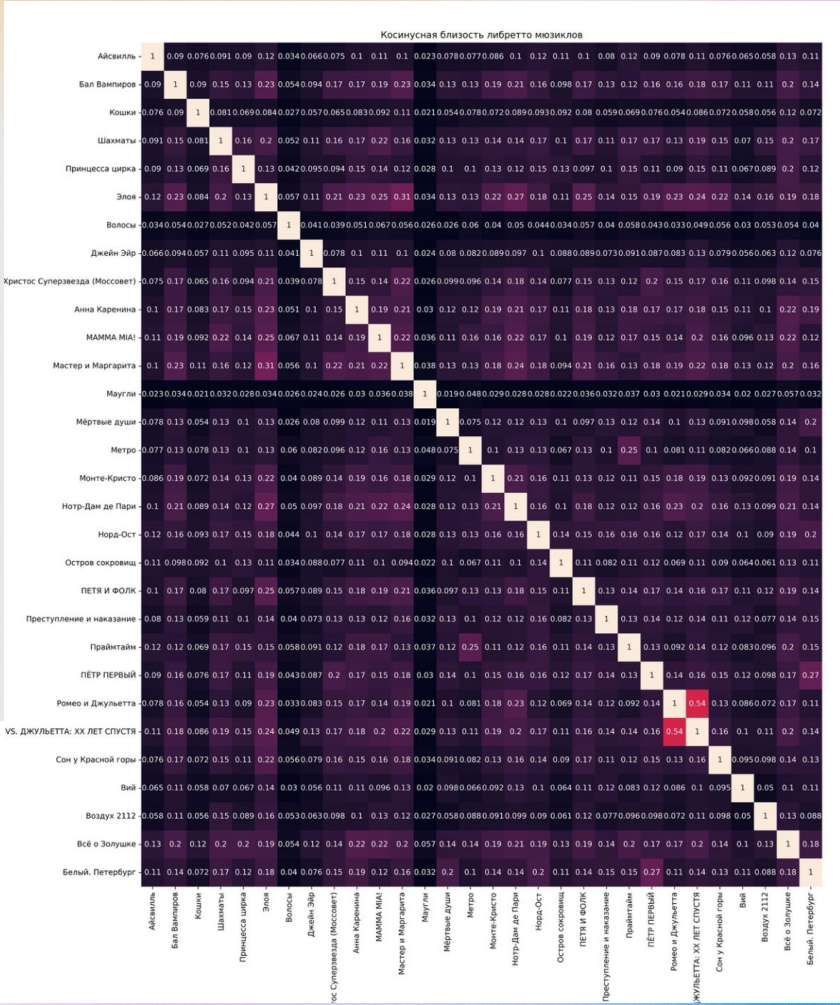
⇒ **чтобы сравнивать содержание векторизованных текстов , нужна еще более глубокая предобработка**

```
{('АББА', 383),
 ('Алёхин', 4317129024397789502),
 ('Анатолий', 4317129024397789502),
 ('Анатолий Сергиевский', 4317129024397789502),
 ('Анатолия', 4317129024397789502),
 ('Англии', 385),
 ('Андерсен', 385),
 ('Ареной идеологической борьбы', 383),
 ('Ах', 4317129024397789502),
 ('Бангкок', 385),
 ('Бангкоке', 385),
 ('Безумствуй', 4317129024397789502),
 ('Ботвинник', 4317129024397789502),
 ('Будапешт', 385),
 ('Будапеште', 385),
 ('Будь', 4317129024397789502),
 ('Быть', 4317129024397789502),
 ('Бьём', 4317129024397789502),
 ('Важно', 4317129024397789502),
 ('Ваш', 4317129024397789502),
 ('Вдруг', 4317129024397789502),
 ('Ведь', 4317129024397789502),
```


```
  "text": "Врёшь Поздно Лжец Поздно Поздно Сумел смогла одно я Есть ли кто-нибудь",
  "type": "MISC",
  "start_char": 30562,
  "end_char": 30632
},
{
  "text": "Мой триумф",
  "type": "MISC",
  "start_char": 30683,
  "end_char": 30693
},
{
  "text": "Дамы",
  "type": "PER",
  "start_char": 30715,
  "end_char": 30719
},
{
  "text": "Анатолий Сергиевский",
  "type": "PER",
  "start_char": 30730,
```

Косинусная близость текстов либретто



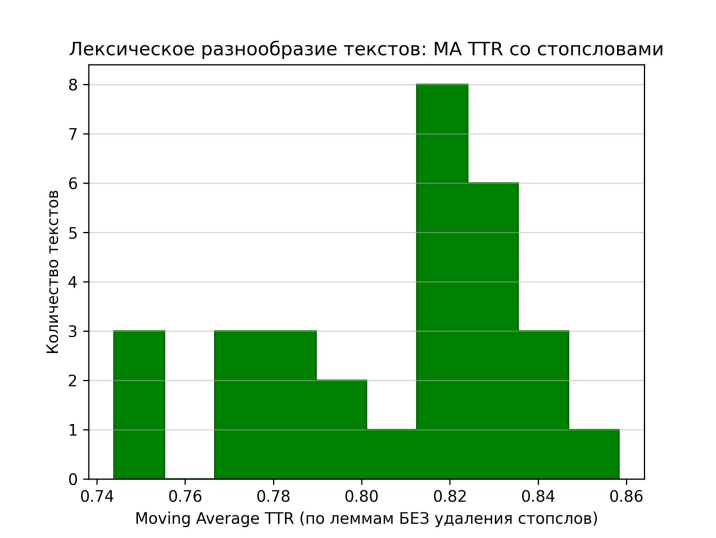


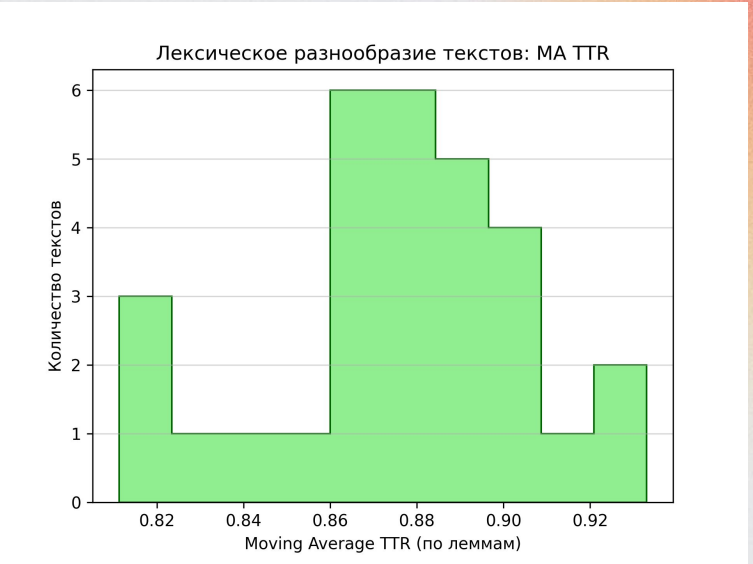
0.54257 - косинусная близость РОМЕО И ДЖУЛЬЕТТА и РОМЕО VS. ДЖУЛЬЕТТА: ХХ ЛЕТ СПУСТЯ
0.30709 - косинусная близость ЭЛОЯ и МАСТЕР И МАРГАРИТА
0.27332 - косинусная близость ПЁТР ПЕРВЫЙ и БЕЛЫЙ. ПЕТЕРБУРГ
0.26868 - косинусная близость ЭЛОЯ и НОТР-ДАМ ДЕ ПАРИ
0.24766 - косинусная близость МЕТРО и ПРАЙМТАЙМ
0.24691 - косинусная близость ЭЛОЯ и ПЕТЯ И ФОЛК
0.24638 - косинусная близость ЭЛОЯ и MAMMA MIA!
0.24461 - косинусная близость ЭЛОЯ и РОМЕО VS. ДЖУЛЬЕТТА: ХХ ЛЕТ СПУСТЯ
0.24274 - косинусная близость МАСТЕР И МАРГАРИТА и НОТР-ДАМ ДЕ ПАРИ

?

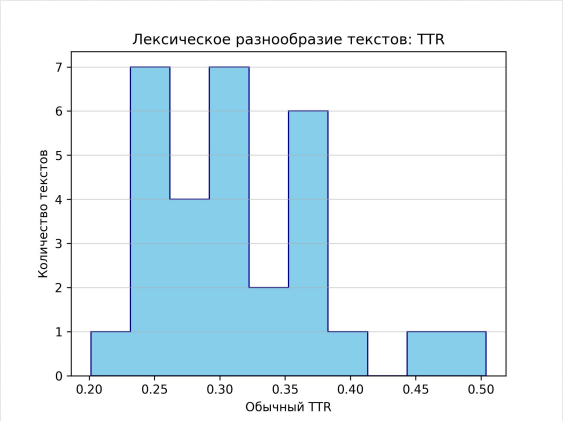
Окей, посмотрим что-то еще

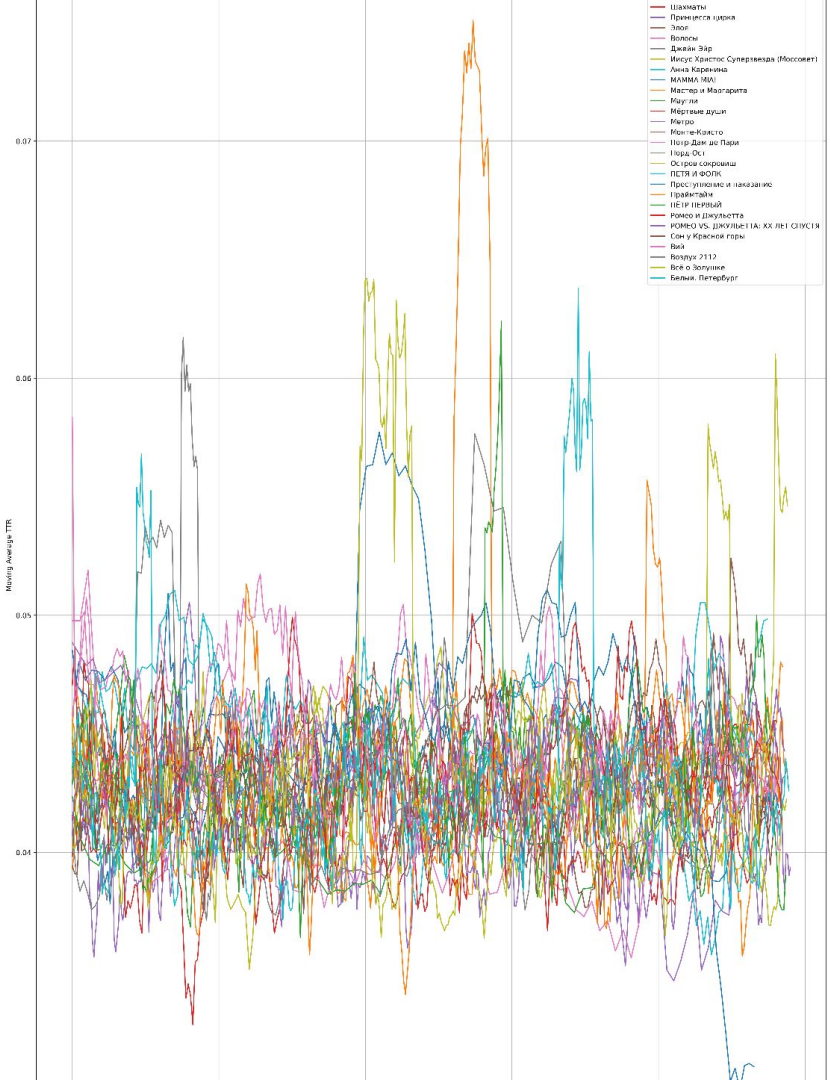
	Название	N токенов	N уникальных лемм	TTR	MA TTR	MA TTR (леммы)	MA TTR (леммы со стопсловами)
0	Айсвилль	5616	1654	0.294516	0.863413	0.877921	0.819641
1	Бал Вампиров	6231	1874	0.300754	0.863986	0.901561	0.820450
2	Кошки	4434	1636	0.368967	0.856078	0.899166	0.826878
3	Шахматы	5777	1534	0.265536	0.874654	0.922497	0.826201
4	Принцесса цирка	10282	2070	0.201323	0.870605	0.880896	0.824300
5	Элоя	6003	1475	0.245710	0.827608	0.871410	0.775724
6	Волосы	1407	709	0.503909	0.870692	0.874671	0.835228
7	Джейн Эйр	3698	1219	0.329638	0.871044	0.902343	0.822532
8	Иисус Христос Суперзвезда (Моссовет)	4056	1169	0.288215	0.831270	0.875627	0.782511
9	Анна Каренина	3931	1157	0.294327	0.786425	0.814414	0.743854
10	MAMMA MIA!	2429	740	0.304652	0.839529	0.866393	0.771655
11	Мастер и Маргарита	4034	1481	0.367129	0.897882	0.933068	0.858334
12	Маугли	1207	570	0.472245	0.869585	0.862733	0.826511
13	Мёртвые души	5556	1964	0.353492	0.861812	0.887866	0.831589

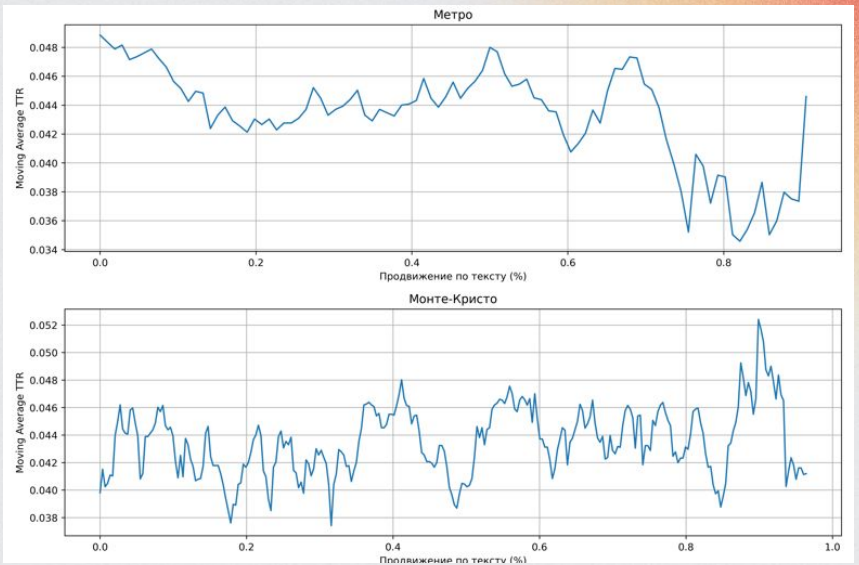




Лексическое разнообразие текстов




Шахматы
Принцесса цирка
Элог
Волосы
Джейн Эйр
Иисус Христос Суперзвезда (Моссовет)
Анна Каренина
МАММА МИА!
Мастер и Маргарита
Мы-ли
Мертвые души
Метро
Монте-Кристо
Нотр-Дам де Пари
Норд-Ост
Остров сокровищ
ПЭТЯ И ФОЛК
Преступление и наказание
Прайдстайн
ПЕТР ПЕРВЫЙ
Ромео и Джульетта
РОМЕО VS. ДЖУЛЬЕТТА: ХХ ЛЕТ СПУСТЯ
Сон у Красной горы
Вий
Воздух 2112
Всё о Золушке
Белый. Петербург
Moving Average TTR
0.07
0.06
0.05
0.04


Метро
Moving Average TTR
0.048
0.046
0.044
0.042
0.040
0.038
0.036
0.034
0.0
0.2
0.4
0.6
0.8
Продвижение по тексту (%)
Монте-Кристо
Moving Average TTR
0.052
0.050
0.048
0.046
0.044
0.042
0.040
0.038
0.0
0.2
0.4
0.6
0.8
1.0
Продвижение по тексту (%)