

ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ И СЕТЕВОЙ АНАЛИЗ ТЕКСТОВ ТРИЛОГИИ «ВЛАСТЕЛИН КОЛЕЦ» ДЖ.Р.Р.ТОЛКИНА

Надежда Фадеева

НИУ ВШЭ ДПО

Компьютерная лингвистика

СОДЕРЖАНИЕ

- Актуальность проекта
- Цели проекта
- Задачи проекта
- Алгоритм подготовки данных и NER
- Глобальное тематическое моделирование
- Локальное тематическое моделирование
- Сетевой анализ персонажей
- Инструменты
- Трудности
- Выводы

Актуальность проекта заключается в создании системы, которая автоматически извлекает социальную структуру и скрытые смыслы из масштабных текстов, превращая неструктурированную информацию в готовую базу данных для анализа.

Цель проекта

Провести комплексный анализ структуры трилогии «Властелин колец» с помощью алгоритмов машинного обучения для объективной реконструкции сюжета и системы персонажей.



ЗАДАЧИ ПРОЕКТА

1. Предобработка и извлечение сущностей

Очистка и лемматизация текста с автоматическим поиском персонажей (NER). Использование словаря маппинга для объединения вариаций имен в единые объекты анализа.

2. Глобальное тематическое моделирование

Выявление ключевых тем всей трилогии, а также связи героев с конкретными сюжетами, чтобы определить, в каких сюжетах чаще всего задействован каждый конкретный герой.

3. Локальное тематическое моделирование

Создание отдельных моделей для каждой книги. Сопоставление локальных тем с глобальными через косинусное сходство для анализа динамики сюжета.

4. Сетевой анализ персонажей

Построение графа связей и расчет метрик центральности. Математическое измерение популярности и реального влияния героев на структуру повествования.

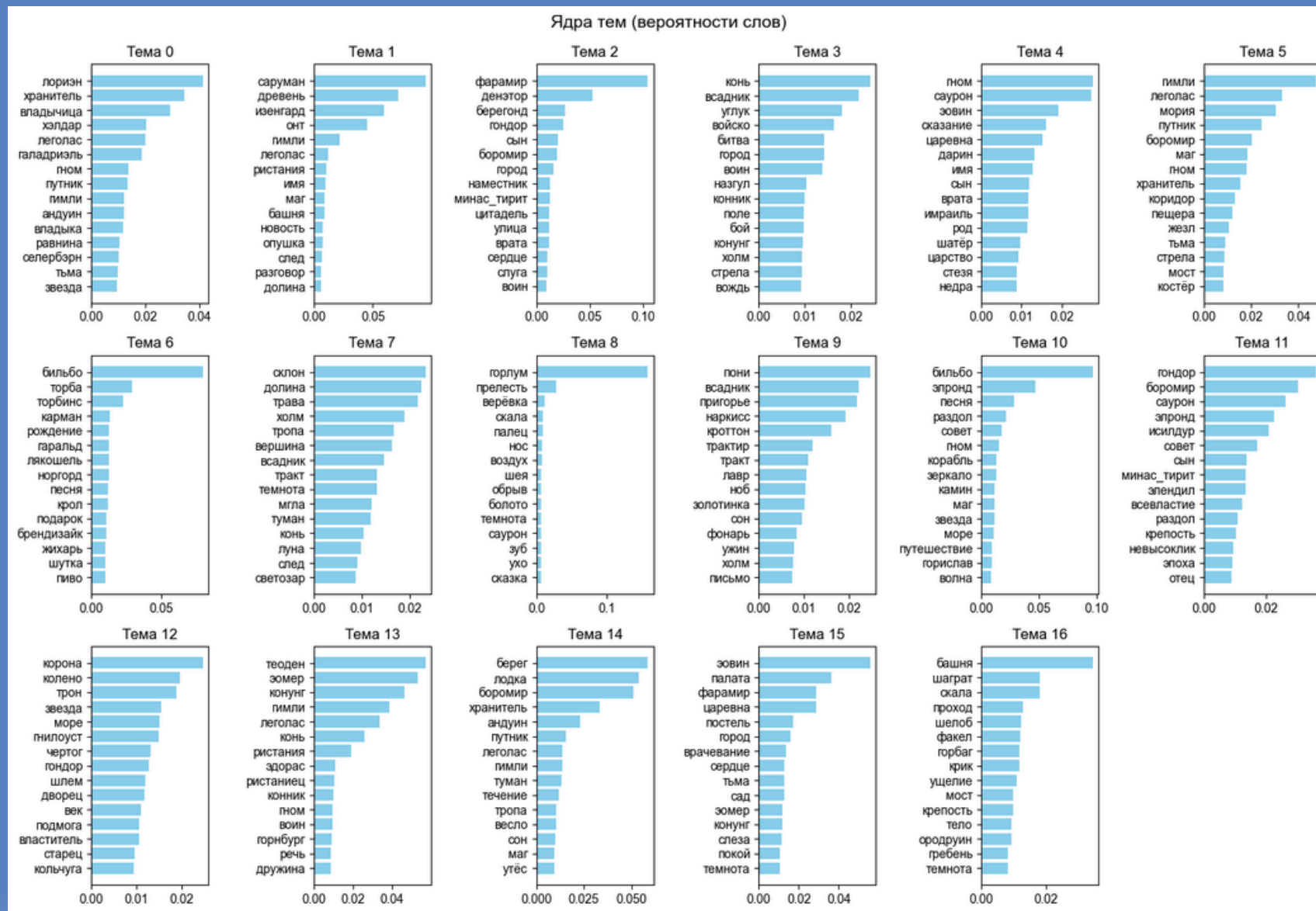
АЛГОРИТМ ПОДГОТОВКИ ДАННЫХ И NER

- **Многоуровневая сегментация:** Автоматическое деление текста на книги и главы. Нарезка глав на сегменты по 200 слов с перекрытием (50 слов) для сохранения связности сюжета при анализе.
- **Извлечение и нормализация сущностей (Natasha):** Автоматический поиск персонажей и локаций. Создание словаря замен для объединения прозвищ и склонений в единые токены.
- **Унификация составных названий:** Склеивание слов в именах и топонимах через нижнее подчеркивание для превращения их в неделимый объект.
- **Семантическая фильтрация (PyMorphy3):** Очистка текста от шума с сохранением только имен существительных. Это позволяет моделям LDA фокусироваться на объектах и смыслах, а не на действиях.

```
'гэндальфом': 'гэндальф',  
'гэндальфе': 'гэндальф',  
'гэндальфа': 'гэндальф',  
'гэндальф': 'гэндальф',  
'гэндальфу': 'гэндальф',  
'митрандир': 'гэндальф',  
'мирандиром': 'гэндальф',  
'митрандиру': 'гэндальф',  
'митрандира': 'гэндальф',  
'олорином': 'гэндальф',  
'старый маг': 'гэндальф',  
'гэндальф серый': 'гэндальф',  
'гэндальфа серого': 'гэндальф',  
'арагорн': 'арагорн',  
'арагорне': 'арагорн',  
'арагорну': 'арагорн',  
'арагорном': 'арагорн',  
'арагорна': 'арагорн',  
'колоброта': 'арагорн',  
'бродяжник': 'арагорн',  
'бродяжнику': 'арагорн',  
'бродяжником': 'арагорн',
```

```
'дол-амрота': 'дол_амрота',  
'дол амрота': 'дол_амрота',  
'боба': 'боб',  
'боб': 'боб',  
'бобом': 'боб',  
'бобу': 'боб',  
'брендизайки': 'брендизайк',  
'брендидуимскому мосту': 'брендидуимский_мост',  
'брендидуимском мосту': 'брендидуимский_мост',  
'брендидуимского моста': 'брендидуимский_мост',
```

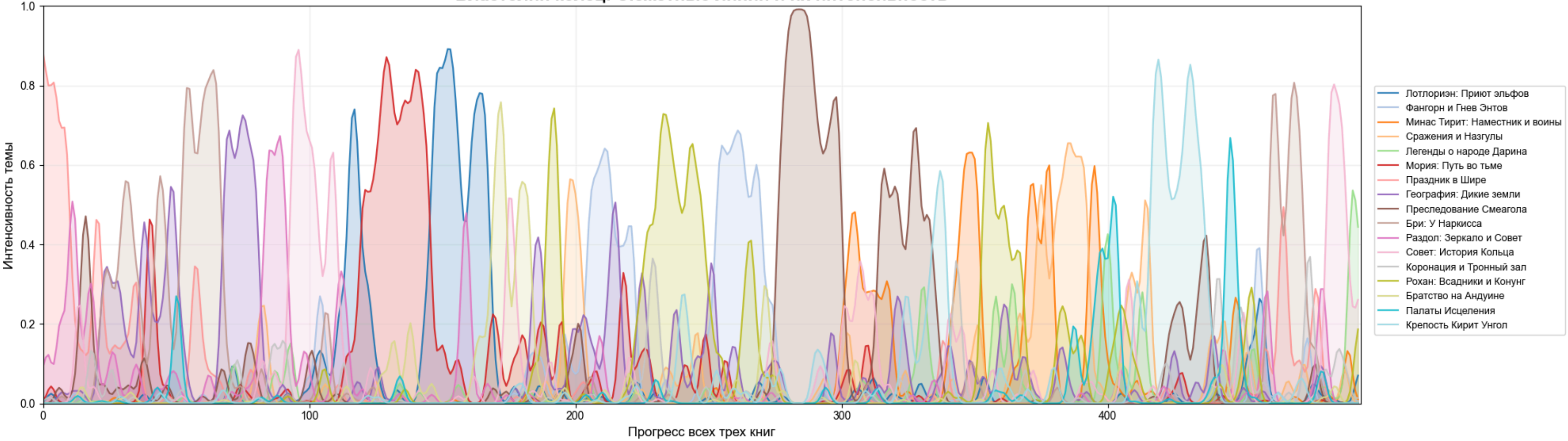
ГЛОБАЛЬНОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ



- **Оптимизация модели:** С помощью кросс-валидации (80/20) и анализа метрик когерентности (0.53) и перплексии (-6.43) выбрано оптимальное количество тем - 17.
- **Фильтрация шума:** Из словаря удалены слова, встречающиеся более чем в 30% сегментов и менее чем в 10 документах. Это позволило сфокусироваться на уникальных маркерах сюжета.
- **Валидация:** Topic Diversity (0.658): Умеренно высокая уникальность выделенных тем.
- **Анализ фоновых слов:** Подтверждено отсутствие семантического «мусора» во всех 17 темах.

Разница в длине полосок подтверждает, что модель нашла уникальные маркеры для каждой сюжетной линии, исключением можно назвать 7-ю тему, вероятно, это связано с тем, что тема является описательной (погода, природа).

Властелин колец: Сюжетные линии и их интенсивность



Последовательная смена разноцветных блоков показывает, что “Хранители” - линейное путешествие.

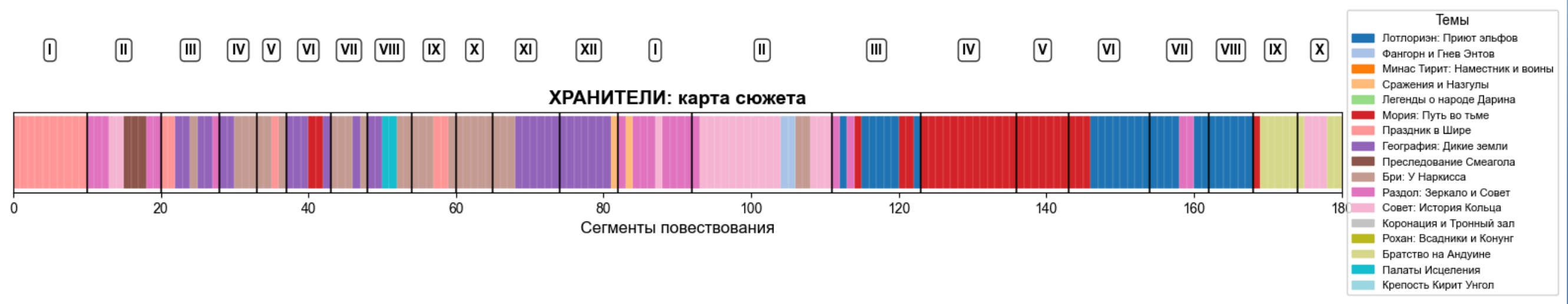
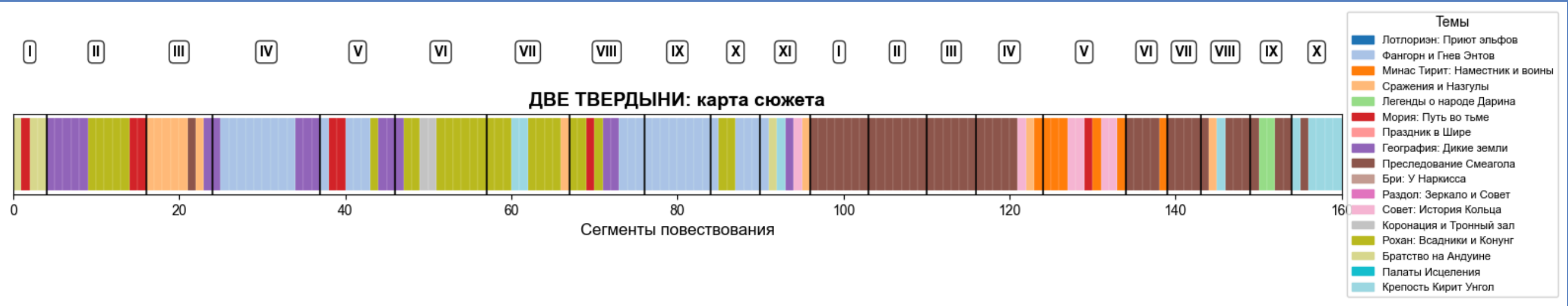
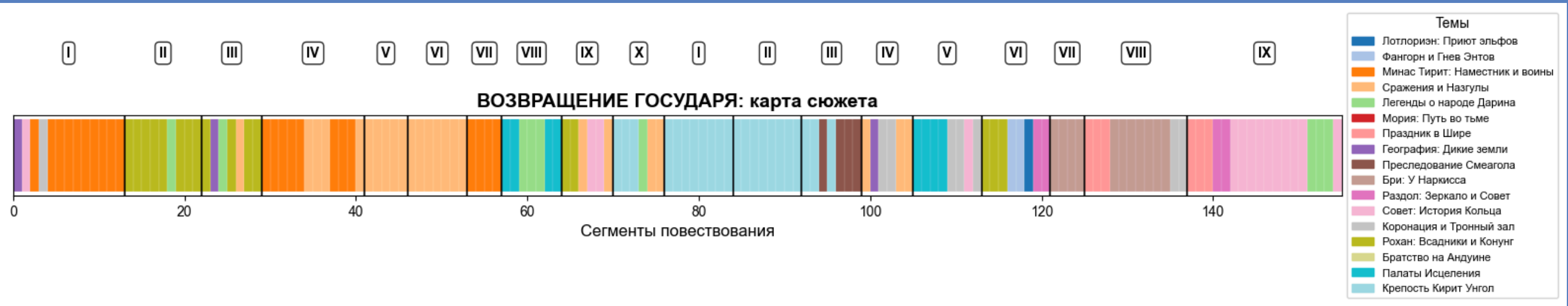
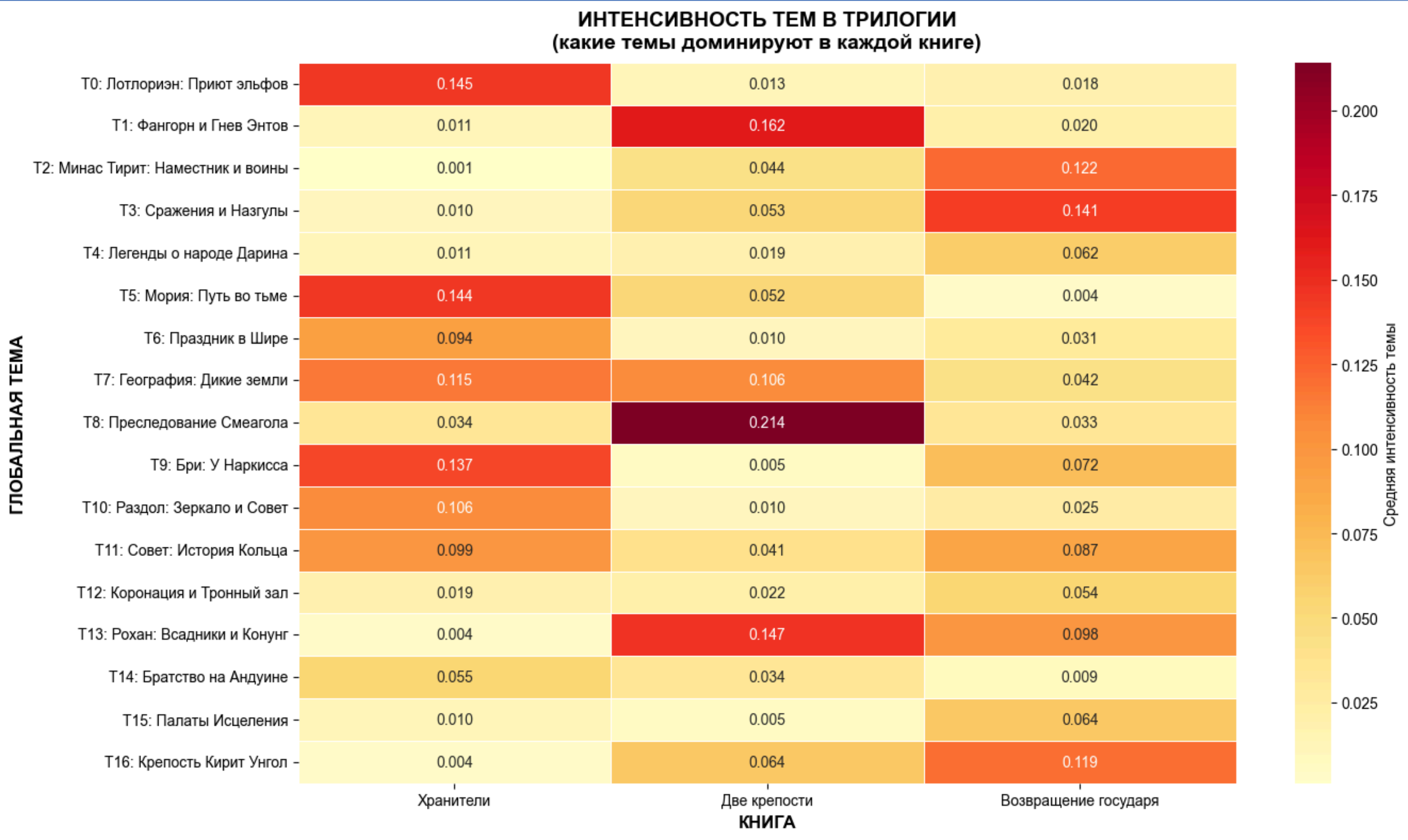


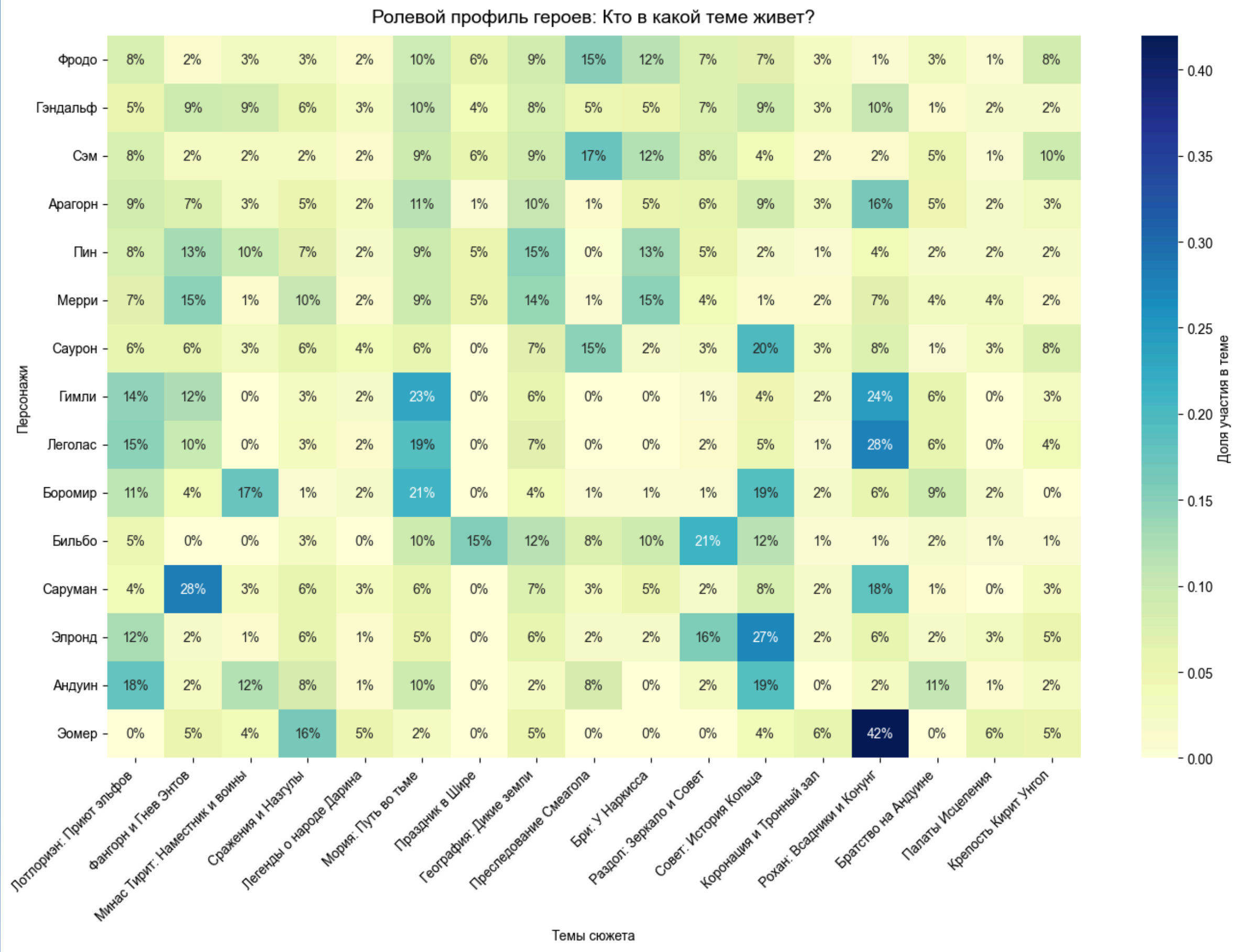
График четко распадается на две части: первая половина посвящена войне в Рохане, вторая - пути Фродо и Сэма.



Карта становится «полосатой», демонстрируя хаос войны, где разные темы перемешаны внутри одной главы.







ЛОКАЛЬНОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Для каждой книги подобрано индивидуальное число тем на основе пиков когерентности:

- “Хранители” - 7 тем (coherence = 0.4839)
- “Две твердыни” - 6 тем (coherence = 0.4840)
- “Возвращение Государя” - 7 тем (coherence = 0.4941)

Результат: Это позволило выявить уникальные сюжетные подтемы (например, «Очищение Шира»), которые в глобальном масштабе «забивались» более крупными событиями.

Вероятность до 99.6% в примерах чанков говорит о том, что темы получились «чистыми» - текст внутри сегмента почти полностью принадлежит к одной сюжетной линии.

ХРАНИТЕЛИ

ТЕМА 0: Совет в Ривенделле и история Кольца

Ключевые слова: кольцо, элронд, враг, эльф, бильбо, сила, мордор, саруман, саурон, совет

Всего чанков в теме: 38

1. Чанк #94 (Вероятность: 99.5%):

Текст: выход битва враг жизнь битва враг кольцо кольцо элронд раздол встреча смута кольцо судьба кольцо пора кольцо элронд элронд эпоха кольцо властелин мордор саурон

2. Чанк #93 (Вероятность: 99.5%):

Текст: тяга перемена гном одинокая_гора худо пристанище разговор царство язык гном казад дума труд отец владение мория глоин мечта гном край предок богатство слава нед

ТЕМА 1: Хоббитон: Праздник и наследство Бильбо

Ключевые слова: бильбо, год, кольцо, торба, хоббитания, мерри, торбинс, пора, лякошель, пин

Всего чанков в теме: 16

1. Чанк #2 (Вероятность: 99.4%):

Текст: сэм праздник потеха потеха хоббитания год пора крол норгорд фургон ящик торба хоббит темень лошадь гном капюшон песня торба неделя сентябрь брендиумский_мост

2. Чанк #5 (Вероятность: 99.4%):

Текст: шум бильбо цель цель тишина крол хоббит год гомон одобрение половина половина хлопок рождение гул племянник наследник фродо совершеннолетие владение имущество ф



КАК ГЛОБАЛЬНЫЕ ТЕМЫ РАСКЛАДЫВАЮТСЯ НА ЛОКАЛЬНЫЕ

Локальная тема 0: Совет в Ривенделле и история Кольца

Связана с глобальными темами:

Тема 11 (0.37): Совет: История Кольца

Тема 10 (0.34): Раздол: Зеркало и Совет

Локальная тема 1: Хоббитон: Праздник и наследство Бильбо

Связана с глобальными темами:

Тема 6 (0.83): Праздник в Шире

Тема 10 (0.67): Раздол: Зеркало и Совет

Локальная тема 2: Лотлориэн: Владычица и дары эльфов

Связана с глобальными темами:

Тема 0 (0.71): Лотлориэн: Приют эльфов

Тема 5 (0.37): Мория: Путь во тьме

Локальная тема 3: Бри: Следопыт и 'Гарцующий Пони'

Связана с глобальными темами:

Тема 9 (0.34): Бри: У Наркисса

Локальная тема 4: Дорога в Пригорье и Переправа

Связана с глобальными темами:

Тема 9 (0.53): Бри: У Наркисса

Локальная тема 5: Мория: Путь во тьме и гибель Мага

Связана с глобальными темами:

Тема 5 (0.68): Мория: Путь во тьме

Тема 14 (0.36): Братство на Андуне

Локальная тема 6: Странствия по Дикому Краю

Связана с глобальными темами:

Тема 7 (0.42): География: Дикие земли

Локальная тема 0: Мертвые болота: Фродо и прелесть Горлума

Связана с глобальными темами:

Тема 8 (0.67): Преследование Смеагола

Локальная тема 1: Рохан: Король Теоден и Гнилоуст

Связана с глобальными темами:

Тема 13 (0.63): Рохан: Всадники и Конунг

Тема 5 (0.34): Мория: Путь во тьме

Локальная тема 2: Итилия: Фарамир и память о Боромире

Связана с глобальными темами:

Тема 2 (0.59): Минас Тирит: Наместник и воины

Тема 11 (0.30): Совет: История Кольца

Локальная тема 3: Изенгард: Палантир и железная башня

Связана с глобальными темами:

Тема 1 (0.31): Фангорн и Гнев Энтов

Локальная тема 4: Фангорн: Древень и Гнев Энтов

Связана с глобальными темами:

Тема 1 (0.60): Фангорн и Гнев Энтов

Локальная тема 5: Плен: Урук-хаи и погоня по степи

Связана с глобальными темами:

Локальная тема 0: Осада Минас-Тирита: Наместник и воины

Связана с глобальными темами:

Тема 2 (0.51): Минас Тирит: Наместник и воины

Локальная тема 1: Стезя Мертвецов и Рохиррим

Связана с глобальными темами:

Тема 13 (0.58): Рохан: Всадники и Конунг

Локальная тема 2: Кирит Унгол и Роковая Гора

Связана с глобальными темами:

Локальная тема 3: Финал Эпохи: Корабли в Гавани

Связана с глобальными темами:

Тема 11 (0.45): Совет: История Кольца

Локальная тема 4: Палаты Исцеления: Врачевание и мир

Связана с глобальными темами:

Тема 15 (0.65): Палаты Исцеления

Тема 2 (0.40): Минас Тирит: Наместник и воины

Локальная тема 5: Очищение Ширы: Возвращение домой

Связана с глобальными темами:

Локальная тема 6: Совет Вождей и путь к Мордору

Связана с глобальными темами:

Тема 13 (0.50): Рохан: Всадники и Конунг

Тема 3 (0.34): Сражения и Назгулы

Хранители

Две твердыни

Возвращение Государя

СЕТЕВОЙ АНАЛИЗ ПЕРСОНАЖЕЙ

Сетевые метрики персонажей: Популярность vs Влияние			
Персонаж	Популярность	Связующая роль	Близость к центру
26 саурон	0.275	0.229	0.576
19 пин	0.475	0.187	0.650
52 лучиэнь	0.044	0.155	0.424
9 сэм	0.500	0.111	0.664
16 горлум	0.194	0.110	0.544
49 элронд	0.306	0.109	0.586
0 бильбо	0.356	0.086	0.597
61 балин	0.088	0.082	0.495
62 гимли	0.419	0.075	0.632
99 эомер	0.294	0.066	0.578
152 шаркич	0.056	0.064	0.473
34 брендидуим	0.075	0.061	0.498
28 орк	0.144	0.053	0.523
64 леголас	0.394	0.052	0.615
13 гэндальф	0.644	0.050	0.737
74 саруман	0.288	0.046	0.573
101 медусельд	0.069	0.041	0.475
29 арагорн	0.500	0.041	0.661
109 скоростень	0.056	0.040	0.471
3 фродо	0.650	0.039	0.741

Саурон имеет среднюю Популярность (0.275), но занимает 1-е место по Связующей роли (0.229) во всей трилогии. Он почти не появляется «в кадре» лично, но является главным узлом, через который связаны все остальные группы героев.

Хранители Две твердыни Возвращение государя			
Персонаж			
пин	0.076	0.129	0.287
сэм	0.278	0.108	0.019
элронд	0.026	0.297	0.041
саурон	0.064	0.000	0.210
фродо	0.101	0.069	0.087
бильбо	0.244	0.000	0.012
горлум	0.022	0.233	0.000
балин	0.220	0.000	0.000
леголас	0.055	0.150	0.015
гэндальф	0.076	0.051	0.092
арагорн	0.063	0.018	0.123
саруман	0.029	0.160	0.000
мерри	0.048	0.106	0.029
гимли	0.002	0.144	0.032
элендил	0.019	0.000	0.155

В «Двух Твердынях» пик влияния у Элронда (0.297) и Горлума (0.233). Это подтверждает, что в середине пути сюжет держится на предыстории и проводнике в Мордор. Влияние Сэма падает с 0.278 до 0.019 в конце: в финале Сэм отрезан от мира и замкнут только на Фродо.

ИНСТРУМЕНТЫ

Язык: Python

NLP и NER: Natasha, PyMorphy3, NLTK

Тематическое моделирование: Gensim (LDA), pyLDAvis, Scikit-learn.

Сетевой анализ: NetworkX, Itertools.

Анализ и данные: Pandas, NumPy, Matplotlib, Seaborn, Pickle.

ТРУДНОСТИ

1. Проекту нет конца и края

Улучшать модель можно вечно: любое изменение фильтров (вроде `no_below` или `no_above`) или добавление новых стоп-слов мгновенно меняет когерентность. Всегда казалось, что можно добавить ещё одну деталь, чтобы еще сильнее поднять когерентность.

2. Проблемы с TF-IDF

Метод TF-IDF, который должен был помочь, на деле всё испортил. Темы перемешались, и в них стали повторяться одни и те же слова. В итоге от него было решено отказаться, чтобы не терять чистоту и различия между сюжетами.

3. Неудачный сентимент-анализ

Готовые модели для русского языка обучались на коротких отзывах и твитах. Они оказались совершенно не готовы к языку Толкина и выдавали плохие результаты (в частности, корреляция была всего 0.006), поэтому от этой идеи пришлось отказаться.

ВЫВОДЫ

Качество и валидация глобальной модели

- Метрика когерентности (Coherence Score: 0.53) и перплексия (-6.43) указывают на то, что модель успешно выделила интерпретируемые темы без «мусора».
- Показатель Topic Diversity: 0.66 и отсутствие фоновых слов подтверждают, что алгоритм четко разграничивает локации и сюжетные линии.

Релевые профили героев

- Анализ подтвердил функциональное разделение ролей: Фродо и Сэм максимально сосредоточены в темах пути к Кольцу, в то время как Гэндальф и Арагорн равномерно распределены по ключевым военно-политическим сюжетам.
-

Качество и валидация локальных моделей

- Средний показатель когерентности ~0.49 подтверждает высокую содержательную плотность локальных кластеров.

Сетевой анализ

- Обладая умеренной популярностью (Degree=0.275), Саурон имеет наивысший показатель связующей роли (0.229) во всей трилогии.
- Гэндальф остается самым популярным персонажем по количеству связей (Degree=0.644).
- Лидером влияния становится Элронд (0.297), поскольку герои в разных частях Средиземья постоянно ссылаются на его решения и Совет. При этом Горлум (0.233) становится ключевым «мостом» для Фродо и Сэма.

СПАСИБО ЗА ВНИМАНИЕ!

Проект на Github:

<https://github.com/NadFadeeva/LOTR-LDA-and-Sentiment-Analysis>