Тематическое моделирование: выявление ключевых тем и их героев в романах Терри Пратчетта

КОРБАН АНДРЕЙ

28-06-2025

АКТУАЛЬНОСТЬ ТЕМЫ:

Анализ больших текстовых корпусов становится всё более востребованным.
Романы Терри Пратчетта содержат сложные смысловые структуры, темы и персонажей, которые сложно выявить традиционными методами филологического анализа.
Тематическое моделирование с помощью машинного обучения позволяет автоматизировать анализ, раскрыть скрытые паттерны и закономерности в построении литературных нарративов, а также получить новые инструменты для лингвистических и литературоведческих исследований.

ЦЕЛИ ИССЛЕДОВАНИЯ:

Автоматизированное выявление и анализ ключевых тем и связанных с ними персонажей в романах Терри Пратчетта с использованием современных методов тематического моделирования.

ОБЪЕКТ ИССЛЕДОВАНИЯ:

Корпус англоязычных художественных романов Терри Пратчетта, выбранных для анализа (20 произведений), а также ключевые темы и персонажи, выявляемые с помощью методов машинного обучения.

ИСПОЛЬЗУЕМЫЕ БИБЛИОТЕКИ PYTHON

РАБОТА С ДАННЫМИ:

- pandas: Анализ и манипуляция табличными данными, создание DataFrame для хранения корпуса текстов и метаданных
- numpy: Поддержка многомерных массивов и математических функций для обработки данных
- collections (Counter, defaultdict): Специализированные структуры данных для подсчета частоты элементов и создания вложенных словарей
- os, re: Базовые модули для работы с файловой системой и регулярными выражениями

ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА:

- spacy: Комплексный фреймворк для NLP для извлечения именованных сущностей (NER)
- nltk (Natural Language Toolkit): набор инструментов для для токенизации и работы со стоп-словами
- sentence_transformers: Библиотека для создания векторных представлений предложений, преобразование текстов в числовые векторы
- bertopic: Реализация тематического моделирования на основе BERT, основной инструмент для выделения тем в текстах

МАШИННОЕ ОБУЧЕНИЯ И КЛАСТЕРИЗАЦИЯ:

- sklearn (CountVectorizer, normalize): Преобразование текста в числовые признаки и нормализация данных
- umap-learn: Алгоритм снижения размерности для визуализации данных
- hdbscan: Алгоритм кластеризации, используемый для группировки текстов по темам

ВИЗУАЛИЗАЦИЯ:

- matplotlib и seaborn: Создание статических визуализаций, графиков и тепловых карт
- plotly: Создание интерактивных визуализаций результатов тематического моделирования
- networkx: Анализ и визуализация сетей связей между персонажами и темами
- wordcloud: Создание облаков слов для визуализации ключевых слов в темах
- tqdm: Отображение прогресс-баров для длительных операций обработки

ОБОСНОВАНИЕ ПОДХОДА:

1. Комплексный анализ:
Код объединяет различные методы NLP и машинного обучения для многостороннего анализа литературных произведений.

2. Использование современных моделей:
Применение BERT-подобных моделей обеспечивает высокое качество векторизации текстов и выделения семантически значимых тем.

3. Визуализация результатов:
Различные методы визуализации позволяют наглядно представить сложные взаимосвязи и паттерны в текстах.

4. Масштабируемость:
Подход может быть применен к другим корпусам текстов с минимальными изменениями.

КЛЮЧЕВЫЕ ЭТАПЫ ОБРАБОТКИ В КОДЕ:

1. Подготовка данных:
- Чтение текстовых файлов
- Разделение текстов на структурные элементы
- Извлечение именованных сущностей (персонажей)

2. Создание тематической модели:
- Настройка BERTopic с оптимальными параметрами
- Векторизация текстов с помощью трансформеров
- Снижение размерности с помощью UMAP
- Кластеризация с помощью HDBSCAN
- Выделение ключевых слов для каждой темы

3. Анализ связей:
- Создание матрицы связей между персонажами и темами
- Нормализация данных для лучшей интерпретации
- Анализ распределения тем по книгам

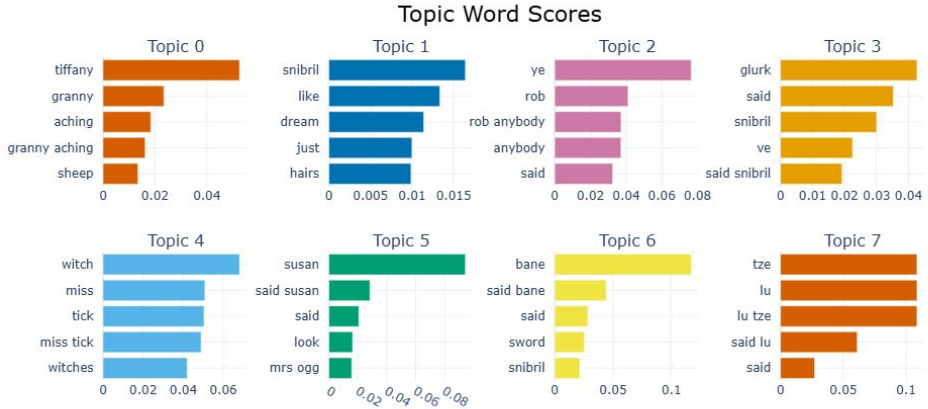
4. Визуализация результатов:
- Интерактивные визуализации тем и их иерархии
- Тепловые карты связей персонажей и тем
- Сетевые графы для отображения связей
- Анализ динамики тем по книгам

АНАЛИЗ ТЕМАТИЧЕСКИХ КЛАСТЕРОВ

ВЫВОДЫ:

1. Персонажецентричность тем:
Большинство тем сформированы вокруг конкретных персонажей (Тиффани, Снибрил, Лу-Цзе, Сьюзен),
что подтверждает важность персонажей в структуре повествования.

2. Четкое разделение по циклам:
Алгоритм успешно выделил темы, соответствующие разным циклам:
- Цикл о Тиффани Болен (Topics 0, 4)
- Трилогия о Номах (Topics 1, 3)
- Роман "Вор времени" (Topic 7)
- Цикл о Смерти и его внучке Сьюзен (Topic 5)



3. Пересечение тематик: Некоторые темы показывают интересные пересечения, например, появление "snibril" в Topic 6 вместе с "bane" и "sword", что может указывать на тематические связи между разными произведениями.

4. Стилистические особенности: В некоторых темах (например, Topic 2) выделяются стилистические особенности речи персонажей, что говорит о способности алгоритма улавливать не только сюжетные, но и лингвистические паттерны.

5. Доминирующие темы: По длине горизонтальных баров можно судить о значимости каждого слова в теме. Наиболее высокие значения имеют "tze" и "lu" в Topic 7, "witch" в Topic 4, и "tiffany" в Topic 0, что может указывать на центральность этих персонажей в соответствующих текстах.

Эта визуализация демонстрирует, как автоматические методы анализа текста могут выявлять тематические структуры в литературных произведениях и подтверждать литературоведческие наблюдения о взаимосвязи персонажей и сюжетных линий.

АНАЛИЗ ТЕПЛОВОЙ КАРТЫ СВЯЗЕЙ ПЕРСОНАЖЕЙ И ТЕМ

АНАЛИЗ ТЕМАТИЧЕСКИХ КЛАСТЕРОВ:

1. Ведьмовская тема (тема 0): Объединяет персонажей Granny, Granny Aching, Tiffany, что, вероятно, соответствует цикл книг о ведьмах Плоского мира. Сюда также попадают Wentworth и Baron, что указывает на их связь с сюжетными линиями ведьм.

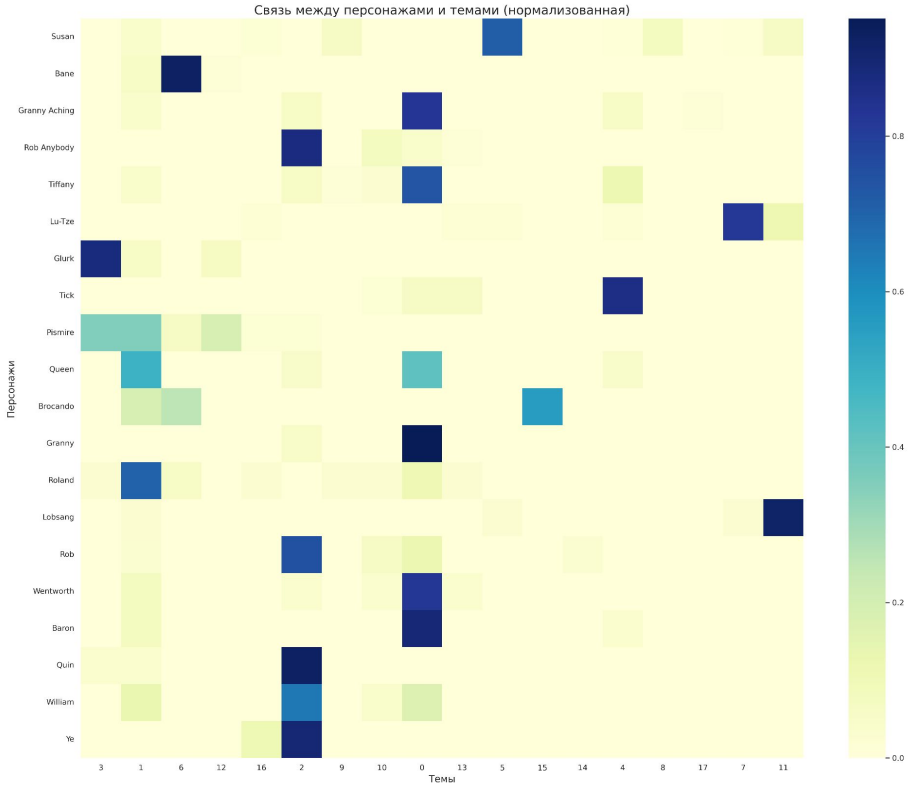
2. Нак Мак Фигли (тема 2): Объединяет Rob Anybody и других персонажей, связанных с маленьким свободным народцем из серии о Тиффани Боль.

3. Тема времени (темы 11 и 17): Связь Susan и Lobsang с темой 11, а Lu-Tze с темой 17 может указывать на сюжетные линии, связанные с Аудиторами реальности и монахами Времени (из романов "Вор времени" и "Санта-Хрякус").

4. Городские персонажи (темы 1, 3, 4): Персонажи вроде Glurk, Roland, Queen связаны с темами, которые могут относиться к городским сюжетным линиям, возможно к Анк-Морпорку или другим городам Плоского мира.

ВЫВОДЫ:

1. Многие персонажи имеют четкую тематическую привязку, что характерно для его писательского стиля.
2. Видны отчетливые тематические кластеры, соответствующие основным сюжетным циклам Плоского мира:
3. Некоторые персонажи имеют связи с несколькими темами, что отражает их участие в различных сюжетных линиях на протяжении серии книг.
4. Интенсивность связей (темно-синий цвет) показывает, насколько глубоко персонаж ассоциируется с определенной темой, что может указывать на его ключевую роль в развитии соответствующих сюжетных линий.



АНАЛИЗ ДЕНДРОГРАММЫ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

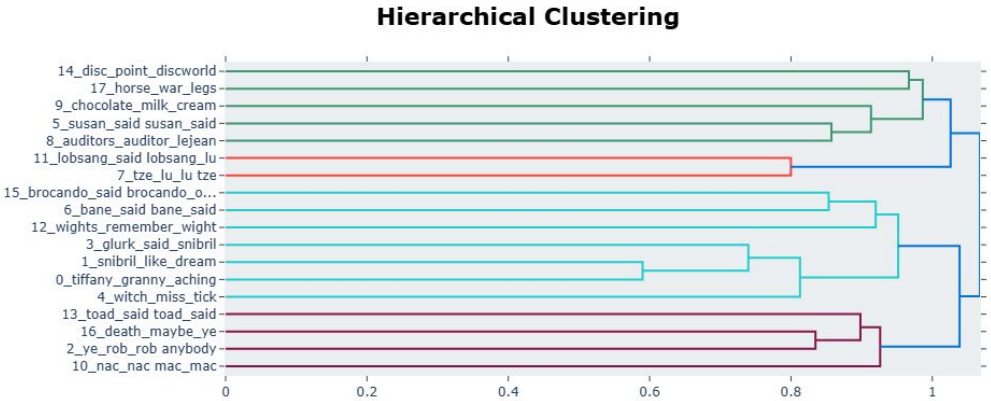
ОСНОВНЫЕ ВЫЯВЛЕННЫЕ КЛАСТЕРЫ

Кластер 1 (голубой цвет, верхняя часть). Элементы: 14, 17, 9, 5
Этот кластер объединяет элементы, связанные с общей вселенной Discworld и Сьюзен Смерть. Присутствие темы "лошади Войны" может указывать на связь с всадниками Апокалипсиса из цикла книг.

Кластер 2 (красный цвет, средняя часть). Элементы: 8, 11, 7, 15
Этот кластер явно связан с романом Thief of Time, где фигурируют Аудиторы, Лобсанг, Лу-Цзе. Это показывает, что тексты, связанные с этим романом, имеют достаточно уникальную лексику и тематику.

Кластер 3 (голубой цвет, средняя часть). Элементы: 6, 12, 3, 1, 0, 4
Здесь объединены различные сюжетные линии.

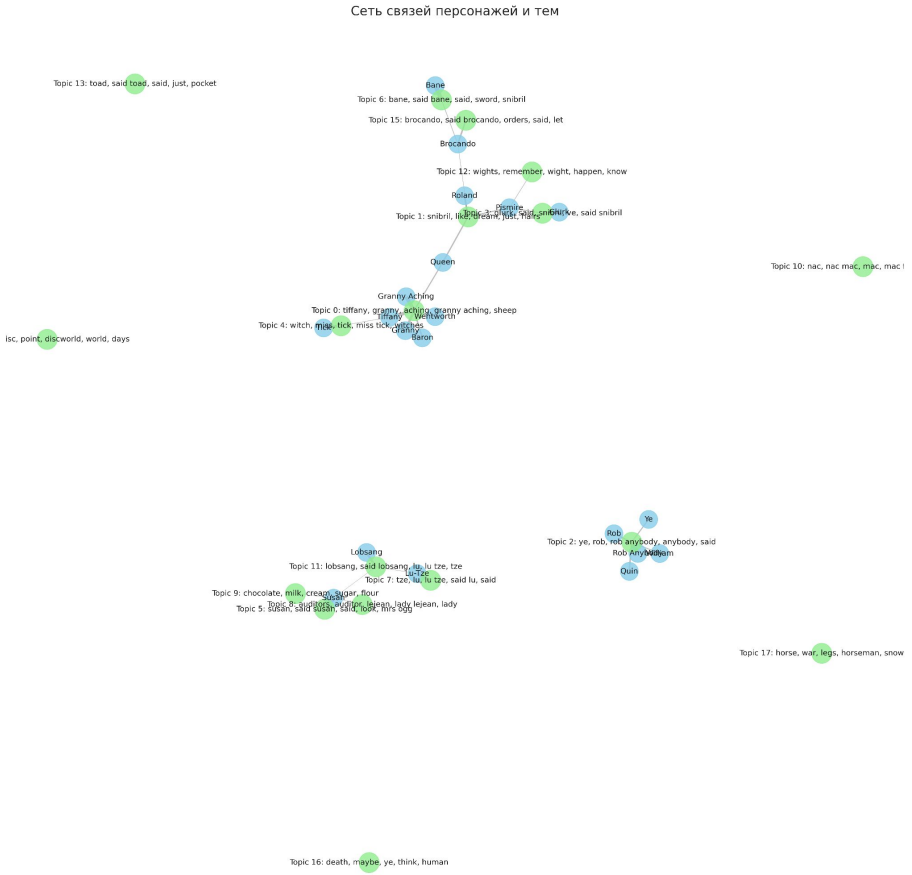
Кластер 4 (пурпурный цвет, нижняя часть). Элементы: 13, 16, 2, 10
Этот кластер сложнее интерпретировать



ВЫВОДЫ ИЗ АНАЛИЗА

1. Структурированность произведений Пратчетта:
различные циклы и романы имеют свою уникальную лексику и стиль, что позволяет их четко дифференцировать.
2. Тематические связи:
Можно наблюдать, как персонажи и темы, относящиеся к одним и тем же произведениям, группируются вместе, подтверждая связность повествования в рамках отдельных циклов.
3. Сложность вселенной Плоского мира:
Несмотря на то, что все элементы относятся к произведениям одного автора, иерархическая структура показывает многослойность и разнообразие созданного Пратчеттом мира.
4. Персонажецентричность:
Многие кластеры формируются вокруг имен персонажей, что подчеркивает важность персонажей в произведениях Пратчетта.

АНАЛИЗ СЕТЕВОЙ ДИАГРАММЫ СВЯЗЕЙ ПЕРСОНАЖЕЙ И ТЕМ



СТРУКТУРНЫЕ ОСОБЕННОСТИ СЕТИ:

1. Центральные и периферийные темы:
   - Центральные темы (с большим количеством связей): Topic 0, Topic 2, Topic 11
   - Периферийные темы (изолированные): Topic 13, Topic 14, Topic 16

2. Плотность связей:
   - Наиболее плотные связи в кластере ведьм и кластере Нак Мак Фиглей
   - Некоторые персонажи имеют связи с несколькими темами (например, Lu-Tze)

3. Изолированные узлы:
   - Некоторые темы представлены без прямой связи с персонажами в сети, что может указывать на темы, распределенные по многим персонажам, но не характерные для какого-то одного

ВЫВОДЫ:

1. Тематическая согласованность: персонажи группируются вокруг определенных тематических циклов, что отражает структуру серии "Плоский мир".

2. Циклы Плоского мира: можно четко выделить основные циклы - ведьмы, монахи Времени, истории о Смерти и его внучке, Нак Мак Фигли.

3. Лексические особенности: ключевые слова в темах часто включают имена персонажей и диалоговые маркеры ("said"), что указывает на важность диалогов в произведениях Пратчетта.

4. Уникальные темы: Некоторые темы, такие как еда (Topic 9) или общие описания Плоского мира (Topic 14), часть всех произведений, не привязываясь к конкретным персонажам.

ПРИМЕНЕНИЕ ИНТЕРАКТИВНОЙ INTERTOPIC DISTANCE MAP

Intertopic Distance Map (карта расстояний между темами) — это инструмент визуализации для анализа тематического моделирования, особенно при использовании LDA.

ОНА ПОЗВОЛЯЕТ:

1. Наглядно представить взаимосвязи между выявленными темами.
2. Определить основные тематические кластеры
3. Изучить расположение тем: (темы, расположенные близко друг к другу, формируют тематические кластеры).
4. Идентифицировать основные группы (группы близко расположенных тем, которые могут представлять собой более широкие тематические области)
5. Исследовать конкретные темы и их состав
6. Провести анализ семантических связей
7. Изучить дистанции: удаленность тем друг от друга указывает на их семантическое различие.
8. Найти мосты
9. Для литературного анализа определить, какие темы характерны для определенных циклов или персонажей или найти неочвидные связи

ПРАКТИЧЕСКАЯ И ТЕОРЕТИЧЕСКАЯ ЗНАЧИМОСТЬ

Данный код представляет собой полноценный исследовательский инструмент для анализа литературных произведений, сочетающий современные методы NLP и машинного обучения с классическими подходами к анализу текстов.