Анализ лексического разнообразия текста

на материале романов Терри Пратчетта

Корбан Андрей

01.03.2025

Цель проекта

1. Отработать на практике методы предобработки текстов и кластеризации токенов

2. Рассчитать метрики лексического разнообразия в коллекции текстов

3. Использовать полученные метрики для ранжирования единиц и выводов

Немного о творчестве писателя

Терри Пратчетт (1948-2015)
считается одним из самых читаемых авторов Великобритании 1990-х годов. Всего им написано более **70 книг**, которые переведены на **37 языков** мира. Общий тираж его книг на разных языках превышает **80 млн. экз.**

15 книг Пратчетта вошли в список **«200 лучших книг по версии BBC»**, составленный в 2003 году по результатам опроса, в котором приняли участие около **1 млн** человек.

ВЫВОД:
Широкая аудитория носителей языка высоко оценила романы.
Там где ирония, там есть широкий выбор материала для изучения языка.

Цикл «Плоский мир»
(1983—2015, насчитывающий 41 роман, 5 рассказов, 4 карты и атласа, 10 справочников и одну поваренную книгу) — это сатира, начинавшаяся как пародия на фэнтези и превратившаяся в совершенно независимое от жанра произведение. Тут можно найти шутки практически на любую тему от истории пирамид Египта до Голливуда и классической философии.

Выбраны книги из цикла Discworld (City Watch)

Причина выбора именно этого набора книг - «стабильность».
Серия большая, одни и те же персонажи, единый стиль и атмосфера и, как следствие, более наглядный анализ изменений в стиле писателя.

Используемые инструменты

Среда разработки:
Google Colab

Библиотеки:
- pandas
- numpy
- matplotlib
- nltk
- spacy
- Wordcloud

Доп.инструмент визуализации:
Power BI (MS Office)
Построение мер на языке DAX

Этапы подготовки материала:

1. Импорт библиотек и инициализация списка стоп-слов

2. Предобработка текстов (токенизация, лемматизация)

3. Расчеты, включающие
   - создание частотных словарей по каждому тексту,
   - частиречный анализ,
   - расчет TF-IDF,
   - определение биграмм и кластеризация по POS1-POS2,
   - расчет метрик Likelihood Ratio, PMI, T-score (Student_t),
   - их сравнение,
   - составления рейтинга,
   - построение визуализаций

4.Экспорт расчетов в csv файлы

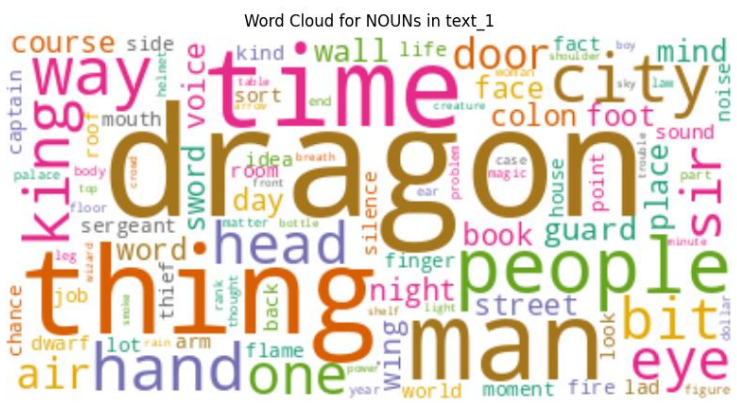
5. Анализ результатов в Power BI

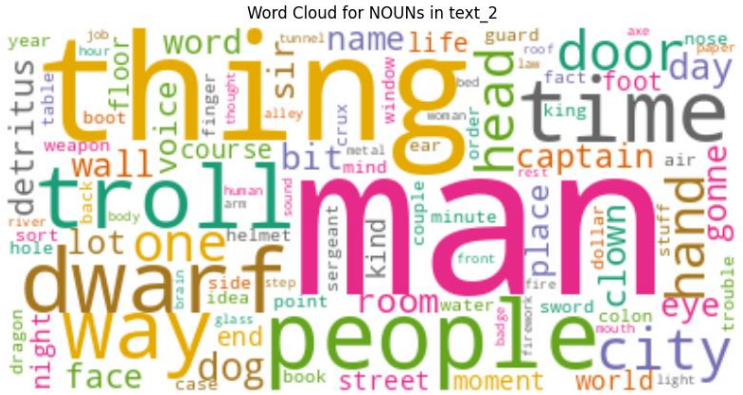
Демонстрация Power BI отчета

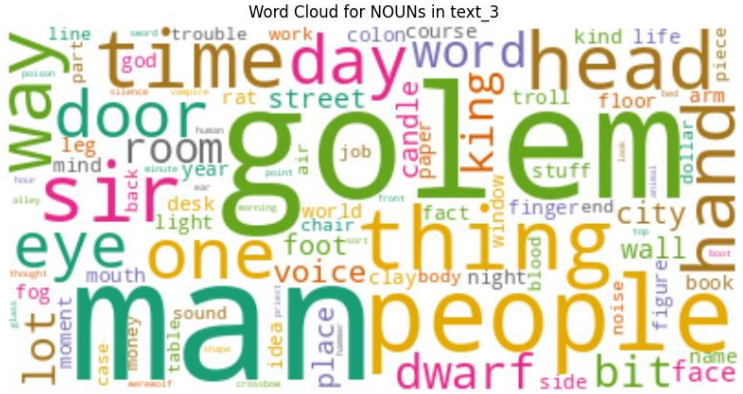
ВИЗУАЛИЗАЦИИ

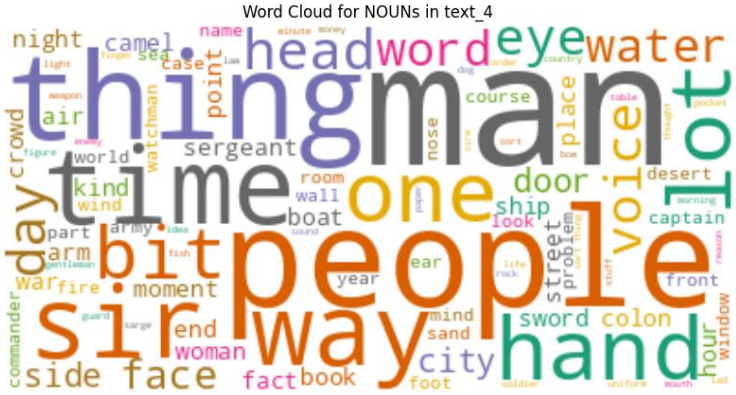
ОБЛАКА СЛОВ И ГРАФИКИ

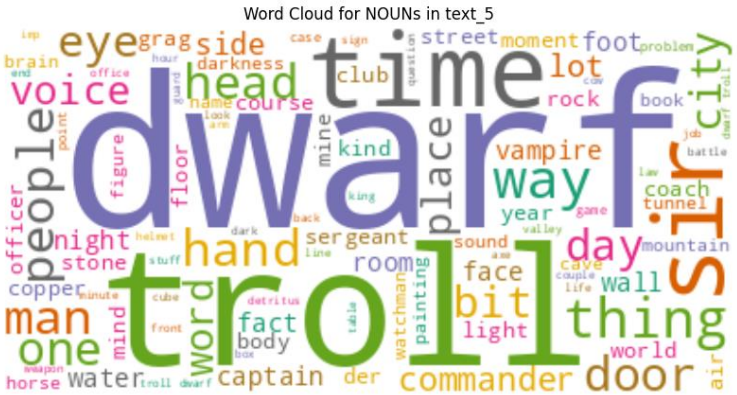
Существительные









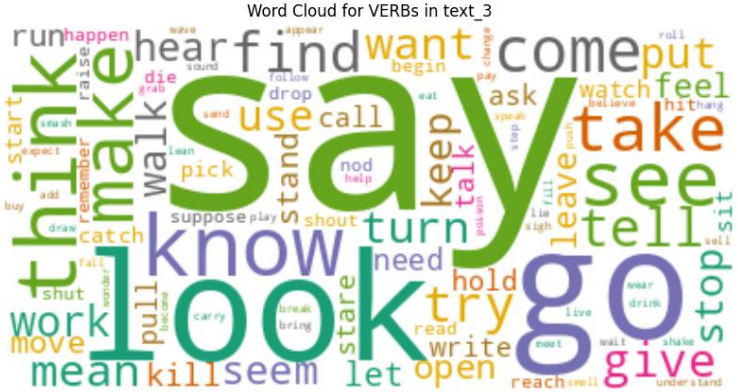


Глаголы

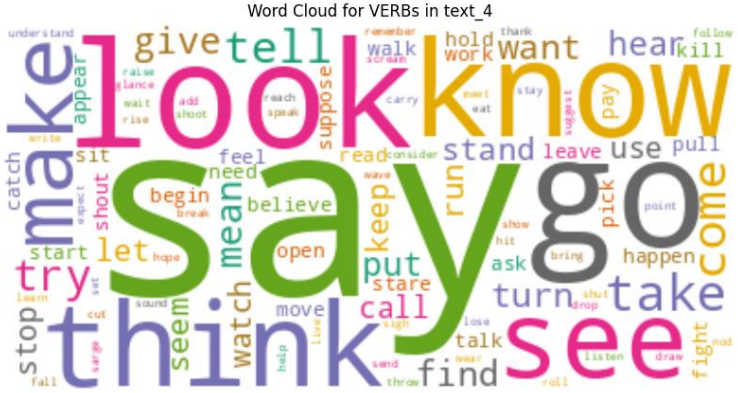
Word Cloud for VERBs in text_1

Word Cloud for VERBs in text_2

Word Cloud for VERBs in text_3



Word Cloud for VERBs in text_4



Word Cloud for VERBs in text_5

Прилагательные

Word Cloud for ADJs in text_1

Word Cloud for ADJs in text_2

Word Cloud for ADJs in text_3

Word Cloud for ADJs in text_4

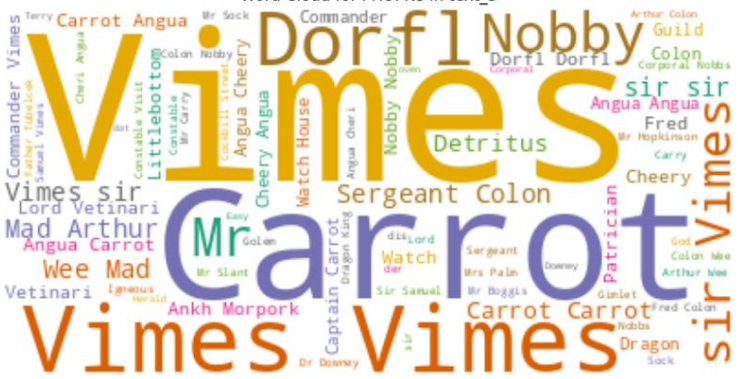
Word Cloud for ADJs in text_5

Имена собственные

Word Cloud for PROPNs in text_1

Word Cloud for PROPNs in text_2

Word Cloud for PROPNs in text_3



Word Cloud for PROPNs in text_4

Word Cloud for PROPNs in text_5

Top 30 - слов по текстам

ВЫВОД:
Книги максимально наполнены диалогами персонажей (споры, простая грубоватя речь стражников)