

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

A word cloud visualization of words from the Harry Potter series. The words are arranged in a dense, overlapping manner, with colors ranging from dark purple to bright yellow. The most prominent words, shown in larger fonts, include "time", "dragon", "man", "city", "thing", "people", "eye", "bit", "sir", "king", "voice", "head", "night", "wing", "hand", "one", "wall", "life", "door", "fact", "day", "mind", "noise", "mouth", "idea", "room", "day", "silence", "finger", "book", "guard", "place", "minute", "street", "look", "world", "moment", "fire", "lad", "figure", "year", "flame", "arm", "lot", "rain", "dwarf", "chance", "job", "stella", "rank", "thought", "back", "shelf", "light", "power", "creature", "face", "sky", "lake", "colon", "foot", "house", "point", "trouble", "part", "case", "magic", "palace", "body", "leg", "sergeant", "word", "throat", "breath", "end", "sort", "kind", "captain", "side", "course", "shoulder", "fact", "day", "mind", "noise". Smaller words like "sword", "mat", "bottle", "ear", "problem", "case", "magic", "house", "point", "trouble", "part", "case", "magic", "house", "point", "trouble", "part", "case", "magic", "house", "point", "trouble", "part" are also visible. The overall composition is a vibrant collage of terms associated with the magical world of Harry Potter.

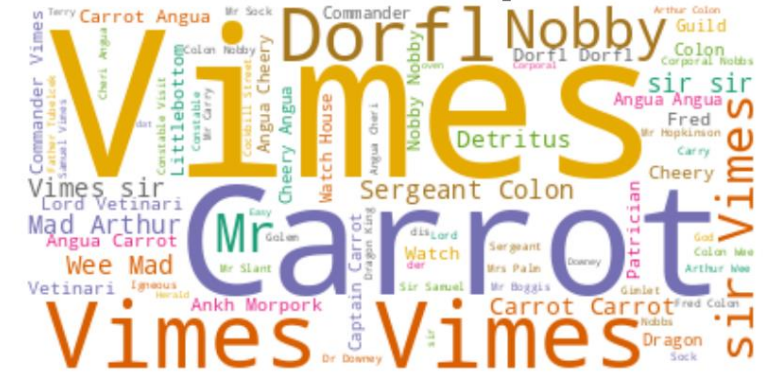
A dense word cloud featuring various terms associated with the Harry Potter universe. The most prominent words are "thing", "man", "people", "city", "time", "dwarf", "way", "troll", "dw", "word", "life", "guard", "door", "day", "voice", "course", "bit", "metal", "mind", "head", "captain", "air", "place", "minute", "front", "fire", "sword", "color", "eye", "moment", "world", "light", "book", "street", "dog", "case", "face", "night", "dragon", "hole", "sort", "lot", "one", "side", "idea", "step", "point", "glass", "end", "room", "water", "badger", "fairwork", "you", "th", "hand", "stuff", "gonne", "arm", "helmet", "couple", "kind", "sergeant", "human", "sound", "wall", "back", "body", "weapon", "detritus", "year", "job", "hour", "flood", "table", "sir", "thought", "alley", "window", "woman", "order", "king", "fact", "foot", "nose", "axe", "roof", "lase", "tunnel", "name". The words are arranged in a chaotic, overlapping manner with varying font sizes and colors, creating a vibrant and textured visual representation of the source material's vocabulary.

[illegible]

[illegible][illegible][illegible]

[illegible]

Word Cloud for PROPNS in text_3



Word Cloud for PROPNS in text_5

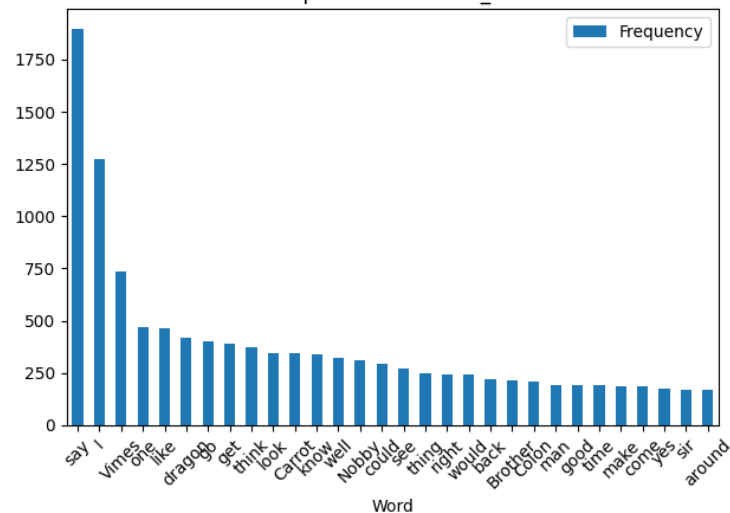


Top 30 - слов по текстам

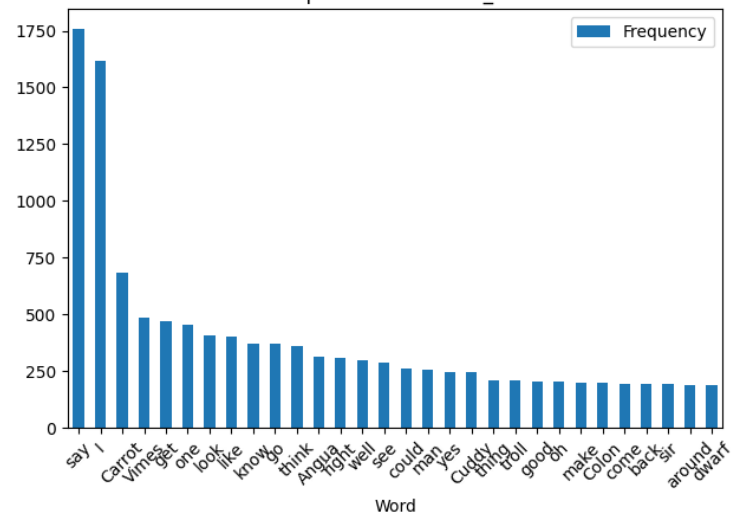
ВЫВОД:

Книги максимально наполнены диалогами персонажей (споры, простая грубоватая речь стражников)

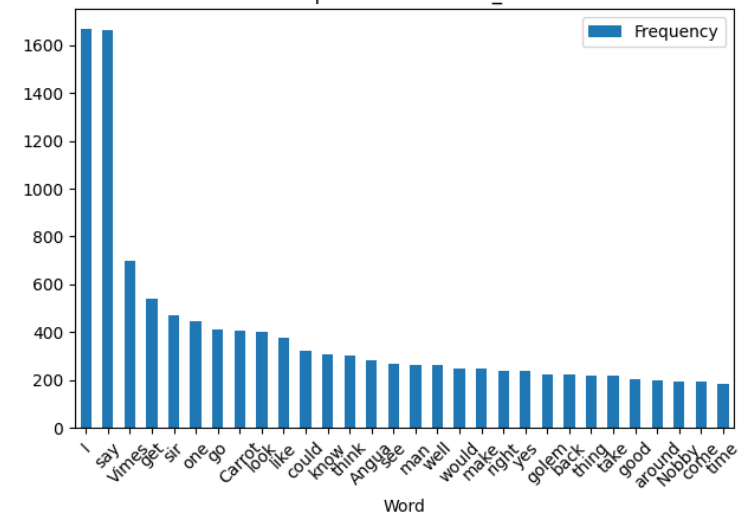
Top 30 words in text_1



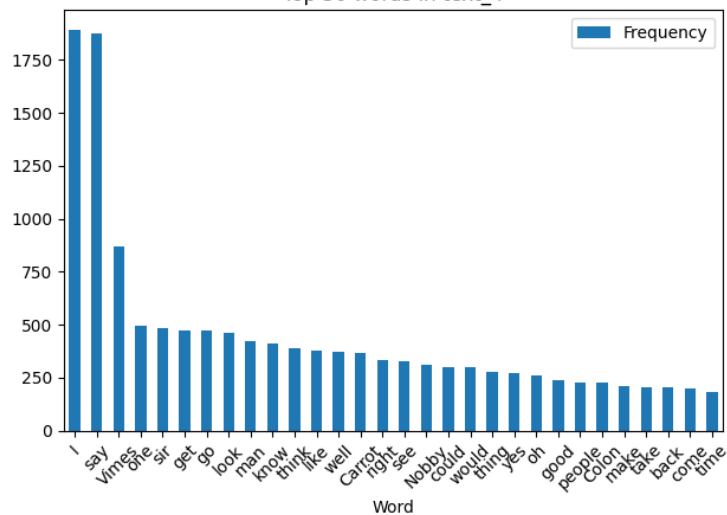
Top 30 words in text_2



Top 30 words in text_3



Top 30 words in text_4



Top 30 words in text_5

