Отчет

Ли Сынгжу 325

**Номер и формулировка задачи**

Задача №5

|  |
| --- |
| Написать программу, позволяющую проверять для текста выполнение закона Ципфа-Мальдеброта. Программа должна:   * осуществлять морфологический анализ словоформ текста; * подсчитывать частоты и ранги различных словоформ и лемм; * выводить списки самых частотных словоформ и лемм; * строить графики (реальный и ожидаемый) зависимости относительной частоты от ранга словоформы и леммы. * определение стиля/жанра текста |

|  |  |  |
| --- | --- | --- |
| **Модули** | nltk |  |
| pymorphy2 |  |
| string |  |
| collection | Counter(list) |
| operator |  |
| matplotlib.pyplot | Чтобы рисовать график |
| tqdm | для распознания процесс работы, сколько осталось (не объязательно, прикладываю версию без этого модуля |



**Морфологический анализ текста**

**Выбранный текст**:

Сергей Мельник - Попаданец. Гексалогия (СИ)

100 строк из введения

<https://libking.ru/books/sf-/sf-fantasy/570101-sergey-melnik-popadanets-geksalogiya-si.html>

**Результат анализа**:

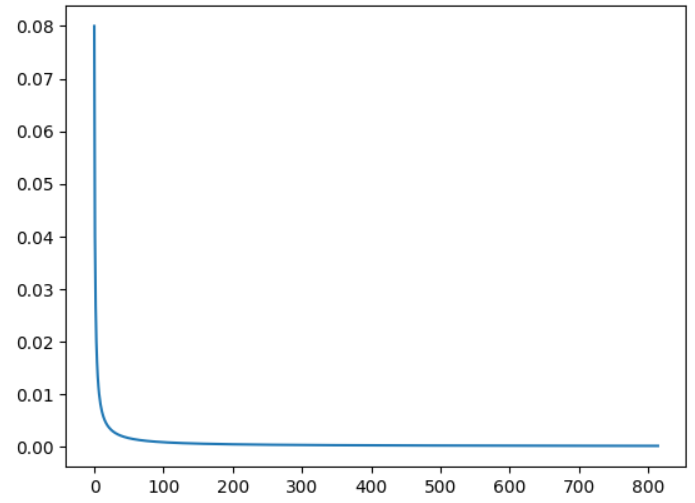
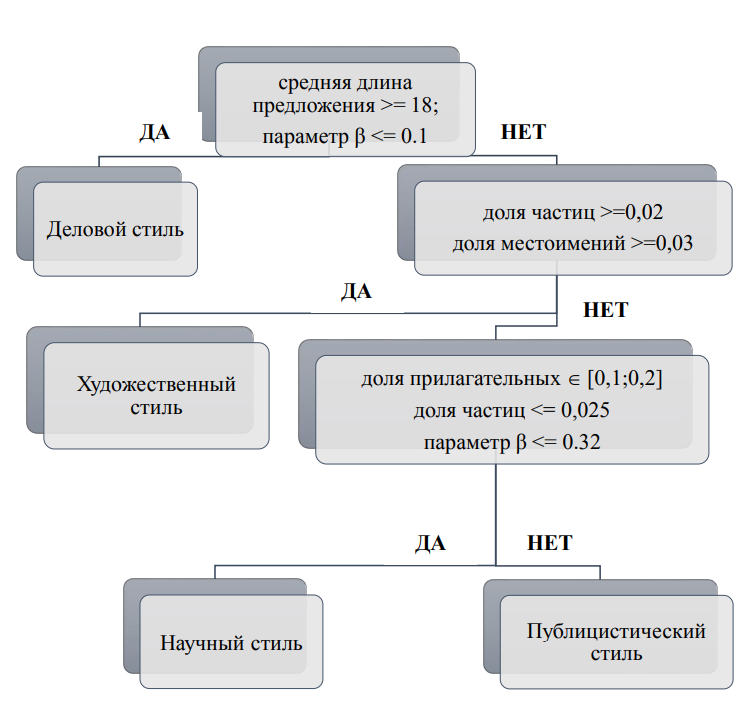


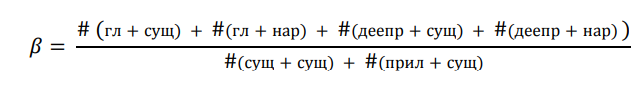
График распределения частот

|  |  |
| --- | --- |
| The highest 9 ranks and frequencies | |
| лес | 0.08 |
| лагерь | 0.04 |
| палатка | 0.027 |
| дождь | 0.02 |
| лялька | 0.016 |
| словно | 0.013 |
| труд | 0.0114 |
| рюкзак | 0.01 |
| собака | 0.008 |

**Задача о распознании жанра текста**:

**Идея решения:**





Дубовик Анна Романовна

Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам

На всякие случая прекрипляю сам файл

**Результат анализа**:

|  |  |
| --- | --- |
| Коэффициенты бета значении | |
| средняя длина предложения | 15.405 |
| доля частиц | 0.0356 |
| доля местоимении | 0.0221 |
| доля прилагательных | 0.2491 |
|  | |
| The type of text | Публицистический стиль |

