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 |  |

**Выбранный текст**:

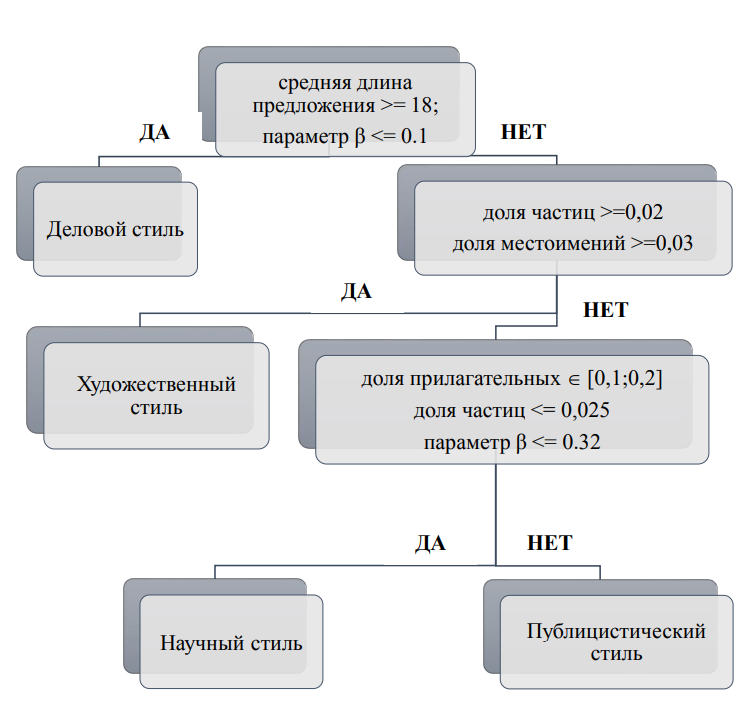
Сергей Мельник - Попаданец. Гексалогия (СИ)

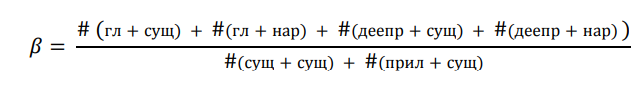
100 строк из введения

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

**Идея решения**

**Задача о распознании жанра текста**





Дубовик Анна Романовна

Автоматическое определение стилистической принадлежности текстов по их статистическим параметрам

На всякие случая прекрипляю сам файл

**Задача подсчёта Закона Ципфа**

|  |
| --- |
|  |

Константа C=0.08 по НКРЯ для теоретического Fri

Константу C для фактического Fri определил как

C=

В морфологическом анализе для удаления стоп-слов завел новый файл stop\_words.txt т.к. не хватило слов в библиотеке nltk.corpus

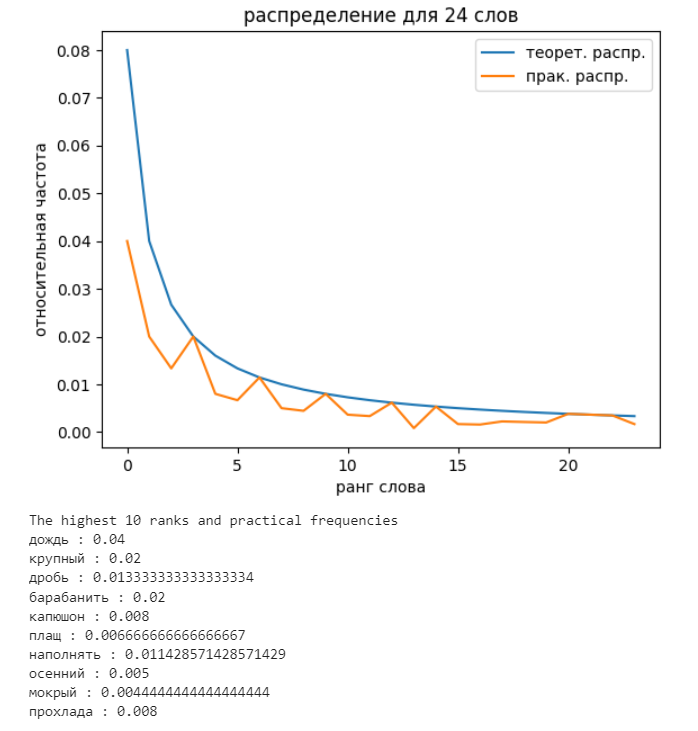
|  |
| --- |
| stop\_words\_file=open("stop\_words.txt",'r',encoding="utf-8")  stop\_words\_list=stop\_words\_file.read()  stop\_words\_file.close()  stop\_words=stop\_words\_list.split('\n') |

Также добавил некоторые пунктуации

|  |
| --- |
| special\_words=string.punctuation  special\_words+='»'+'«'+'–'+'”'+'“'+'—' |

**Результат анализа**

**Морфологический анализ для начальных 22 слов**



**Определение жанра текста**

|  |  |
| --- | --- |
| Коэффициенты бета значении | |
| Beta | 0.44396551724137934 |
| средняя длина предложения | 15.404580152671755 |
| доля частиц | 0.0098159509202454 |
| доля местоимении | 0.014723926380368098 |
| доля прилагательных | 0.18650306748466258 |
|  | |
| The type of text | Публицистический стиль |

|  |
| --- |
|  |

Получается не очень корректно, если были бы большее дата сет, то могло бы использовать метод машинного обучения, но, к сожалению, такой дата сет не найден.

Хотя такой статический метод очень развлекательно. В том числе, корректность распознания жанра пропорциональна его объему. Чем больше слов в тексте, тем лучше распознает.

**Вывод**

Из всех задач получил предвиденные результаты.

Хотя в большом объеме разница графика относительной частоты стало больше, но все-таки ее отклонение сходилось к одному. От того что график становится длиннее, в большом объеме смотреть разницу относительной частоты между теоретической и фактической конечно же, затруднительно.

Самый интересный факт, это то, что анализируемость двух задач обратно пропорционально между собою. Из начально рассчитывал на то что они будут пропорциональными.