АНАЛИЗ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ ПАКЕТОВ PYMORPHY2 И NLTK-RAKE

Выполнили Гайдук Юлия и Панкратова Анна

Работа выполняется на материале произведения Е. И. Рерих "Мир Огненный"

Объем текста 2 Мб

Подготовка

- Предобработка текста: удаление пунктуации, лемматизация с помощью пакета pymorphy2
- Загрузка стоп-слов для русского языка (предлоги, союзы, служебные части речи и др.)

Подсчет частот

- Создание функций, возвращающих списки абсолютных и относительных частот
- Наиболее часто встречающиеся леммы:
-

Лемма 'огненный' встречается в тексте 2168 раз.
Лемма 'мир' встречается в тексте 1705 раз.
Лемма 'нужно' встречается в тексте 1291 раз.
Лемма 'можно' встречается в тексте 1163 раз.
Лемма 'человек' встречается в тексте 1144 раз.
Лемма 'дух' встречается в тексте 980 раз.
Лемма 'такой' встречается в тексте 966 раз.
Лемма 'огонь' встречается в тексте 883 раз.
Лемма 'сердце' встречается в тексте 881 раз.
Лемма 'который' встречается в тексте 881 раз.
Лемма 'каждый' встречается в тексте 826 раз.
Лемма 'энергия' встречается в тексте 813 раз.
Лемма 'когда' встречается в тексте 809 раз.
Лемма 'тонкий' встречается в тексте 747 раз.
Лемма 'самый' встречается в тексте 710 раз.
Лемма 'сознание' встречается в тексте 691 раз.
Лемма 'высокий' встречается в тексте 664 раз.
Лемма 'свой' встречается в тексте 643 раз.
Лемма 'мысль' встречается в тексте 634 раз.
Лемма 'явление' встречается в тексте 614 раз.

Можно предположить, что раз произведение озаглавлено "Мир Огненный", то данному вопросу будет посвящена значительная часть содержания, что и отражается в частотном списке лемм: среди наиболее часто встречающихся лемм мы видим такие элементы как "огненный", "мир", "человек", "дух". Также мы видим наверху частотного списка такие предикативы как "можно" и "нужно", что указывает на обилие в тексте фраз с модальным значением долженствования, необходимости, возможности. Можно предположить, что эти слова являются сказуемыми безличных предложений.

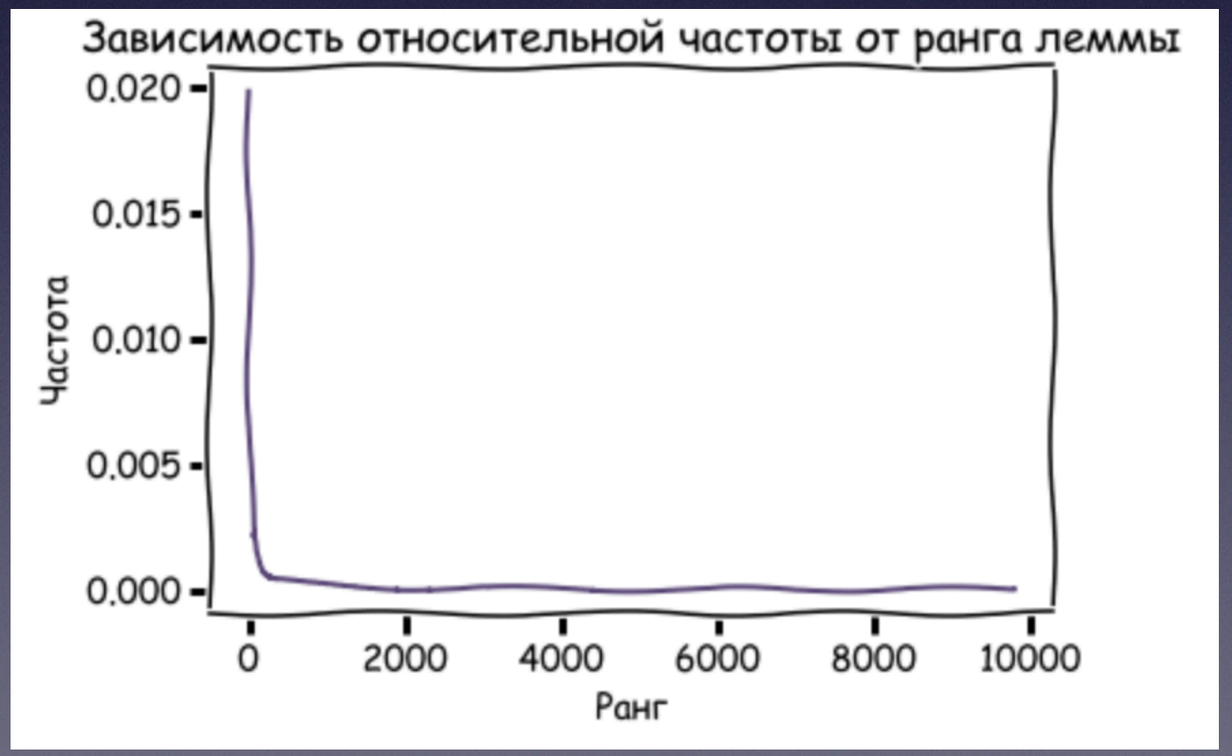
Подсчет рангов

Ранг — порядковый номер слова в списке,
упорядоченном по убыванию частот

Лемма 'огненный' имеет ранг 1.
Лемма 'мир' имеет ранг 2.
Лемма 'нужно' имеет ранг 3.
Лемма 'можно' имеет ранг 4.
Лемма 'человек' имеет ранг 5.
Лемма 'дух' имеет ранг 6.
Лемма 'такой' имеет ранг 7.
Лемма 'огонь' имеет ранг 8.
Лемма 'сердце' имеет ранг 9.
Лемма 'который' имеет ранг 10.

Зависимость частоты от ранга

Положим частоты на ось y, а ранги — на ось x. Построим график и посмотрим, что получилось. Мы видим, что частота обратно пропорциональна рангу. Чем выше частота, тем меньше ранг, и наоборот. График иллюстрирует эмпирический закон Ципфа.



Применение алгоритма RAKE

Алгоритм был применен на лемматизированном и на нелемматизированном текстах, сравнивались словосочетания длины в 2 и 3 слова. Мы видим, что в составе наиболее часто встречающихся словосочетаний лежат леммы, которые pymorphy2 определил как наиболее частотные.

Наиболее часто встречающиеся словосочетания длины 2 по лемматизированному тексту:

огненный дух
дух человек
только человек
огненный сознание
сердце огненный
огненный сердце
сознание человек
огонь дух
огненный мир
мир огненный

Наиболее часто встречающиеся словосочетания длины 2 по нелемматизированному тексту:

может сердце
будет нужно
только могут
жизни может
сознание может
нужно сердце
можно будет
огонь может
огненного нужно
только тонкого

Выводы

- Анализ списка частотности лемм показал, что в выбранном произведении в топе списка частот оказались слова, тесно связанные с тематикой произведения (например *огненный*, *человек*, *дух*. Можно предположить, что в текстах узкой направленности или специализированных текстах в списке частотности лемм будут присутствовать слова, связанные с тематикой текста.

- При построении графиков зависимости частоты от ранга выяснилось, что чем выше частота, тем меньше ранг. Частота обратно пропорциональна рангу и для словоформ, и для лемм, что подтверждает закон Ципфа.

- Анализ наиболее часто встречающихся словосочетаний показал, что в их основе лежат леммы, выделенные как наиболее частотные на раннем этапе.

Благодарим за внимание!