Отчёт по заданию
Лебедев Андрей, группа 424

Цель работы:

- Создать текстовую коллекцию из трёх статей Википедии.
- Применить два морфологических анализатора (PyMorphy и нейросетевой анализатор).
- Создать частотные списки лемм из всех статей.
- Выполнить сравнительный анализ анализаторов по скорости, качеству лемматизации и другим критериям.

Исходные данные:

Были взяты три статьи Википедии:
1. Гватемальско-испанские отношения
2. Бобровый
3. Булгакова Варвара Михайловна

Все тексты объединены для дальнейшего анализа.

Инструменты:

- PyMorphy2
- SpaCy (нейросетевая модель)

Частотный анализ лемм

1. Результаты PyMorphy2:

Частотный список лемм, полученных с использованием PyMorphy2:

Лемма	Частота
в	139
и	97
на	69
остров	56
быть	45
год	39
с	21
гватемала	18
по	18
который	18

Время выполнения: 1.66 секунд.

2. Результаты SpaCy:

Частотный список лемм, полученных с использованием SpaCy:

Лемма	Частота
,	231
.	171
в	141
и	97
на	69
]	67
\xa0	66
остров	64
год	56
(	39

Время выполнения: 3.30 секунд.

Анализ и сравнение

1. Скорость выполнения:
PyMorphy2 оказался быстрее, обработав текст за 1.66 секунд, в то время как SpaCy потратил 3.30 секунд на выполнение задачи.

2. Качество лемматизации:
PyMorphy2 не включает в частотные списки знаки препинания и специальные символы, что делает его более точным для чисто текстовой обработки. SpaCy, напротив, включает в список лемм символы пунктуации (например, запятые, точки, круглые скобки) и даже специальные символы (например, \xa0 — символ неразрывного пробела), что может затруднять дальнейший анализ текста.

3. Частотные списки и семантика:
Несмотря на разницу в обработке пунктуации и спецсимволов, оба анализатора согласны по основным леммам. В обеих системах предлоги, союзы и семантически значимые слова (например, остров, год) занимают верхние позиции.

4. Статистический анализ:
Корреляция Пирсона между частотами лемм двух анализаторов составляет 0.98, что указывает на высокую степень схожести в результатах. Оба анализатора выдают частотные списки лемм, близкие друг к другу.
t-тест показал, что различия в средних значениях частот лемм статистически незначимы ($p = 0.6754$), что подтверждает, что результаты двух анализаторов можно считать схожими.

Выводы:

1. PyMorphy2 работает быстрее и точнее обрабатывает текст, не включая знаки препинания и специальные символы.
2. SpaCy, хотя и медленнее, включает пунктацию, что может быть полезно для более специфичных задач NLP.
3. Оба инструмента показывают очень близкие результаты в частотном

анализе (корреляция 0.98), что говорит о их схожей эффективности в базовых задачах морфологического анализа.