МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В.Ломоносова

Факультет вычислительной математики и кибернетики

Анализ неструктурированных данных.

Тема №9 Законы текста

Отчет о выполнении большого домашнего задания

студента 325 учебной группы факультета ВМК МГУ
Жигалова Никиты Сергеевича

гор. Москва
2024 г.

1 Постановка задачи

Была взята тема 9 **"Законы текста"** :

Написать программу, позволяющую проверять для заданного текста выполнение закона Ципфа-Мальдеброта и/или Хипса. После выполнения морфологического анализа словоформ текста программа должна:

- подсчитывать частоты и ранги различных словоформ/лемм;
- подсчитывать частоты и ранги различных словоформ/лемм;
- выводить по запросу 5-20 самых частотных словоформ/лемм;

строить график зависимости относительной частоты от ранга словоформы-/леммы и/или график закона Хипса. Для проверки закона Хипса следует рассмотреть в обрабатываемом тексте текстовые фрагменты последовательно увеличивающегося размера (в токенах/словоупотреблениях) и построить график зависимости числа уникальных лемм фрагмента от его размера.

Отчет: программа с комментариями, указание обработанных текстов, подсчитанная статистика, комментарии к способу ее подсчета, графики и выводы.

Прикладные задачи: оценка естественности текста, определение стиля/-жанра текста.

2 Реализация

Данная программа решала несколько задач:

1. Предварительная обработка текста: замена ё на е, удаление пунктуации, приведение текста к нижнему регистру и токенизация
2. Полный морфологический анализ слов текста с выведением результатов в файл Excel
3. Подсчет ранга, абсолютной и относительной частоты употребления словоформ и лемм с выводом в файл и на экран
4. Работа с графиками и вывод графиков кол-ва слов по частям речи, а также закона Хипса для лемм и для словоформ, и график появления новых словоформ в тексте

Работа с программой осуществляется через терминал, ответом на запросы программы является y/n (вне зависимости от регистра и пробелов по краям) или же число, где об этом сказано прямо. В случае некорректного пользовательского ввода программа аварийно прекращет свое выполнение!

Библиотеки:

- `pymorphy3` - морфологический анализ слов
- `matplotlib.pyplot` - построение графиков частоты встречи слов разных частей речи, а также для отрисовки графика закона Хипса
- `numpy` - для задания осей в графиках
- `pandas` - построение таблицы морфологического анализа и вывод в .xlsx файл
- `openpyxl` - библиотека для работы с .xlsx файлами, нужна для вывода морфологического анализа в таблицу
- `sys` - более опрятное завершение программы при ошибках ввода
- `re` - используем регулярные выражения для проверки форматов файлов
- `nltk.tokenize` - использовался для токенизации текста
- `tqdm` - красивые прогресс-бары при длительном выполнении некоторых функций (например, морфологический анализ)

Программа состоит из трех модулей:

1. `morph.py` - все функции необходимые для морфологического анализа таких частей речи как: существительные, глаголы, прилагательные, местоимения, причастия, деепричастия, числительные и наречия (остальные части речи не анализируются)
2. `text_job.py` - функции, необходимые для предобработки текста, подсчета статистики по тексту и вывода графиков
3. `main.py` - основная работа с консолью, вывод в файлы, импортирует две предыдущие

3 Проделанная работа

Изначально текст предобрабатывается с помощью функций в `text.py`, функция `form_stop_words` формирует список стоп-слов, расширяя его большинством местоимений и служебных частей речи, `filter_punct` - заменяет ё на е, а также удаляет из текста всю пунктуацию

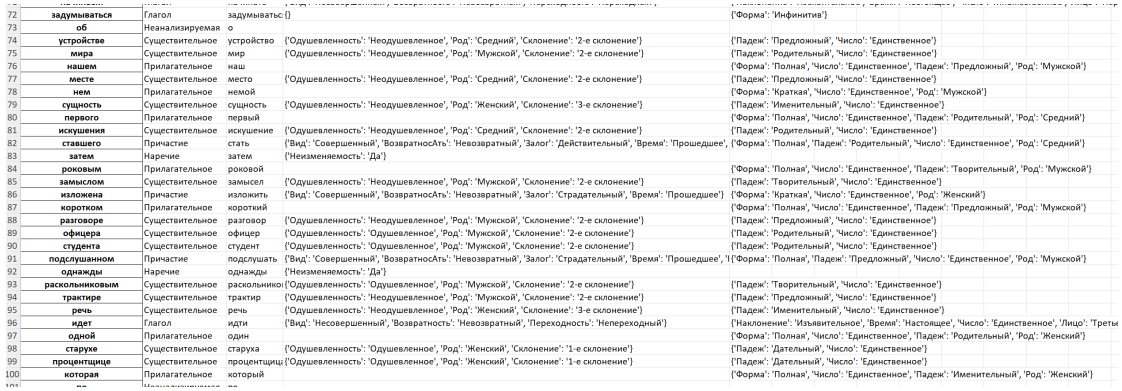
После выполняется подсчет лемм и словоформ, встречающихся в тексте, с помощью функций `ranged_wordforms` и `ranged_lemmas` соответственно

После чего с помощью `relative_freq` вычисляется относительная частота словоформ и лемм

Также в программе реализован полный морфологический анализ текста, за это отвечает модуль `morph.py` За основу была взята работа с библиотекой pymorphy3 и тегами, которые анализатор присваивает каждому слову. Морфологический анализ выполняется в функции `morphological_analysis`, в зависимости от поступившей части речи, определяемой параметром tag.POS. Сразу скажу, что проблема омонимии в моей задаче не решена, рассматривается наиболее вероятное словоупотребление от парсера (с индексом 0). Далее в зависимости от части речи выполняется морфологический разбор. Рассматриваются имена существительные, глаголы, инфинитивы, полные/краткие/-сравнительные прилагательные, числительные, наречия, местоимения, причастия и деепричастия. В каждой функции обработки части речи создается словарь result, внутри которого в зависимости от морфологического разбора находится 3 или 4 ключа. У всех частей речи присутствуют ключи «Часть речи» и «Начальная форма», у большинства – присутствуют «Постоянные признаки» и «Непостоянные признаки», оформленные в виде словарей. Дальше по каждому возможному признаку формируются эти два словаря. Так, у причастия словарь постоянных признаков состоит из вида, возвратности, залога, времени и переходности, а словарь непостоянных признаков – из формы, падежа, числа и рода. Все формируется в зависимости от наличия того или иного тэга в обработанной парсером версии слова. В основном все признаки отбирались с помощью тегов pymorphy, но, например, склонение у существительных и возвратность у глаголов искалось вручную без библиотеки. Помимо перечисленных выше частей речи и их обработчиков присутствует обработчик неанализируемых частей речи (предлогов, союзов, частиц, междометий и др.). Они для анализа не нужны, для прикладной задачи тоже.

После выполнения анализа выводится bar plot самых частых частей речи с их количеством (функция `graph_morph`)

Примерный вид таблицы морфологического анализа:



72	**задумываться**	Глагол	задумываться	{}	{'Форма': 'Инфинитив'}
73	**об**	Неанализируемая	о		
74	**устройстве**	Существительное	устройство	{'Одушевленность': 'Неодушевленное', 'Род': 'Средний', 'Склонение': '2-е склонение'}	{'Падеж': 'Предложный', 'Число': 'Единственное'}
75	**мира**	Существительное	мир	{'Одушевленность': 'Неодушевленное', 'Род': 'Мужской', 'Склонение': '2-е склонение'}	{'Падеж': 'Родительный', 'Число': 'Единственное'}
76	**нашем**	Прилагательное	наш		{'Форма': 'Полная', 'Число': 'Единственное', 'Падеж': 'Предложный', 'Род': 'Мужской'}
77	**месте**	Существительное	место	{'Одушевленность': 'Неодушевленное', 'Род': 'Средний', 'Склонение': '2-е склонение'}	{'Падеж': 'Предложный', 'Число': 'Единственное'}
78	**нем**	Прилагательное	немой		{'Форма': 'Краткая', 'Число': 'Единственное', 'Род': 'Мужской'}
79	**сущность**	Существительное	сущность	{'Одушевленность': 'Неодушевленное', 'Род': 'Женский', 'Склонение': '3-е склонение'}	{'Падеж': 'Именительный', 'Число': 'Единственное'}
80	**первого**	Прилагательное	первый		{'Форма': 'Полная', 'Число': 'Единственное', 'Падеж': 'Родительный', 'Род': 'Средний'}
81	**искушения**	Существительное	искушение	{'Одушевленность': 'Неодушевленное', 'Род': 'Средний', 'Склонение': '2-е склонение'}	{'Падеж': 'Родительный', 'Число': 'Единственное'}
82	**ставшего**	Причастие	стать	{'Вид': 'Совершенный', 'ВозвратноСть': 'Невозвратный', 'Залог': 'Действительный', 'Время': 'Прошедшее',	{'Форма': 'Полная', 'Падеж': 'Родительный', 'Число': 'Единственное', 'Род': 'Средний'}
83	**затем**	Наречие	затем	{'Неизменяемость': 'Да'}	
84	**роковым**	Прилагательное	роковой		{'Форма': 'Полная', 'Число': 'Единственное', 'Падеж': 'Творительный', 'Род': 'Мужской'}
85	**замыслом**	Существительное	замысел	{'Одушевленность': 'Неодушевленное', 'Род': 'Мужской', 'Склонение': '2-е склонение'}	{'Падеж': 'Творительный', 'Число': 'Единственное'}
86	**изложена**	Причастие	изложить	{'Вид': 'Совершенный', 'ВозвратноСть': 'Невозвратный', 'Залог': 'Страдательный', 'Время': 'Прошедшее'}	{'Форма': 'Краткая', 'Число': 'Единственное', 'Род': 'Женский'}
87	**коротком**	Прилагательное	короткий		{'Форма': 'Полная', 'Число': 'Единственное', 'Падеж': 'Предложный', 'Род': 'Мужской'}
88	**разговоре**	Существительное	разговор	{'Одушевленность': 'Неодушевленное', 'Род': 'Мужской', 'Склонение': '2-е склонение'}	{'Падеж': 'Предложный', 'Число': 'Единственное'}
89	**офицера**	Существительное	офицер	{'Одушевленность': 'Одушевленное', 'Род': 'Мужской', 'Склонение': '2-е склонение'}	{'Падеж': 'Родительный', 'Число': 'Единственное'}
90	**студента**	Существительное	студент	{'Одушевленность': 'Одушевленное', 'Род': 'Мужской', 'Склонение': '2-е склонение'}	{'Падеж': 'Родительный', 'Число': 'Единственное'}
91	**подслушанном**	Причастие	подслушать	{'Вид': 'Совершенный', 'ВозвратноСть': 'Невозвратный', 'Залог': 'Страдательный', 'Время': 'Прошедшее',	{'Форма': 'Полная', 'Падеж': 'Предложный', 'Число': 'Единственное', 'Род': 'Мужской'}
92	**однажды**	Наречие	однажды	{'Неизменяемость': 'Да'}	
93	**раскольниковым**	Существительное	раскольников	{'Одушевленность': 'Одушевленное', 'Род': 'Мужской', 'Склонение': '2-е склонение'}	{'Падеж': 'Творительный', 'Число': 'Единственное'}
94	**трактире**	Существительное	трактир	{'Одушевленность': 'Неодушевленное', 'Род': 'Мужской', 'Склонение': '2-е склонение'}	{'Падеж': 'Предложный', 'Число': 'Единственное'}
95	**речь**	Существительное	речь	{'Одушевленность': 'Неодушевленное', 'Род': 'Женский', 'Склонение': '3-е склонение'}	{'Падеж': 'Именительный', 'Число': 'Единственное'}
96	**идет**	Глагол	идти	{'Вид': 'Несовершенный', 'Возвратность': 'Невозвратный', 'Переходность': 'Непереходный'}	{'Наклонение': 'Изъявительное', 'Время': 'Настоящее', 'Число': 'Единственное', 'Лицо': 'Треты
97	**одной**	Прилагательное	один		{'Форма': 'Полная', 'Число': 'Единственное', 'Падеж': 'Родительный', 'Род': 'Женский'}
98	**старухе**	Существительное	старуха	{'Одушевленность': 'Одушевленное', 'Род': 'Женский', 'Склонение': '1-е склонение'}	{'Падеж': 'Дательный', 'Число': 'Единственное'}
99	**процентщице**	Существительное	процентщица	{'Одушевленность': 'Одушевленное', 'Род': 'Женский', 'Склонение': '1-е склонение'}	{'Падеж': 'Дательный', 'Число': 'Единственное'}
100	**которая**	Прилагательное	который		{'Форма': 'Полная', 'Число': 'Единственное', 'Падеж': 'Именительный', 'Род': 'Женский'}

Построение графика возникновения новых лемм/словоформ

Далее осуществлялся подсчет новых лемм/словоформ в последовательно увеличивающихся частях текста. По умолчанию я брал разбиение текста на 100 частей, благодаря чему построение графика не занимает слишком много времени и график получается достаточно ровным. Но если захотим решать такие прикладные задачи как обнаружение самокопирования (копипаста) или суммаризации текста (можно также обнаруживать подведение итогов в статьях, потому, что там обычно повторяются те же слова), тогда может потребоваться более мелкое разбиение текста. `plot_heaps_lm` и `plot_heaps_wf` осуществляют вывод графиков

Дополнительные папки

В папке **texts** прикреплены файлы с тестовыми текстами для анализа текста.
В папке **results** прикреплены файлы с результатами анализа текста.
В папке **tests** содержатся тесты с вручную подсчитанными ответами, которые лежат в "Ответы к тестам.txt"+ здесь также лежат тексты для бонусного задания на проверку копипаста
В папку **graphs** сохраняются сгенерированные графики по итогу работы анализатора

Тестовые тексты

1. Л.Н.Толстой "Война и мир texts/war_and_peace.txt
2. Ф.М.Достоевский "Преступление и наказание"(фрагмент) - texts/crime_and_punishment.
3. Конституция РФ - texts/constitution.txt
4. Статья на Хабре - "Как фармить Kaggle texts/habr_kaggle_farm.txt

5. Доклад "Институциональная школа экономики texts/institutionalism.txt
6. Конспект лекций по курсу "Операционные системы ВМК 2006г. - texts/os_mash.txt
7. Описание этого задания - texts/description_big_task.txt
8. Практическое применение: проверка текста на копипасту ("Преступление и наказание"+ снова первая половина "Преступления и наказания"+ первые пять глав "Война и мир") - texts/copypaste_proof.txt

Калибровка кривой закона Хипса

Закон Хипса:

$$n\_unique = \alpha N^{\beta}$$

Где n_unique - число уникальных лемм/словоформ. N - всего лемм/словоформ в тексте (фрагменте текста) α, β - подбираемые константы

Для русских текстов обычно берут $10 < \alpha < 100$ и $0.4 < \beta < 0.6$. Исходя из этого, а также беря за основу график появления новых лемм/словоформ в тексте "Война и мир"были выбраны константы $\alpha = 45, \beta = 0.48$ - для графика лемм и $\alpha = 35, \beta = 0.58$ - для графика словоформ. Таким образом на основе сравнения с текстом "Война и мир"можно анализировать лексическое разнообразие слов и лемм в различных текстах (а значит и стилистику текста), а также проверять тексты на наличие самокопирований (копипаста)

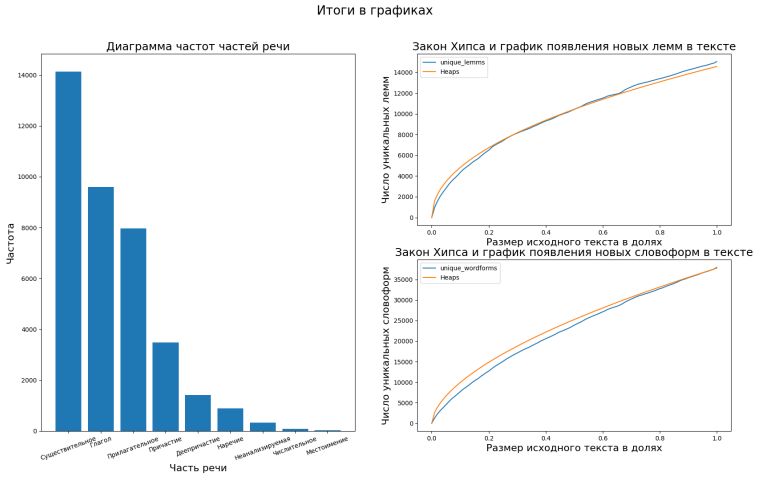
4 Анализ текстов

Художественные тексты

Для текста "Война и мир"были получены результаты:
Средняя длина лемм в тексте: 8.4537

Самые частые 5 лемм:

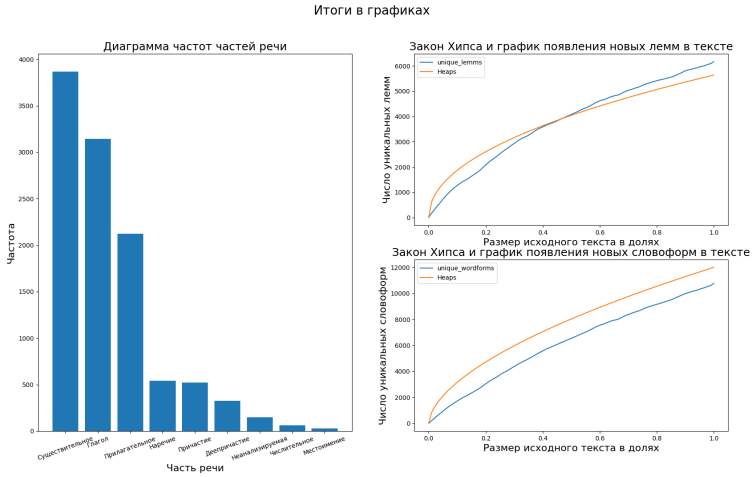
1. КОТОРЫЙ | ранг: 1 | отн. частота (Fr): 0.0140 | абс. частота (Fa): 2392
2. СКАЗАТЬ | ранг: 2 | отн. частота (Fr): 0.0118 | абс. частота (Fa): 2018
3. КНЯЗЬ | ранг: 3 | отн. частота (Fr): 0.0113 | абс. частота (Fa): 1928
4. СВОЙ | ранг: 4 | отн. частота (Fr): 0.0103 | абс. частота (Fa): 1767
5. ГОВОРИТЬ | ранг: 5 | отн. частота (Fr): 0.0088 | абс. частота (Fa): 1507



Для текста "Преступление и наказание"(фрагмент):
Средняя длина лемм в тексте: 8.0366

1. СВОЙ | ранг: 1 | отн. частота (Fr): 0.0070 | абс. частота (Fa): 167
2. ЧЕЛОВЕК | ранг: 2 | отн. частота (Fr): 0.0064 | абс. частота (Fa): 151
3. КОТОРЫЙ | ранг: 3 | отн. частота (Fr): 0.0061 | абс. частота (Fa): 145
4. ОДИН | ранг: 4 | отн. частота (Fr): 0.0058 | абс. частота (Fa): 138
5. СТАТЬ | ранг: 5 | отн. частота (Fr): 0.0054 | абс. частота (Fa): 129

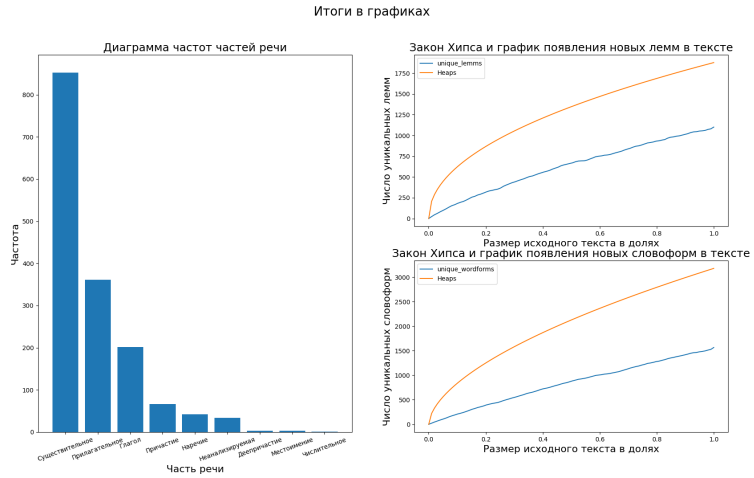
Самые частые 5 лемм:



Доклады/статьи

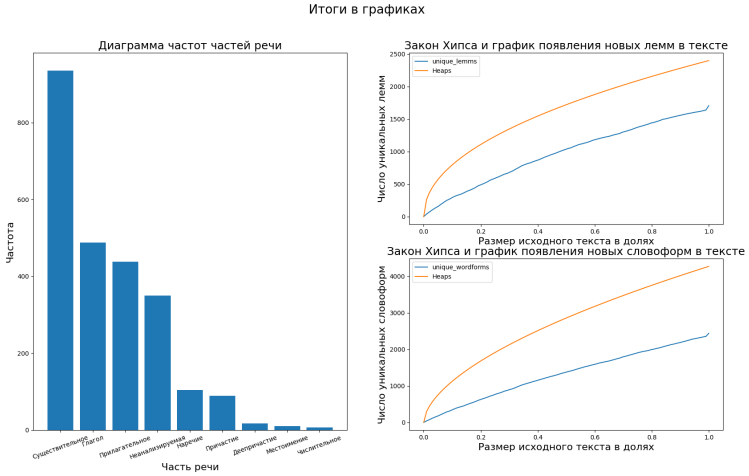
Доклад по институциональной школе экономике
Информация по леммам
Средняя длина лемм в тексте: 8.8937
Самые частые 5 лемм:

1. ЭКОНОМИЧЕСКИЙ | ранг: 1 | отн. частота (Fr): 0.0250 | абс. частота (Fa): 61
2. ЭКОНОМИКА | ранг: 2 | отн. частота (Fr): 0.0143 | абс. частота (Fa): 35
3. ТЕОРИЯ | ранг: 3 | отн. частота (Fr): 0.0119 | абс. частота (Fa): 29
4. ИНСТИТУТ | ранг: 4 | отн. частота (Fr): 0.0115 | абс. частота (Fa): 28
5. КОТОРЫЙ | ранг: 5 | отн. частота (Fr): 0.0107 | абс. частота (Fa): 26



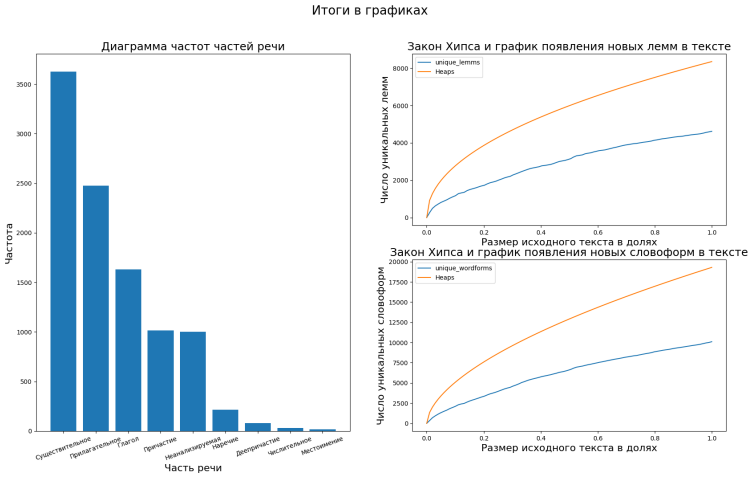
Статья на Хабр "Как фармить Kaggle"
Средняя длина лемм в тексте: 8.0235
Самые частые 5 лемм:

1. СОРЕВНОВАНИЕ | ранг: 1 | отн. частота (Fr): 0.0105 | абс. частота (Fa): 43
2. KAGGLE | ранг: 2 | отн. частота (Fr): 0.0100 | абс. частота (Fa): 41
3. МОДЕЛЬ | ранг: 3 | отн. частота (Fr): 0.0093 | абс. частота (Fa): 38
4. КОТОРЫЙ | ранг: 4 | отн. частота (Fr): 0.0085 | абс. частота (Fa): 35
5. РЕШЕНИЕ | ранг: 5 | отн. частота (Fr): 0.0081 | абс. частота (Fa): 33



Учебная литература
Конспект лекций по операционным системам
Средняя длина лемм в тексте: 8.8601
Самые частые 5 лемм:

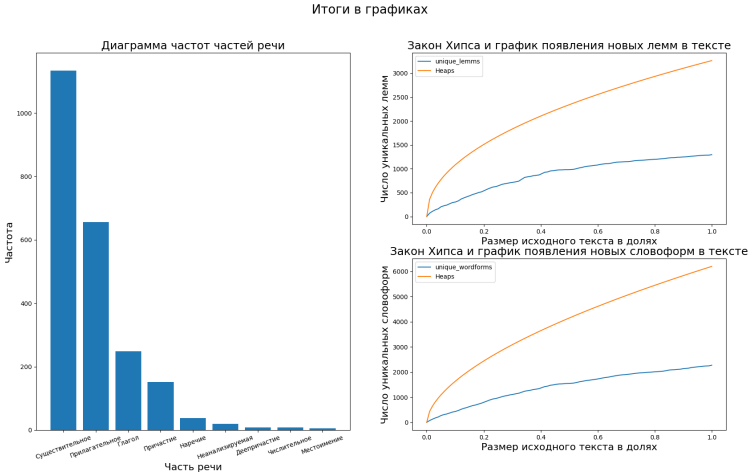
1. СИСТЕМА | ранг: 1 | отн. частота (Fr): 0.0235 | абс. частота (Fa): 1263
2. ПРОЦЕСС | ранг: 2 | отн. частота (Fr): 0.0195 | абс. частота (Fa): 1050
3. КОТОРЫЙ | ранг: 3 | отн. частота (Fr): 0.0135 | абс. частота (Fa): 724
4. ФАЙЛ | ранг: 4 | отн. частота (Fr): 0.0108 | абс. частота (Fa): 579
5. УСТРОЙСТВО | ранг: 5 | отн. частота (Fr): 0.0104 | абс. частота (Fa): 560



Официально-деловые
Конституция РФ
Средняя длина лемм в тексте: 9.2317
Самые частые 5 лемм:

1. ФЕДЕРАЦИЯ | ранг: 1 | отн. частота (Fr): 0.0680 | абс. частота (Fa): 518
2. РОССИЙСКИЙ | ранг: 2 | отн. частота (Fr): 0.0613 | абс. частота (Fa): 467
3. СТАТЬЯ | ранг: 3 | отн. частота (Fr): 0.0248 | абс. частота (Fa): 189
4. ГОСУДАРСТВЕННЫЙ | ранг: 4 | отн. частота (Fr): 0.0240 | абс. частота (Fa): 183

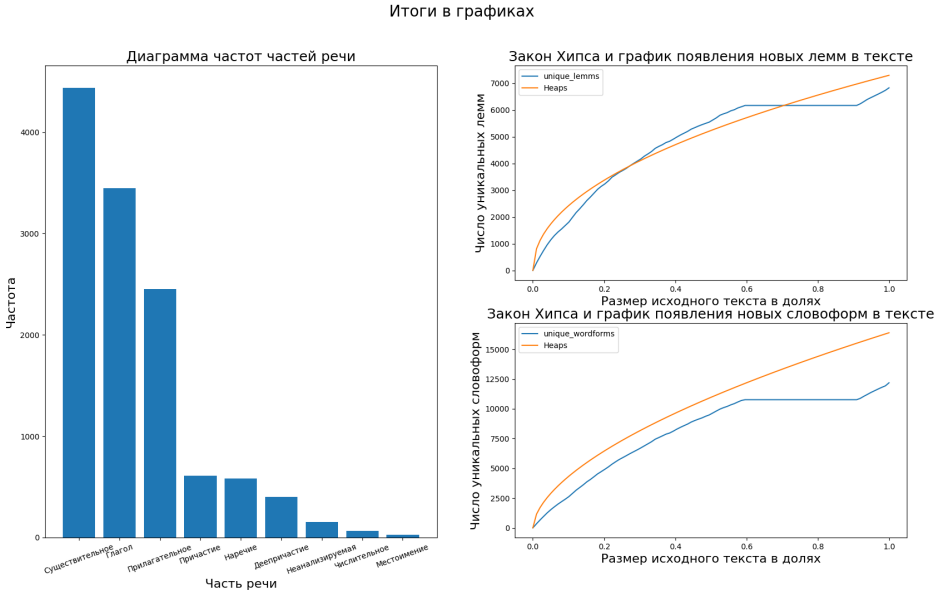
5. ФЕДЕРАЛЬНЫЙ | ранг: 5 | отн. частота (Fr): 0.0220 | абс. частота (Fa): 168



5 Бонусное задание

Решение прикладной задачи: обнаружение самокопирования в тексте (копипаста)

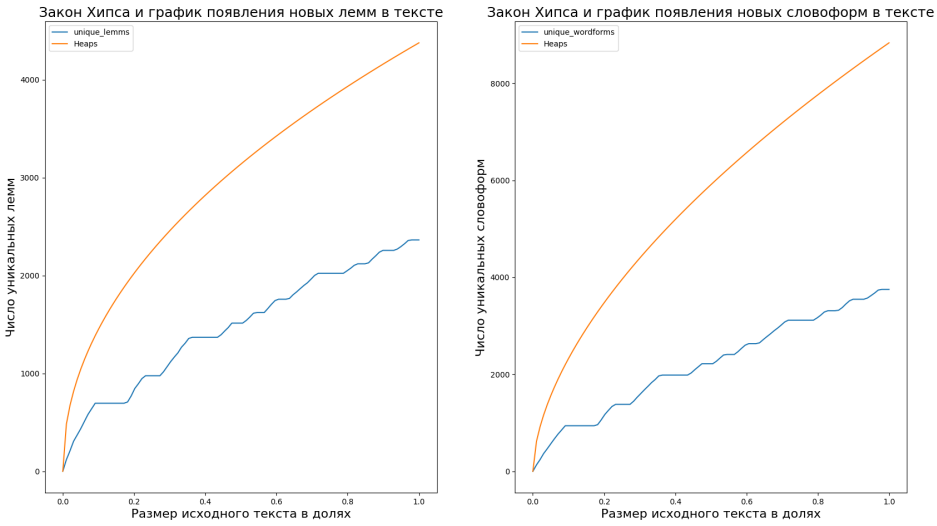
Для задачи был специально сделан тест, составленный из "Преступления и наказания"плюс сразу за ним еще первая половина его же и потом первые 5 глав "Война и мир"
Посмотрим на графики



Легко видеть, что в момент, когда начинается второй текст "Преступления и наказания"новые леммы и словоформы перестают добавлятся и график выглядит, как константная прямая. И когда потом начинается текст "Война и мир"график снова начинает рост, так как новые слова начинают добавляться.

Таким образом можно лекго проверять текста на копипасту, а также если в тексте присутствует суммаризация по вышележащему тексту это тоже будет заметно на графике.

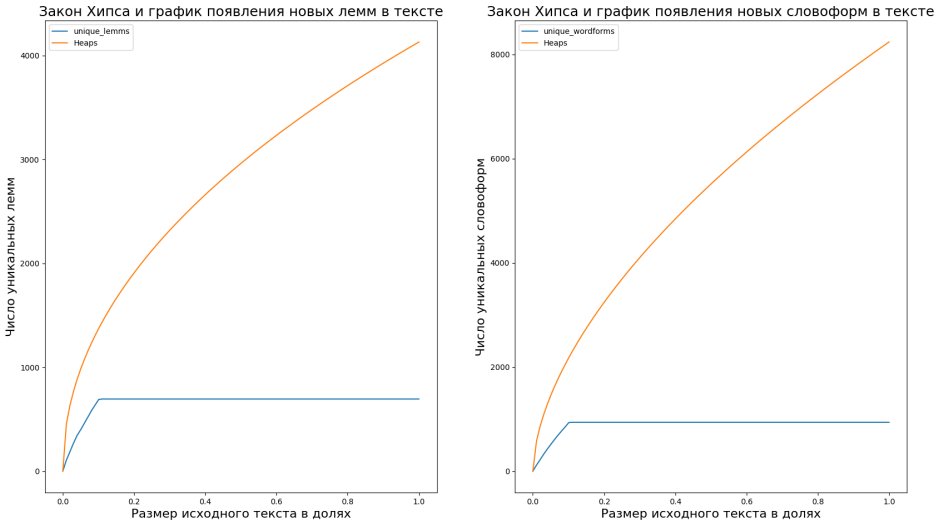
Еще примеры:
$test/cp\_1.txt$
Двойная война и мир - первые десять глав последовательно продублированные



по графику четко видно когда начинается копипаста, из-за чего он похож на лестницу

test/cp_2.txt

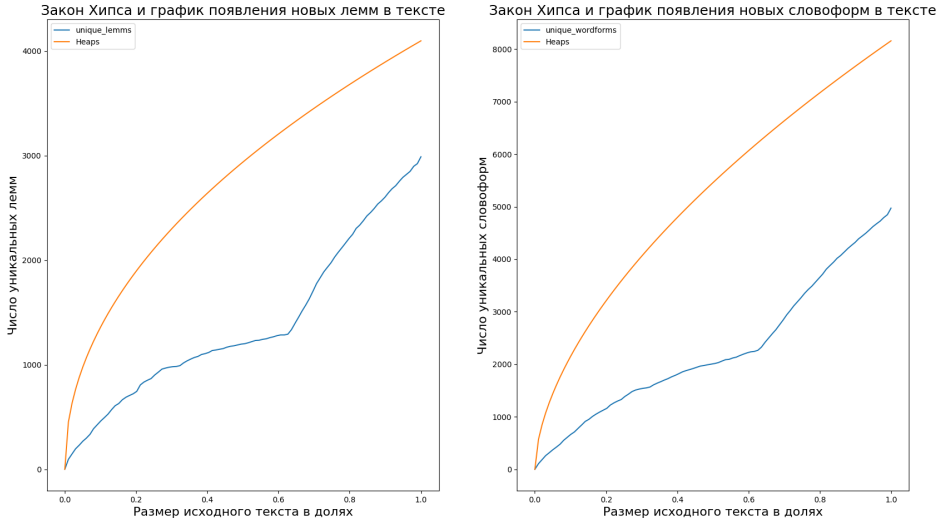
Первая глава Война и мир десять раз



ожидаемо видим, как первые 10% текста новые леммы возникали, после чего их рост прекращется и график имеет вид константы

test/cp_3.txt

Конституция РФ + 5 первых глав "Война и мир"



Тут уже видим, что по началу новые леммы возникали довольно медленно, что хаарктерно официально-деловому тексту, а после следует значительный рост, когда начинается художественный текст Толстого.

6 Вывод

По итогу работы можно сказать, что закон Хипса полезный инструмент для работы с текстами, с его помощью можно проверять текста на плагиат, и даже проверять лексическую принадлежность текста. Но по последнему пункту, к сожалению, это не такой гибкий инструмент и лучше всего он позволяет сравнивать две противоположности: художественные и официально-деловые текста. Разница между ними достаточно значительна и как показано в тесте $test/cp\_3.txt$ с его помощью их можно различать, но к сожалению отличить, например, учебную литературу от официально делового будет намного сложнее, так что тут лучше использовать другие методы.