Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №3 по курсу «Информационный поиск»

Студент: В. М. Филиппов
Преподаватель: А. А. Кухтичев
Группа: М8О-410Б

Дата:
Оценка:
Подпись:

Москва, 2025

Лабораторная работа №3 «Токенизация, Стэмминг, Закон Ципфа»

1 Токенизация

Нужно реализовать процесс разбиения текстов документов на токены, который потом будет использоваться при индексации. Для этого потребуется выработать правила, по которым текст делится на токены. Необходимо описать их в отчёте, указать достоинства и недостатки выбранного метода. Привести примеры токенов, которые были выделены неудачно, объяснить, как можно было бы поправить правила, чтобы исправить найденные проблемы.

В результатах выполнения работы нужно указать следующие статистические данные:

- Количество токенов.
- Среднюю длину токена.

Кроме того, нужно привести время выполнения программы, указать зависимость времени от объёма входных данных. Указать скорость токенизации в расчёте на килобайт входного текста. Является ли эта скорость оптимальной? Как её можно ускорить?

2 Закон Ципфа

Для своего корпуса необходимо построить график распределения терминов по частотностям в логарифмической шкале, наложить на этот график закон Ципфа. Объяснить причины расхождения.

В качестве дополнительного задания, можно (но необязательно) подобрать константы для закона Мандельброта, наложить полученный график на график распределения терминов по частотностям. Привести выбранные константы.

3 Стемминг

Добавить в созданную поисковую систему лемматизацию. В простейшем случае, это просто поиск без учёта словоформ. В более сложном случае, можно давать бонус большего размера за точное совпадение слов.

Лемматизацию можно добавлять на этапе индексации, можно на этапе выполнения поискового запроса. В отчёте должна быть включена оценка качества поиска, после

внедрения лемматизации. Стало ли лучше? Изучите запросы, где качество ухудшилось. Объясните причину ухудшения и как можно было бы улучшить качество поиска по этим запросам, не ухудшая остальные запросы?

1 Описание

В рамках этой лабораторной работы, я реализовал токенизатор и стэммизатор на языке C++. Также построил на Python график для закона Ципфа.

1 Токенизатор

Алгоритм проходит по тексту посимвольно и решает: добавить символ в текущий токен или завершить текущий токен и начать новый.

Базовые символы, то есть буквы, цифры и символ подчёркивания включаются в состав токена.

Специальные символы алгоритм обрабатывает по-умному. Он умеет "склеивать"символы пунктуации с буквами, если они образуют логическую единицу:

- Числа с плавающей точкой (. или ,): Разрешает одну точку или запятую внутри числа (например, 3.14 или 10,5), если по обе стороны находятся цифры.
- Дефисы:
  - В начале слова: если после него идет цифра (отрицательные числа, например -5).
  - В середине слова: если он стоит между двумя буквами/цифрами (составные слова, например high-tech).
- Апострофы: Разрешает апостроф внутри слова, если за ним следует буква или цифра (например, it's или don't).

Такой токенайзер плохо справляется, например, с почтами и ссылками внутри текста. Например строку "Contact me at user@example.com today"он разберёт как {"contact" "me" "at" "user" "example" "com" "today"}. Почта разбилась на несколько отдельных токенов, вместо одного единого. Также некорректно разбираются строки вида "Version 1.2.3 costs $99.99 (50% off)". Из этой строки получится {"version" "1.2" "3" "costs" "99.99" "50" "off"}. Мы теряем знаки доллара и процентов. children3educationwits вот такие странные токены я встретил, кажется это раньше было ссылкой или что-то типо того. Можно запретить токены с символами и цифрами одновременно.

Для моего корпуса документов почты встречаются очень редко, так что не стоит заморачиваться над отдельными правилами для их разбора. Проценты и знаки денег также редкое явление, так что их можно разбирать.

Токенизация для корпуса размером в 305135 документов или 2661.52 МБ заняла 17.9245 секунды, со скоростью 148.485 МБ/с. Средний размер токена вышел 8.84495

символа, общее количество токенов составило 3552265. В среднем длина слова на английском составляет 4.7 символа, у мой токенизатор выдал больше. Я связываю это, во-первых, с тем, что корпус документов содержит в себе спортивные термины, фамилии, названия клубов, которые по длине в среднем больше, чем слова из обихода. Во-вторых, вероятно, это связано с проблемами в выделении текста из HTML страницы: из-за использования специфических тэгов, которые не распознаёт GUMBO слова могут склеиваться.

2 Стеммер

Для стемминга используются алгоритм Портера. Цель стемминга - привести разные формы одного слова (например, connections, connected, connecting) к единой основе (connect). Центральное понятие в этом алгоритме – мера. Любое слово английского языка можно представить в виде $[C](VC)^m[V]$, где

- C — последовательность из одной или более согласных.
- V — последовательность из одной или более гласных.
- m — мера слова.

Алгоритм разрешает отрезать суффикс только в том случае, если оставшаяся часть слова достаточно «длинная» (имеет достаточную меру m). Это нужно, чтобы не превратить короткие слова типа sky или be в бессмысленные обрубки.

Основная часть алгоритма состоит из пяти этапов.

1. Удаление простых окончаний: Сначала убираются множественное число (-s, -es) и простые глагольные формы (-ed, -ing).
2. Сжатие длинных суффиксов: На втором и третьем этапах сложные латинские и технические суффиксы заменяются на более простые.
3. Финальная обрезка: На четвертом этапе удаляются оставшиеся суффиксы типа -ance, -able, -ion, если слово остается узнаваемым.
4. Чистка: В самом конце алгоритм убирает лишние буквы (например, немую -e в конце слова или двойные согласные).

Поскольку алгоритму не нужен огромный словарь всех слов языка, он работает невероятно быстро и занимает минимум памяти.

Однако, поскольку это просто набор правил, то случаются и промахи

- Перебор: Может превратить universal, university и universe в одну основу univers, хотя это разные понятия.
- Недобор: Может не понять, что knew и know — это одно и то же, потому что это неправильный глагол, а алгоритм знает только правила.

Токенизация вместе со стэммингом для корпуса размером в 305135 документов или 2661.52 МБ заняла 191.318 секунды, со скоростью 13.9115 МБ/с. Средний размер токена вышел 8.59844 символа, общее количество токенов составило 3142050. Средний размер токена упал на 2.8%, а количество токенов упало на 13.05%. На самом деле, очень небольшое падение. Вероятно, это связано с тем, что около 900000 токенов - это числа, а еще в корпусе много имён собственных, которые плохо сокращаются стеммером.

3 Закон Ципфа

Закон Ципфа — это эмпирическая закономерность, которая описывает распределение частоты слов в естественном языке. Если говорить просто: в любом достаточно большом тексте (или корпусе текстов) самое частое слово встречается примерно в 2 раза чаще, чем второе по частоте, в 3 раза чаще, чем третье, и так далее.

$$f(r) = \frac{C}{r^s}$$

, где

- $f(r)$ — частота встречаемости слова.
- C — константа (соответствует частоте самого популярного слова, если s=1).

или в развёрнутом виде

$$P_r = \frac{\frac{1}{r^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

, где

- f — частота (вероятность появления) слова с рангом r.
- r — ранг слова (позиция в списке частотности по убыванию).
- N — общее количество слов в словаре (размер лексикона).
- s — показатель степени, характеризующий распределение (в классическом законе Ципфа для естественных языков $s = 1$).

Закон Ципфа для токенов со стеммингом имеет более ярко выраженный горб, который оказывается прижат к линии, если стемминг не использовать. Это легко объяснимо: мы искусственно увеличиваем частоту встречи символа в тексте, поэтому график отклоняется от теоретических значений.

Резкий уход вниз от теоретического закона связан с тем, что в токенах оказывается очень большое количество чисел, зачастую уникальных, поэтому вот такое "падение"графика.

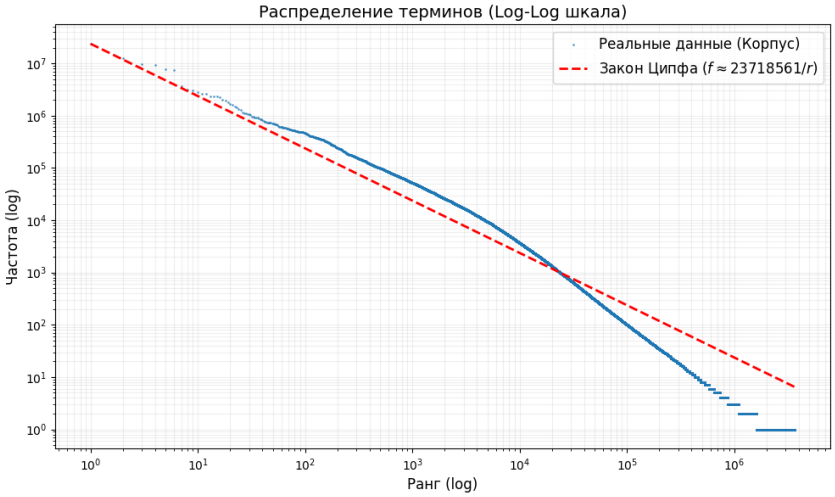

Рис. 1: Распределение Ципфа для токенов без стемминга

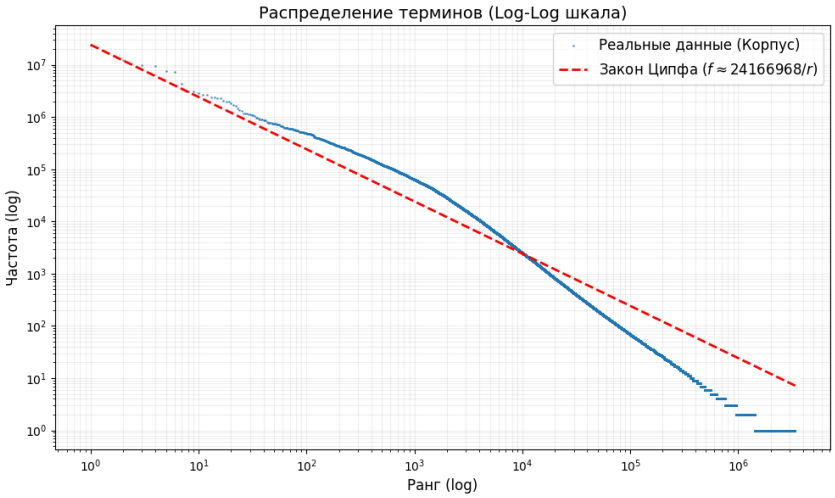

Рис. 2: Распределение Ципфа для токенов со стеммингом

2 Исходный код

Для реализации токенизатора используется класс Tokenizer, внутри которого реализуется метод tokenize, который делает основную токенизацию. Также есть метод getRawTokens, который будет необходим позже для сбора всяких служебных символов из запроса: отрицание, пересечение и объединение.

Для реализации стеммера используется интерфейс IStemmer вместе с реализациями DummyStemmer и PorterStemmer. Интерфейс необходим для того, чтобы легко подменять для тестирования отсутствия стеммера в лице DummyStemmer, который возвращает просто тоже слово при стемминге. PorterStemmer реализует алгоритм Портера, включая методы шагов, измерения меры, проверки CVC последовательностей и прочее.

3 Выводы

При выполнении данной лабораторной работы я узнал, каким образом реализуется стемминг слов в естественном языке. Также я познакомился с законом Ципфа, который меня сильно удивил. Это очень красивая закономерность. С токенизацией я так или иначе сталкивался и до этого, но было полезным еще раз потренироваться в ёё написании.

Список литературы

[1] Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск / Пер. с англ. Д. А. Клюшина. — М.: Вильямс, 2011. — 528 с. — ISBN 978-5-8459-1623-4.

[2] The Porter Stemming Algorithm [Электронный ресурс] // Tartarus. URL: https://tartarus.org/martin/PorterStemmer/ (дата обращения: 15.12.2025).

[3] Как поисковые системы оценивают тексты: закон Ципфа, индекс туманности и прочее [Электронный ресурс] // Pitcher. URL: https://pitcher.agency/blog/zakon-cipfa (дата обращения: 16.12.2025).