Слова — Лексическая статистика

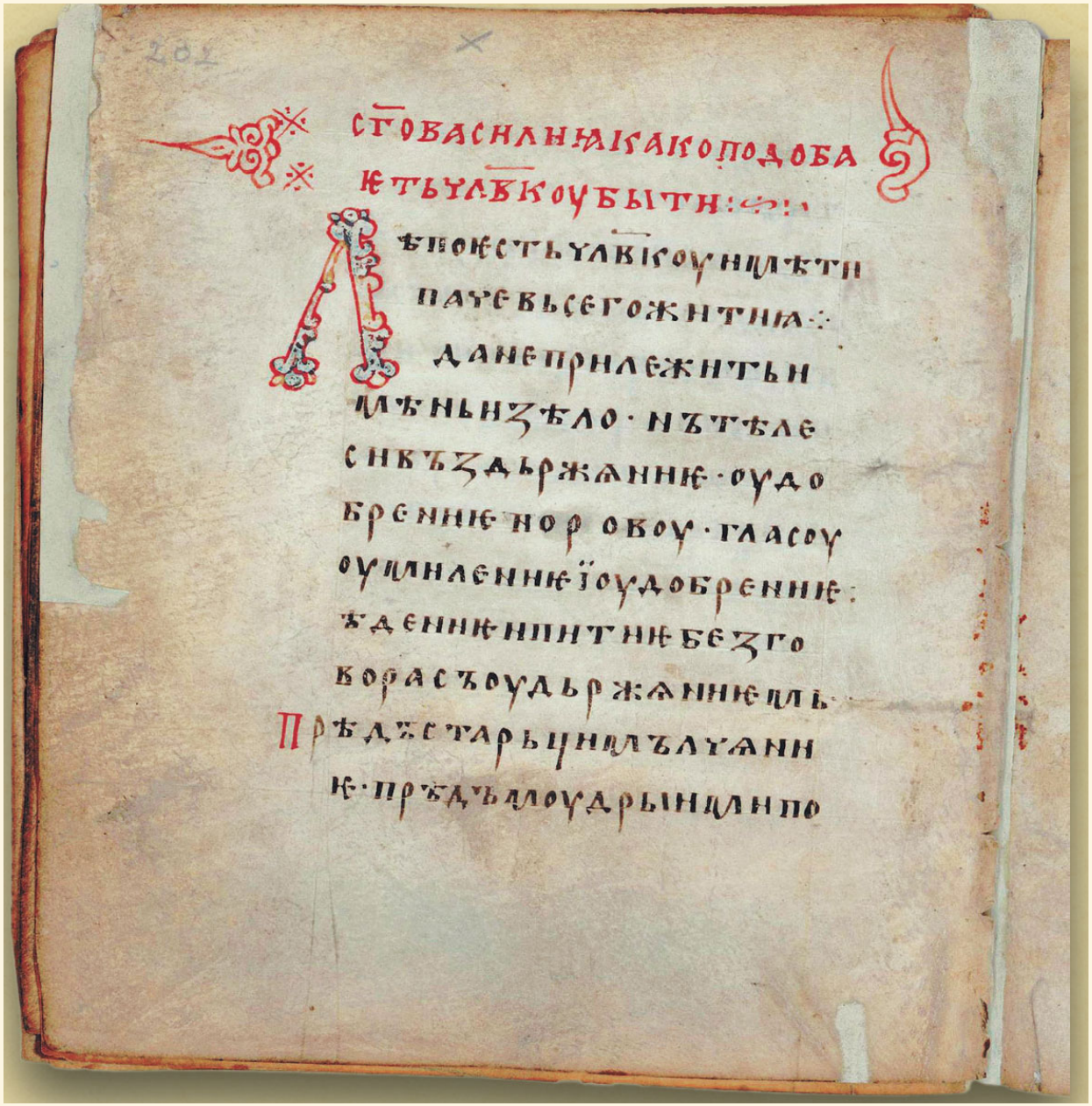
Квантитативный анализ текста

Кирилл Александрович Маслинский

07.02.2022 / 01

Институт русской литературы (Пушкинский Дом) РАН

Предисловие


СТОВАСНЛНѦІѢАКОПОДОБА
ЄТЬУЛѢКОУБЫТИ:
ѢПОЖСТЬУЛѢІСОУНЦЛѢТИ
ПАУЄВЬСЄГОЖИТИА·
ДАНЄПРИЛЄЖИТЬИ
ЦѢНЬИЖѢЛО·НЪТѢЛЄ
СНѢЪЖДЬРЖАННЄ·ОУДО
БРЄННЄНОР ОВОУ·ГЛАСОУ
ОУЦЛНЛЄННЄЇОУДОБРЄННЄ:
ѢДЄННЄНПИТНЄБЄЖГО
ВОРАСЪОУДЬРЖАННЄЦЛЬ·
ПРѢДЪСТАРЬЦНЦЛЪЛУАНИ
Ѥ·ПРѢДЪЛОУДРЬІНЦЛНПО

When Irish monks began separating words in manuscripts by spaces in the seventh century, little could they know that they were performing a central task of computational text analysis.

Andrew Piper, Enumerations: Data and literary study (2018) p. 42

Лексическая статистика

Как начать считать слова

- сколько разных слов в сообщении в мессенджере?
- на странице научной статьи?
- сколько новых слов на каждой следующей странице,
- сколько новых слов во второй половине книги (по сравнению с первой)?
- сколько новых слов в книге по сравнению с другими книгами?
- когда наконец мы перестанем встречать новые слова?

Как начать считать слова

- сколько разных слов в сообщении в мессенджере?
- на странице научной статьи?
- сколько новых слов на каждой следующей странице,
- сколько новых слов во второй половине книги (по сравнению с первой)?
- сколько новых слов в книге по сравнению с другими книгами?
- когда наконец мы перестанем встречать новые слова?

Как начать считать слова

- сколько разных слов в сообщении в мессенджере?
- на странице научной статьи?
- сколько новых слов на каждой следующей странице,
- сколько новых слов во второй половине книги (по сравнению с первой)?
- сколько новых слов в книге по сравнению с другими книгами?
- когда наконец мы перестанем встречать новые слова?

Как начать считать слова

- сколько разных слов в сообщении в мессенджере?
- на странице научной статьи?
- сколько новых слов на каждой следующей странице,
- сколько новых слов во второй половине книги (по сравнению с первой)?
- сколько новых слов в книге по сравнению с другими книгами?
- когда наконец мы перестанем встречать новые слова?

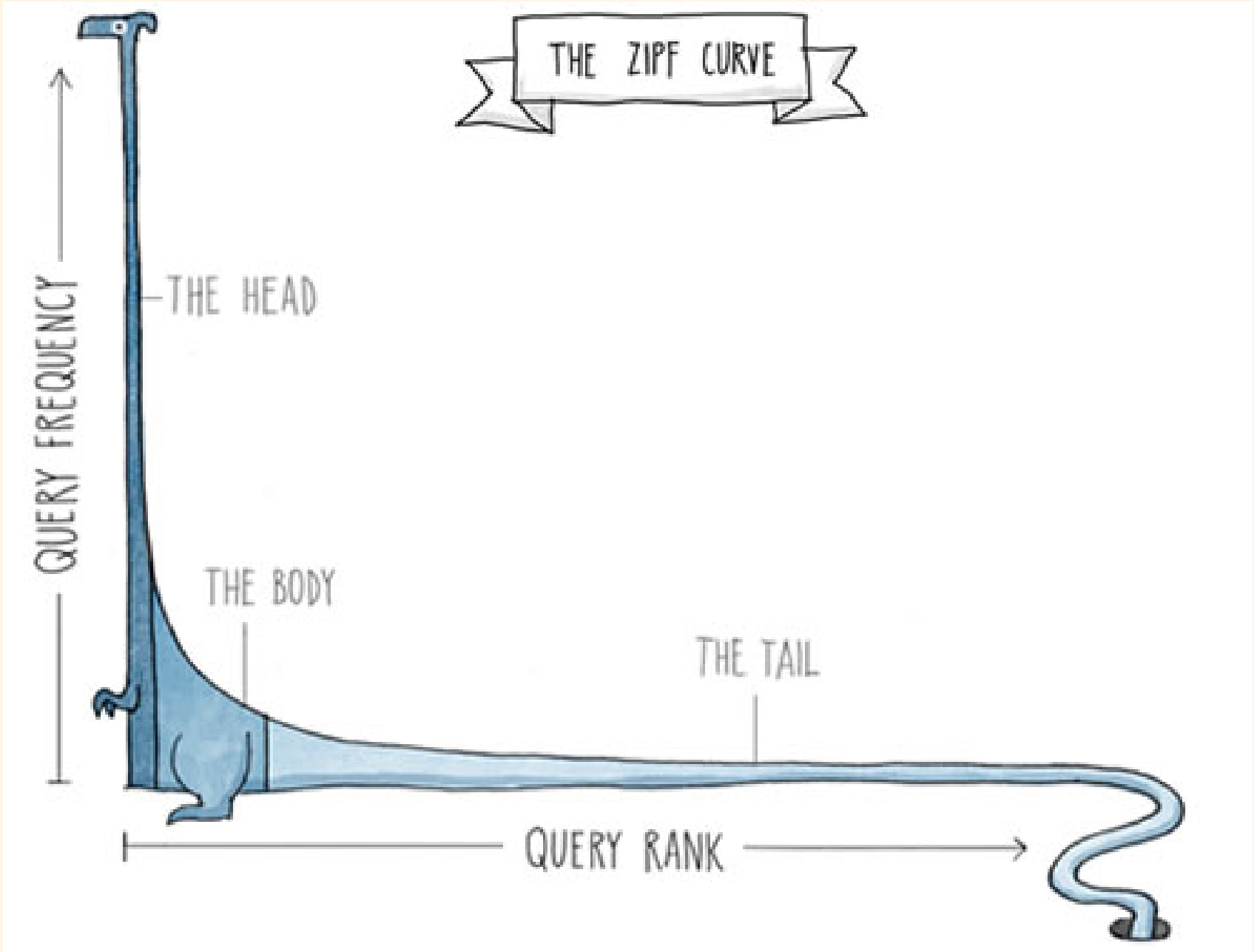
Как начать считать слова

- сколько разных слов в сообщении в мессенджере?
- на странице научной статьи?
- сколько новых слов на каждой следующей странице,
- сколько новых слов во второй половине книги (по сравнению с первой)?
- сколько новых слов в книге по сравнению с другими книгами?
- когда наконец мы перестанем встречать новые слова?

Как начать считать слова

- сколько разных слов в сообщении в мессенджере?
- на странице научной статьи?
- сколько новых слов на каждой следующей странице,
- сколько новых слов во второй половине книги (по сравнению с первой)?
- сколько новых слов в книге по сравнению с другими книгами?
- когда наконец мы перестанем встречать новые слова?

План


THE ZIPF CURVE
QUERY FREQUENCY
THE HEAD
THE BODY
THE TAIL
QUERY RANK

Закон Ципфа

Предсказывает частотность слова по его рангу в частотном списке:

$$f(w) = \frac{C}{r(w)^a} \tag{1}$$

$f(w)$ — частотность слова w

$r(w)$ — ранг слова w в частотном списке

C — константа

a — константа, близкая к 1.

Предсказания закона Ципфа

При $a = 1$, $C = 60000$ закон Ципфа предсказывает:

$$f(w) = \frac{C}{r(w)}$$

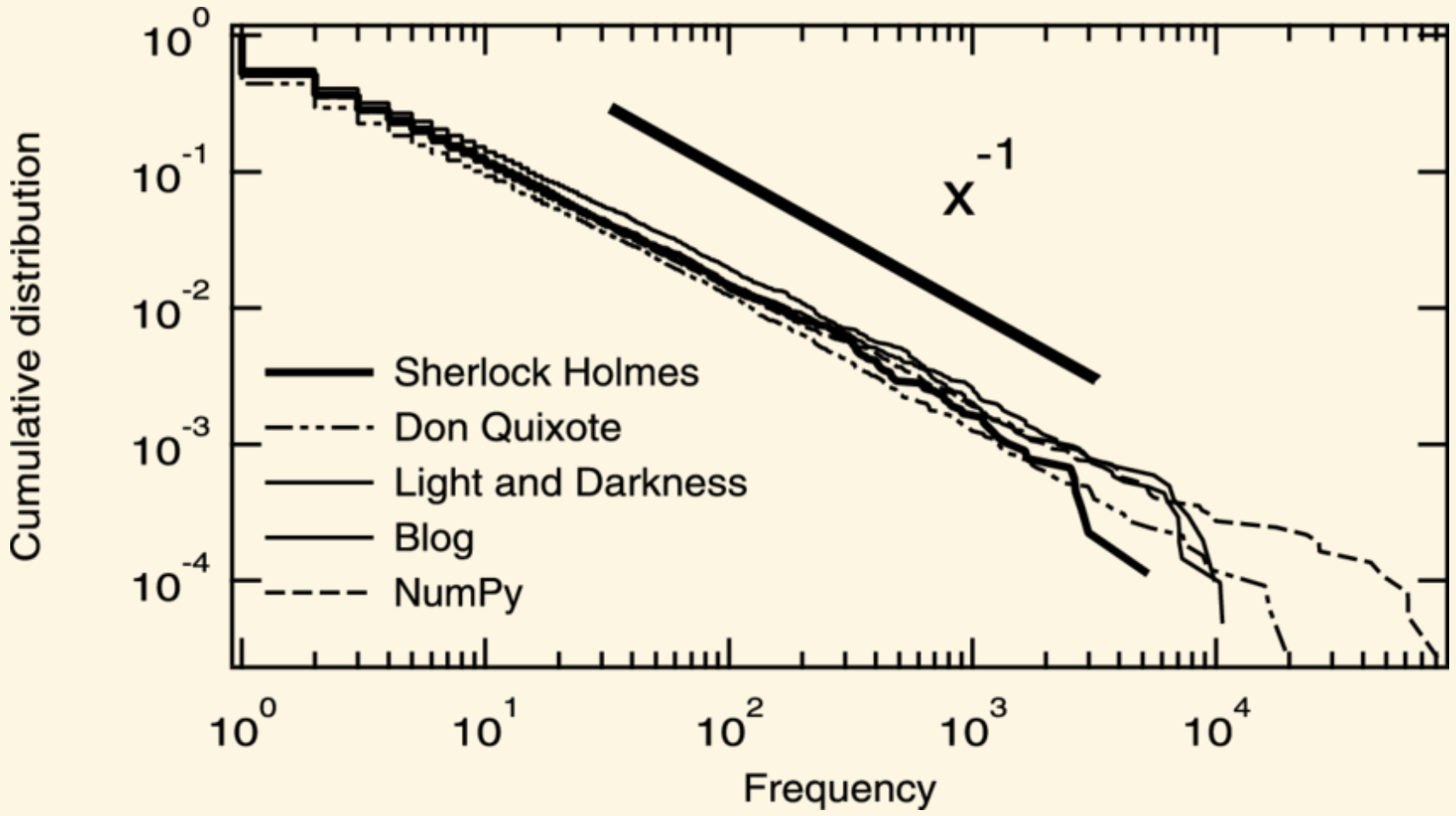
- самое частотное слово встретится $f(w) = C/1 = 60000$ раз
- второе по частотности слово $C/2 = 30000$ раз
- третье по частотности слово $C/3 = 20000$ раз
- сотое $C/100 = 600$ раз
- сто первое $C/101 = 594,06$ раз (около 99% частотности сотого)
- и длинный хвост из 80000 слов с частотностью между $1,5$ и $0,5$.

Логарифмическая форма закона Ципфа

$$\log f(w) = log(C) - a \log r(w) \quad (2)$$

Линейная функция:

$$y = kx + b$$


Cumulative distribution
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
x^{-1}
Sherlock Holmes
Don Quixote
Light and Darkness
Blog
NumPy
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
Frequency

Закон Ципфа-Мандельброта (1953)

$$f(w) = \frac{C}{(r(w) + b)^a} \tag{3}$$

При $C = 60000$, $a = 1$, $b = 1$ предсказанная частотность самого частотного слова:

Закон Ципфа $\frac{C}{1} = \frac{60000}{1} = 60000$

Закон Ципфа-Мандельброта $\frac{C}{r+b} = \frac{60000}{(1+1)} = 30000$

Объяснения закона Ципфа

1. Психолингвистическое (Ципф):
   - экономия усилий говорящего (меньше разных слов);
   - экономия усилий слушающего (больше разных слов).
2. Теоретико-информационное (Мандельброт):
   - минимизация средней стоимости передачи информации в тексте.
3. Процесс, приводящий к подобному распределению:
   - новые слова с константной вероятностью (Simon 1955);
   - «обезьяна и пишущая машинка» (Miller 1957).

Степенное распределение

Few Giants — Many dwarfs

Примеры

- частотности слов;
- размеры городов;
- распределение дохода (закон Парето).

План

Открытые и закрытые классы слов

Словарь языка незамкнут — всё время возникают новые слова.

Function words, closed-class Вершину частотного списка занимают служебные части речи (предлоги, союзы, местоимения). Все единицы перечислимы, пополняется очень медленно. В тексте выполняют прежде всего грамматическую функцию.

Content words, open-class Далее в частотном списке преобладают слова открытых классов (пополняемых), прежде всего существительные. В тексте выполняют прежде всего референтную функцию.

Пример: частотность русской лексики

Единица измерения частотности:

ipm — вхождений на миллион / instances per million

```
1 36358.94 и misc
2 27792.36 в prep
3 20689.51 не misc
4 18942.62 он pron
5 16588.14 на prep
6 15631.11 я pron
7 12546.08 что misc
8 11398.44 тот adjpron
9 11223.99 быть verb
10 11150.72 с prep
11 9808.61 а misc
```

```
32600 1.04 мертветь verb
32601 1.04 сволочной adj
32602 1.04 втыкаться verb
32603 1.04 нахлебник noun
32604 1.04 русоволосый adj
32605 1.04 автопилот noun
32606 1.04 иссечение noun
32607 1.04 бульдожий adj
32608 1.04 бренность noun
32609 1.04 нездоровье noun
32610 1.04 саргасса noun
```

План

Лексическая статистика

- Закон Ципфа
- Частотность и состав лексикона
- Размер и скорость роста словаря
- Размер лексикона/Лексическое разнообразие
- Практические следствия закона Ципфа

Скорость роста словаря

Чем дальше мы читаем текст, тем реже встречаем новые слова.

Оценка Гаральда Баайена (Baayen G):

$$G = \frac{V(1)}{N} \tag{4}$$

где:

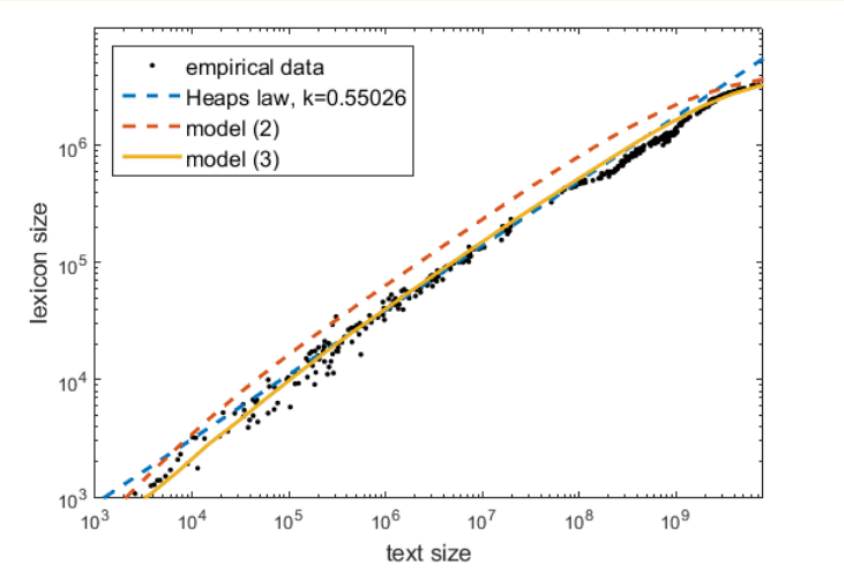
$V(1)$ — количество hapax legomena на N токенов текста

N — количество токенов текста.

Рост словаря

Чем дальше мы читаем текст, тем реже встречаем новые слова.

$$V = kN^{\beta} \tag{5}$$



V — размер словаря

N — размер корпуса

k — константа (обычно 10—100)

β — константа $0 < \beta < 1$ (обычно 0,4—0,6)

Probability to see a word

Maximum Likelihood Estimation (MLE) — based on the observed frequency in the corpus:

$$P = \frac{f(w)}{N}$$

где

P — probability;

$f(w)$ — word frequency w;

N — corpus size.

Probability to see a word

Model Based Estimation (LNRE model) — based on the knowledge of the general properties of word distributions:

$$P = \frac{C}{(r(w) + b)^a}$$

P — probability;

$f(w)$ — word rank w in a frequency list;

a, b — model parameters;

C — normalizing constant.

This is Zipf-Mandelbrot model

План

Коэффициент лексического разнообразия

Одна из первых и широко используемых мер сложности речи/текста.

$$TTR = \frac{V}{N} \tag{6}$$

V — размер словаря, число разных словоформ/лемм в тексте (types)

N — число словоформ в тексте

Обратная величина: средняя частотность слов в тексте

$$F_{mean} = \frac{N}{V} \tag{7}$$

Коэффициент лексического разнообразия

Одна из первых и широко используемых мер сложности речи/текста.

$$TTR = \frac{V}{N} \tag{6}$$

V — размер словаря, число разных словоформ/лемм в тексте (types)

N — число словоформ в тексте

Обратная величина: средняя частотность слов в тексте

$$F_{mean} = \frac{N}{V} \tag{7}$$

TTR: проблемы

Классические применения:

- определение авторства
- оценка сложности (детской) речи — развития речи

Проблемы:

- зависит от длины текста, длиннее текст — ниже TTR (r=0.99).
- зависит от способа выделения types (словоформы/леммы)

Нормализованная версия TTR:

- для сравнения используются фрагменты текста одинаковой длины.

План

Практические следствия закона Ципфа

1. Data sparseness — в сколь угодно большом корпусе:
   - почти все слова встречаются очень редко;
   - небольшая группа частотных слов составляет значительную часть токенов корпуса;
   - LNRE — Large Number of Rare Events.
2. Рост словаря — даже очень большие корпуса не содержат всех слов языка:
   - искаженная оценка вероятности слова по частотности в корпусе;
   - нельзя использовать размер словаря для оценки степени лексического разнообразия текста.
3. Знания о распределении слов в любом тексте можно использовать для оптимизации и построения моделей.

Заключение

Takeaways

- Помни о словах, которые еще не встретились. Делай на них **скидку**.
- Никогда не суди о богатстве словаря автора по количеству разных слов в тексте.
- Откинув небольшое число самых частотных слов, можно резко сократить объем корпуса, сохранив бóльшую часть смысловых слов.