Как и зачем считать частотность слов в текстах?

Александр Пиперски

Малый мехмат, 12.05.2018

Случай с Оливером
(Р. М. Фрумкина)

Заглонитель Ланс Оливер чуть не погиб в результате наплочения турма. Он ехал ласкунно на лошади покровнательно от Мэнсфилда (Австралия) и увидел вахню турмов, в которой было кастожно 15 животных. Столенно, ничего бы и не случилось, если бы собака Оливера не начала порочить на вахню.

Один из турмов — старый, крупный лователь, выбатушенный корочением собаки, бросился за ней. Та отпешила скумановаться за лошадью, на которой сидел Оливер. Тогда турм бросился уже на Оливера. Он схватил подвешенца отмаленными твинами за плечи и вытокнул его на землю.

А. С. Пушкин

Я вас любил: любовь еще, быть может,
В душе моей угасла не совсем;
Но пусть она вас больше не тревожит;
Я не хочу печалить вас ничем.
Я вас любил безмолвно, безнадежно,
То робостью, то ревностью томим;
Я вас любил так искренно, так нежно,
Как дай вам бог любимой быть другим.

А. С. Пушкин: частотный словарь

вас	5	*безмолвно*	1
я	4	*безнадежно*	1
любил	3	*бог*	1
не	3	*больше*	1
быть	2	*в*	1
то	2	…	
так	2		

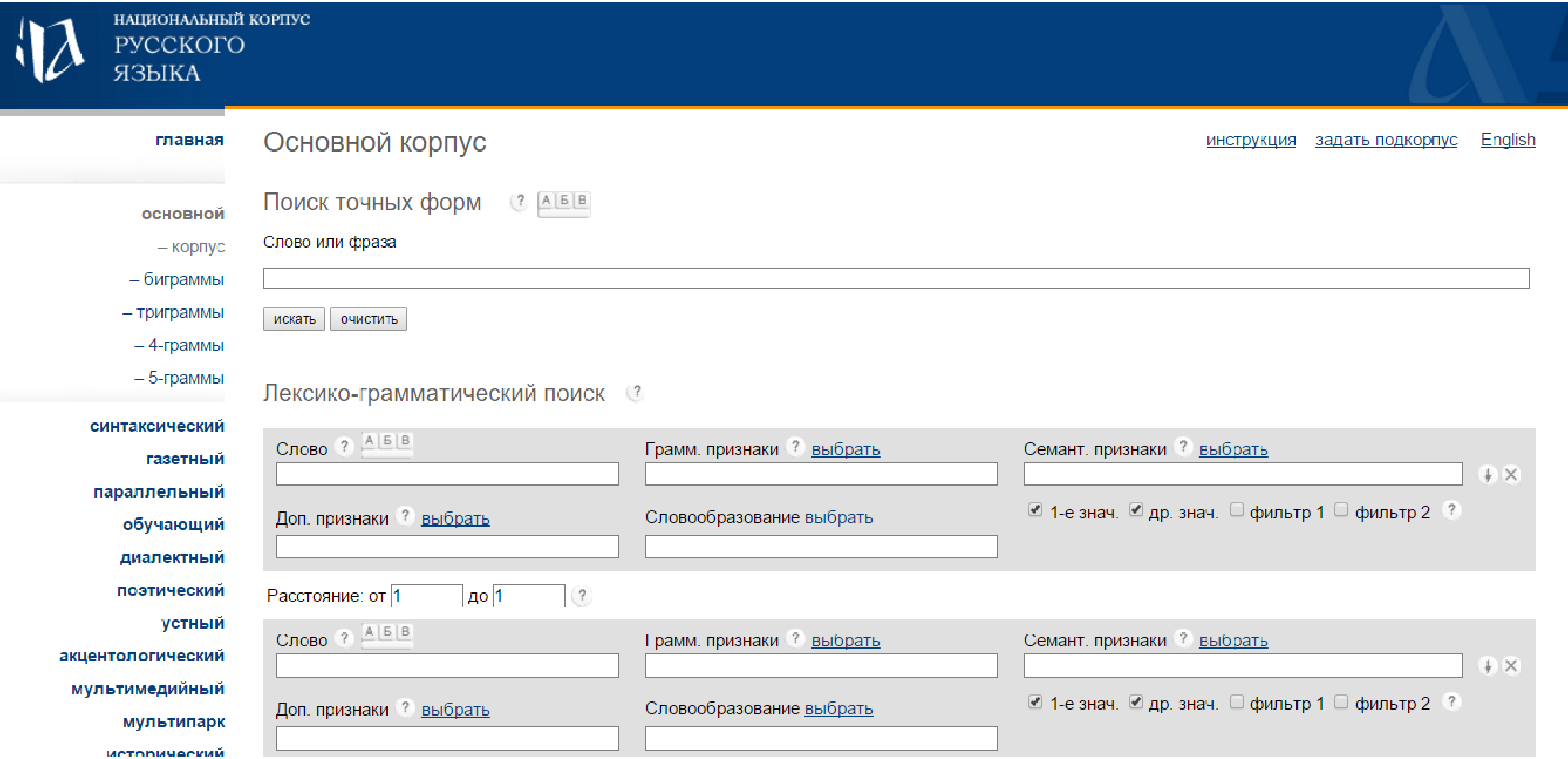
- **(?!)** Слово *любил* встречается чаще, чем слово *не*, а слово *быть* — чаще, чем слова *то* и *как*

Частотные и редкие слова

- *год*, *можно*, *когда* — частотные слова
- *дельный*, *ляжка*, *пакт* — редкие слова
- Как это определить?

Лингвистические корпуса

- Корпус = тексты + разметка + поиск
- Самый известный корпус русского языка — Национальный корпус русского языка (www.ruscorpora.ru)
- Примерно 280 млн словоформ
- XVIII–XXI века



Основной корпус

Поиск точных форм

Слово или фраза

Лексико-грамматический поиск

Слово
Грамм. признаки
Семант. признаки
Доп. признаки
Словообразование
☑ 1-е знач. ☑ др. знач. ☐ фильтр 1 ☐ фильтр 2

Расстояние: от 1 до 1

Слово
Грамм. признаки
Семант. признаки
Доп. признаки
Словообразование
☑ 1-е знач. ☑ др. знач. ☐ фильтр 1 ☐ фильтр 2

Разметка лингвистических корпусов

- Лемматизация — приведение слова к начальной форме (лемме)
- *Мой дядя самых честных правил*
  мой дядя самый честный правило

Частотный словарь

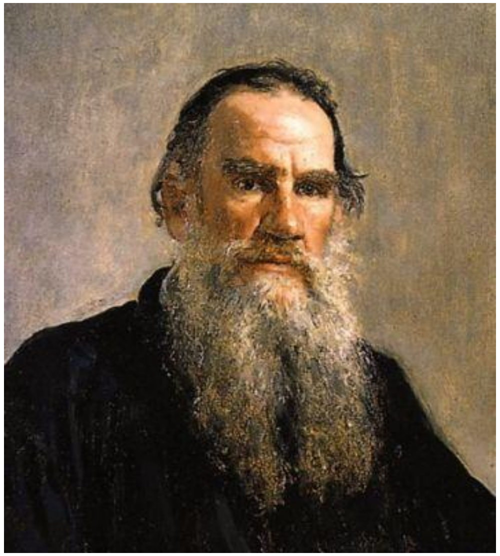
- Если мы знаем, сколько раз в корпусе встречается каждая словоформа / лемма, мы можем построить **частотный словарь** (список слов, в котором каждому слову приписана его частотность)
- Для словоформ — особенно просто (не нужна лемматизация)
- Для лемм — сложнее, но содержательнее

Как считать частотность?

- Корпуса могут быть разного объёма
- *и* встречается в НКРЯ 10 874 602 раза
- Как сравнить частотность этого слова в НКРЯ с частотностью в каком-нибудь корпусе, в котором всего 2 млн словоформ?

Как считать частотность?

- НКРЯ, лемма *телефон*:
  - Лев Толстой — 9 вхождений
  - Анатолий Рыбаков — 6 вхождений
  - У кого чаще?



Как считать частотность?

- **Частотность** =
  = Число вхождений / Объём корпуса (в словах)
- *телефон*
  - Толстой: 9 / 1 906 467 = 0,000005
  - Рыбаков: 6 / 180 583 = 0,000033
- Дробная частотность — неудобно
  ⟹ домножаем на 1 000 000

Частотность (ipm)

- ipm — instances per million words
- *телефон*
  - Толстой: 5 ipm
  - Рыбаков: 33 ipm

Частотные словари русского языка

- Гарри Джоссельсон (Детройт, 1953, 1 млн)
- Эви Штейнфельдт (Таллин, 1963, 400 тыс.)
- Лидия Засорина (ред.; Ленинградский университет + Горьковский университет, 1977, 1 млн)
- Ольга Ляшевская, Сергей Шаров (2009, 92 млн): dict.ruslang.ru/freq.php

Самые частотные русские слова

- Угадайте 10 самых частотных слов русского языка
- Сравниваем со словарём Ляшевской–Шарова

	Лемма	Часть речи	Частота (ipm)
1	и	conj	35801.8
2	в	pr	31374.2
3	не	part	18028.0
4	на	pr	15867.3
5	я	spro	12684.4
6	быть	v	12160.7
7	он	spro	11791.1
8	с	pr	11311.9
9	что	conj	8354.0
10	а	conj	8198.0

Зачем нужны частотные словари?

- Автоматическая обработка текста
- Преподавание языка
    - преподавание носителям
    - преподавание языка как иностранного

Частотность и преподавание иностранного языка

- При изучении иностранного языка надо в первую очередь знать частотные слова
- Сколько слов достаточно, чтобы более-менее понимать текст на иностранном языке?

Случай с Оливером (Р. М. Фрумкина)

Заглонитель Ланс Оливер чуть не погиб в результате наплочения турма. Он ехал ласкунно на лошади покровнательно от Мэнсфилда (Австралия) и увидел вахню турмов, в которой было кастожно 15 животных. Столенно, ничего бы и не случилось, если бы собака Оливера не начала порочить на вахню.

Один из турмов — старый, крупный лователь, выбатушенный корочением собаки, бросился за ней. Та отпешила скумановаться за лошадью, на которой сидел Оливер. Тогда турм бросился уже на Оливера. Он схватил подвешенца отмаленными твинами за плечи и вытокнул его на землю.

Случай с Оливером (Р. М. Фрумкина)

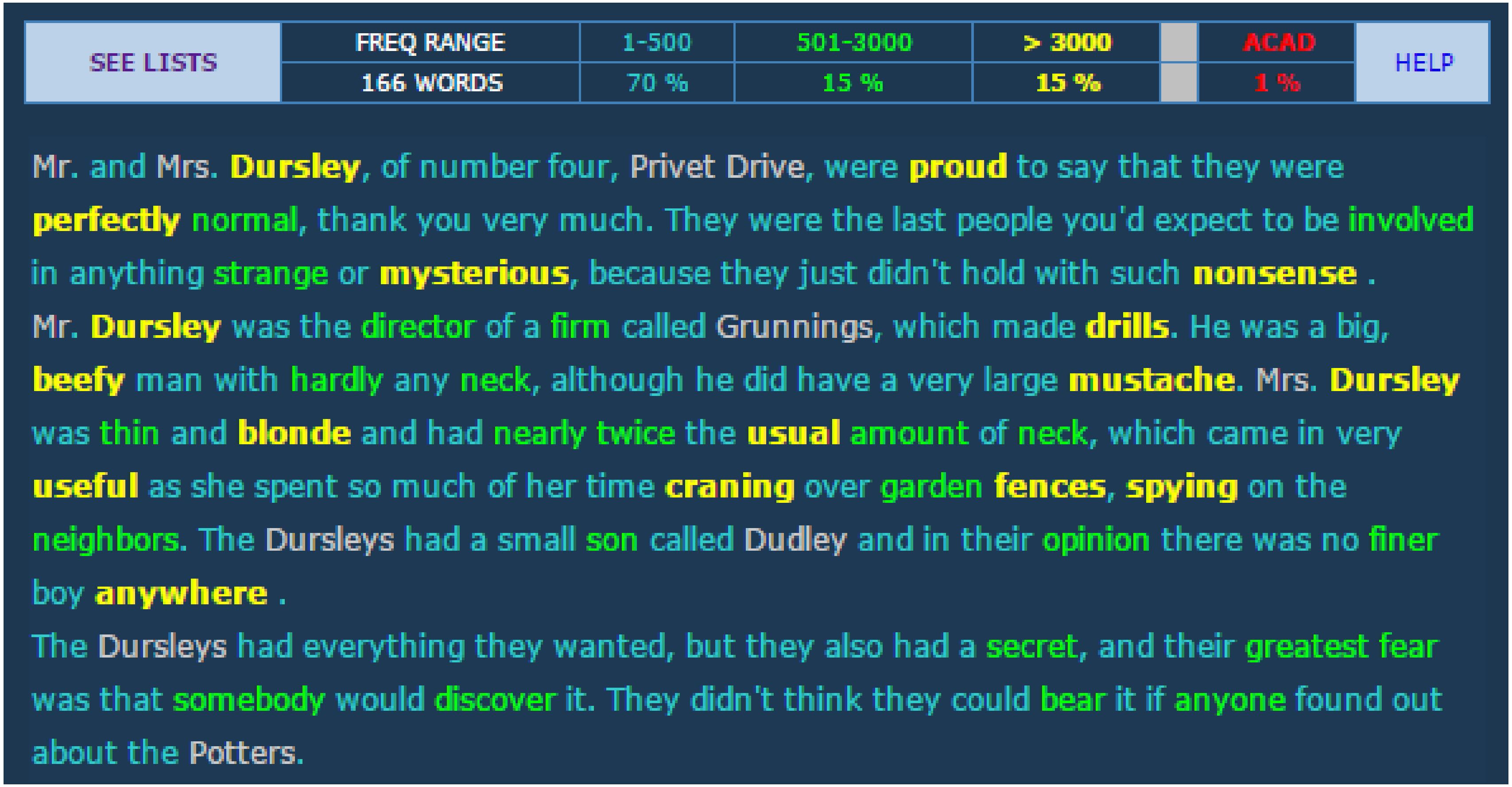
Скотовод Ланс Оливер чуть не погиб в результате нападения кенгуру. Он ехал верхом на лошади неподалеку от Мэнсфилда (Австралия) и увидел стадо кенгуру, в котором было примерно 15 животных. Возможно, ничего бы и не случилось, если бы собака Оливера не начала лаять на стадо.

Один из кенгуру — старый крупный самец, раздраженный лаем собаки, бросился за ней. Та попыталась укрыться за лошадью, на которой сидел Оливер. Тогда кенгуру бросился уже на Оливера. Он схватил всадника передними лапами за плечи и сбросил его на землю.

Случай с Оливером (Р. М. Фрумкина)

- В этом тексте заменены вымышленными слова, которые ниже 2500-го места по частотности в словаре Штейнфельдт
- Но всё равно всё по сути понятно!

wordandphrase.info



SEE LISTS	FREQ RANGE	1-500	501-3000	> 3000		ACAD	HELP
	166 WORDS	70 %	15 %	15 %		1 %	

Mr. and Mrs. **Dursley**, of number four, Privet Drive, were **proud** to say that they were **perfectly** normal, thank you very much. They were the last people you'd expect to be involved in anything strange or **mysterious**, because they just didn't hold with such **nonsense** .

Mr. **Dursley** was the director of a firm called Grunnings, which made **drills**. He was a big, **beefy** man with hardly any neck, although he did have a very large **mustache**. Mrs. **Dursley** was thin and **blonde** and had nearly twice the **usual** amount of neck, which came in very **useful** as she spent so much of her time **craning** over garden **fences**, **spying** on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy **anywhere** .

The Dursleys had everything they wanted, but they also had a secret, and their greatest fear was that somebody would discover it. They didn't think they could bear it if anyone found out about the Potters.

Macmillan Dictionary

★★★ — 1–2500 места в частотном списке (*the*, *animal*)

★★ — 2501–5000 места в частотном списке (*appropriate*, *tragedy*)

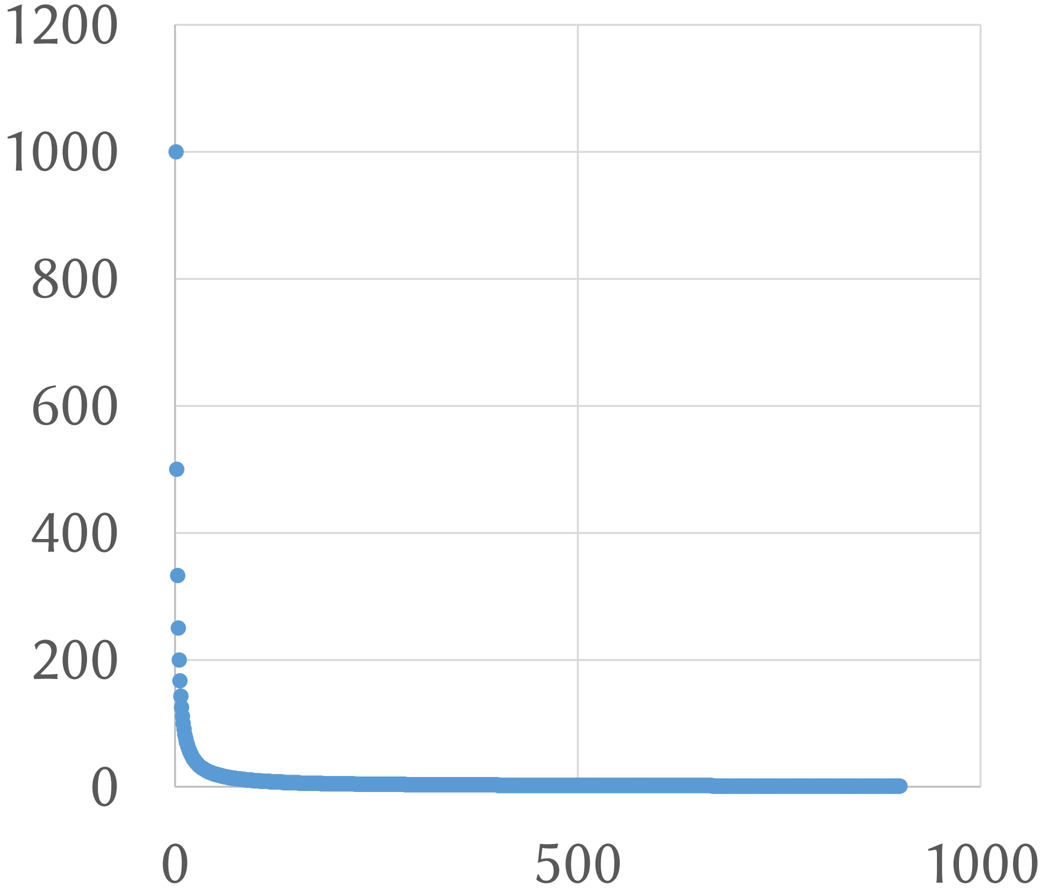
★ — 5001–7500 места в частотном списке (*restriction*, *allegedly*)

без звёдочек — остальные (*crescent*, *thatch*)

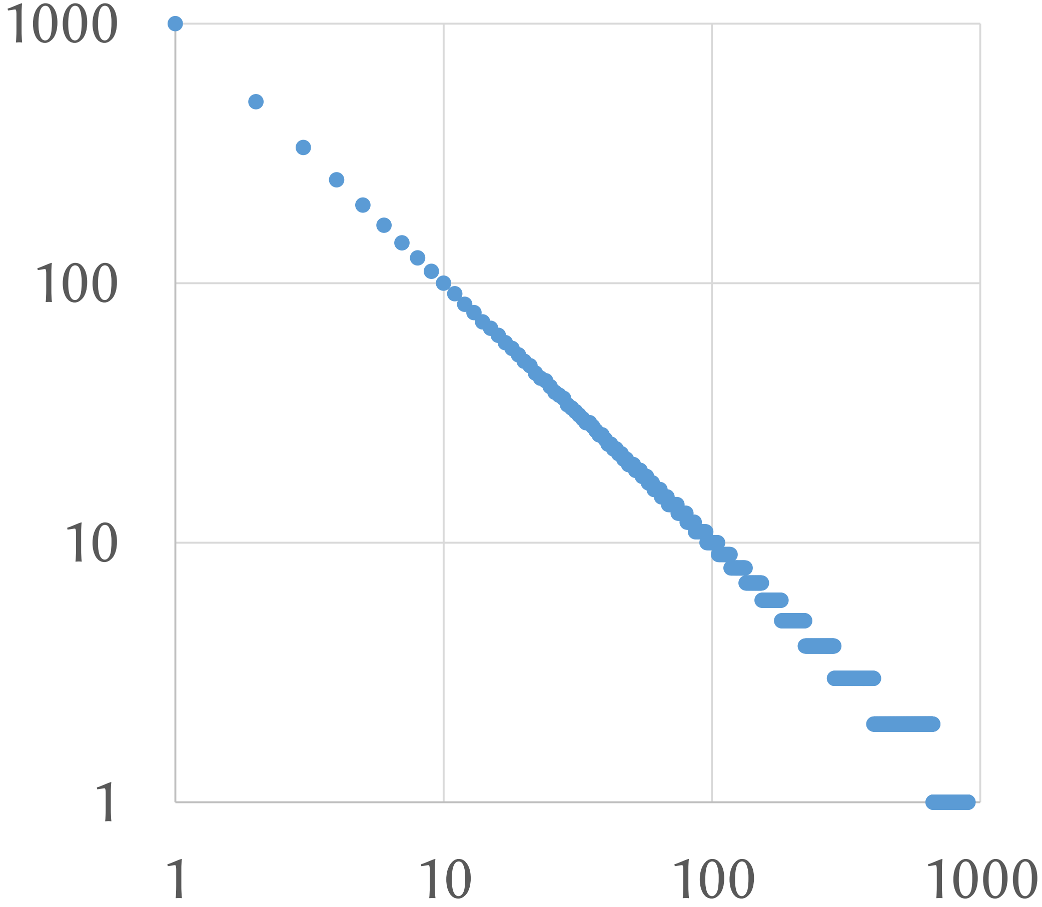
Закон Ципфа

- Джордж Кингсли Ципф (1902–1950)
- Закон (распределение) Ципфа: частотность слова обратно пропорциональна его рангу в частотном списке
- $f(w) = \frac{c}{r(w)}$

Идеальное распределение



Идеальное распределение: логарифмическая шкала



Вопросы на понимание

- Самое частотное слово в корпусе встретилось 60 000 раз
- Сколько раз встретится 2-е по частотности слово? 3-е слово? 100-е слово? 101-е слово? 1000-е слово? 1001-е слово? 60 000-е слово?
- Сколько разных слов мы ожидаем увидеть в таком корпусе?

Пример: Война и мир, том 1

- Длина текста: 110 273 слова
- Слово *и* встречается 4 911 раз
- Всего разных слов: 11 447

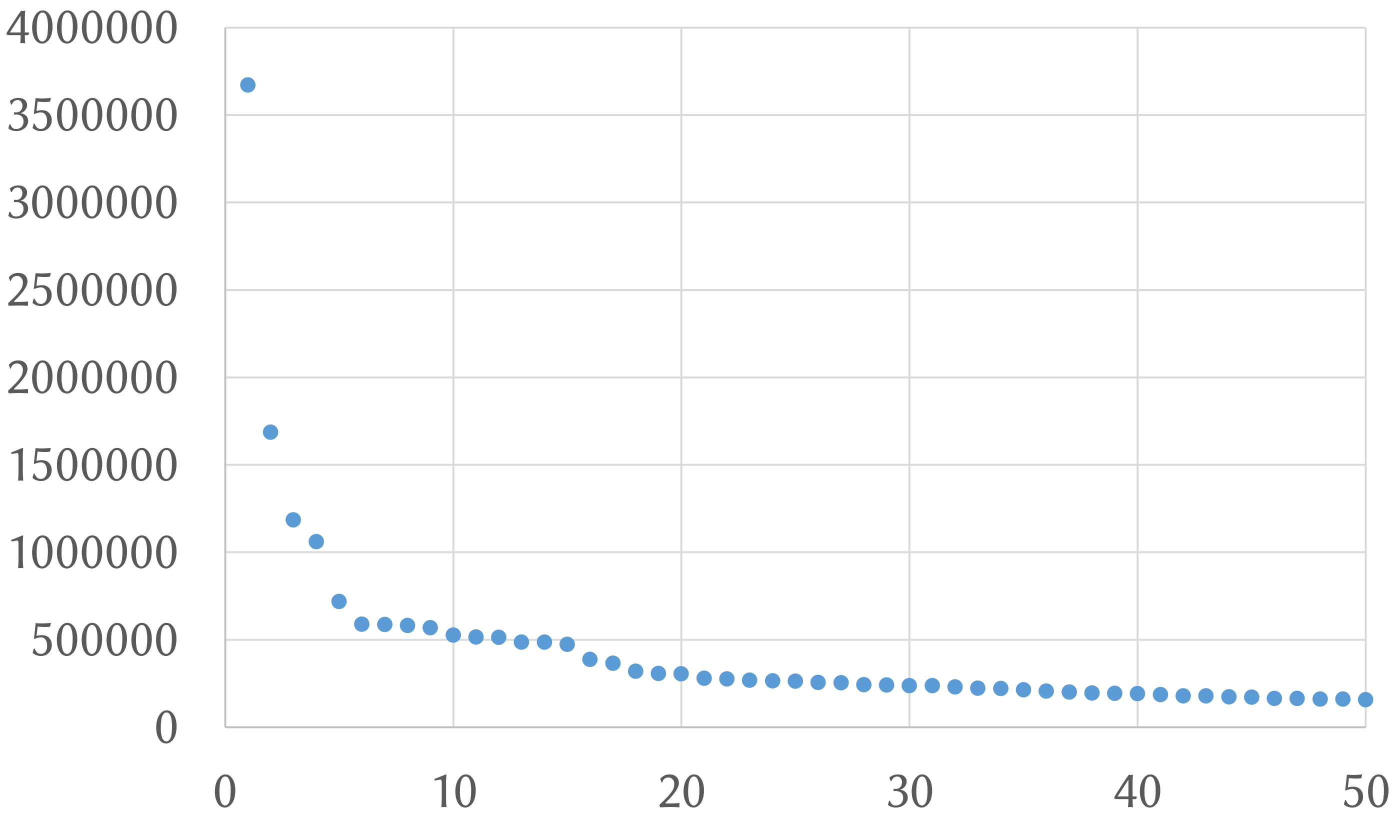
Закон Ципфа в реальной жизни

- Размеры городов
- Количество ссылок на научные статьи и сайты
- …

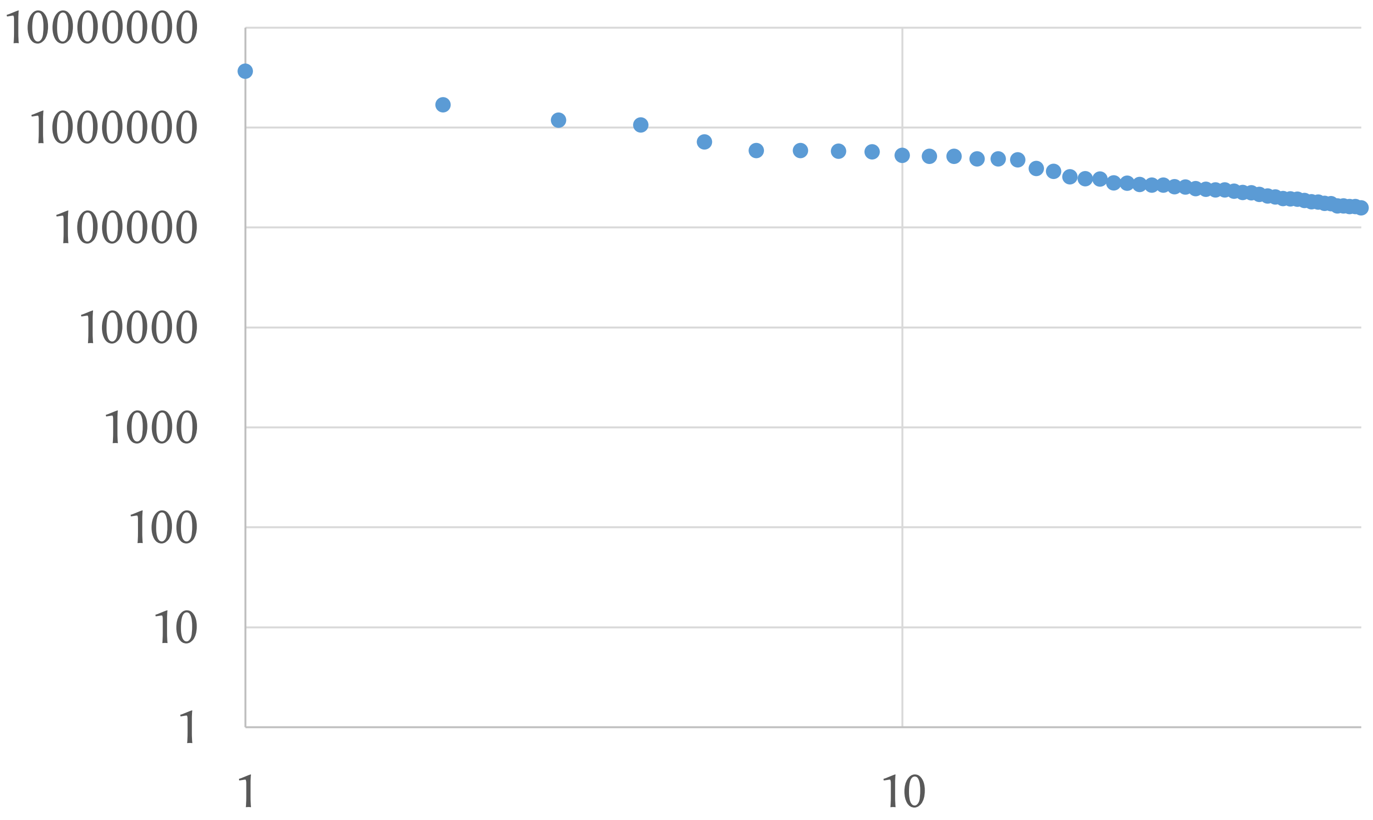
Города Германии

1. Берлин	3 671 000
2. Гамбург	1 686 100
3. Мюнхен	1 185 400
4. Кёльн	1 060 582
5. Франкфурт-на-Майне	720 000
6. Эссен	588 800

Города Германии



Города Германии



Уточнения закона Ципфа

- Прямая может иметь угловой коэффициент $\neq -1$
- Распределение Ципфа в обобщённом виде:

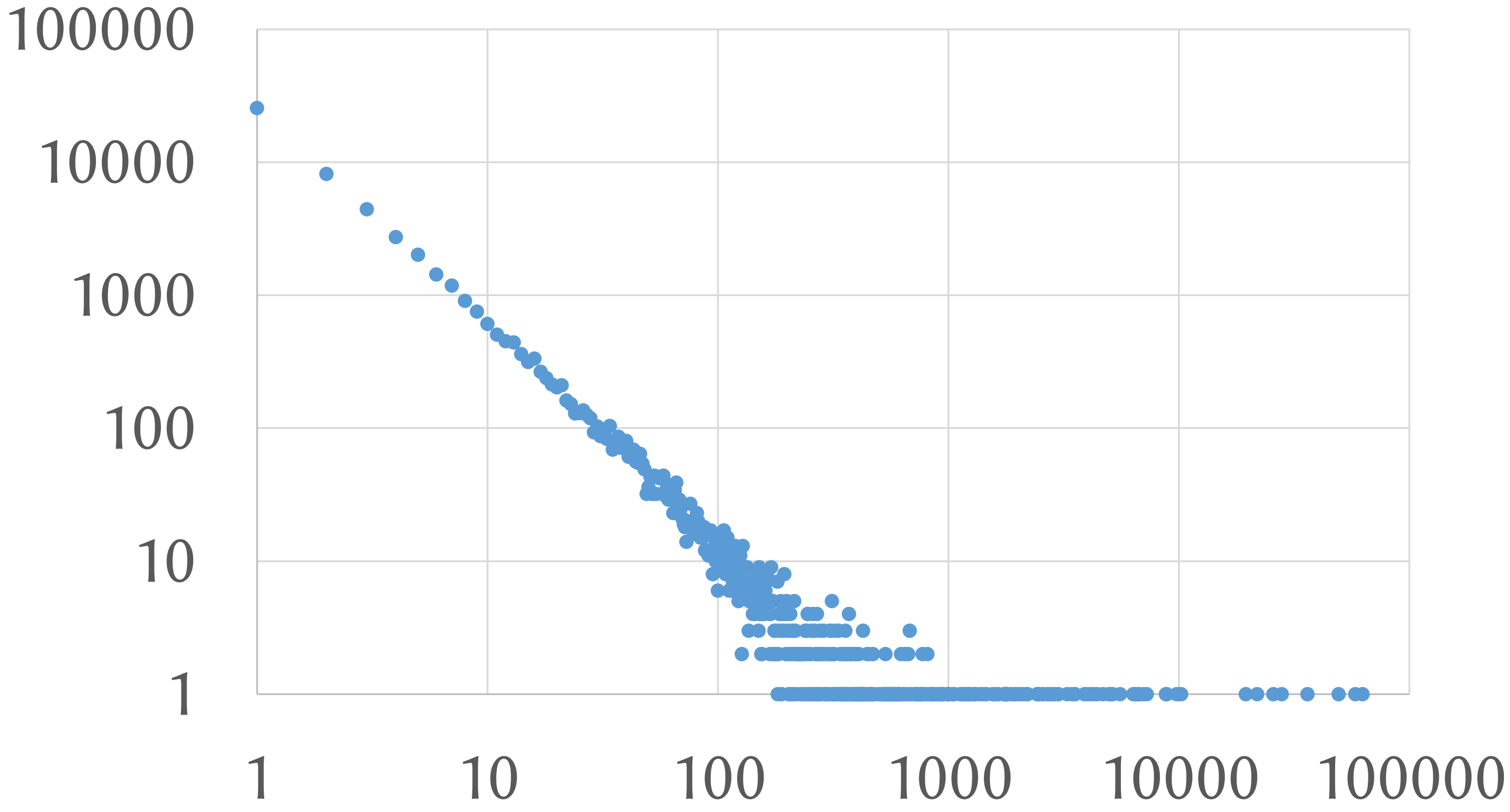
$$f(w) = \frac{C}{r(w)^a}$$

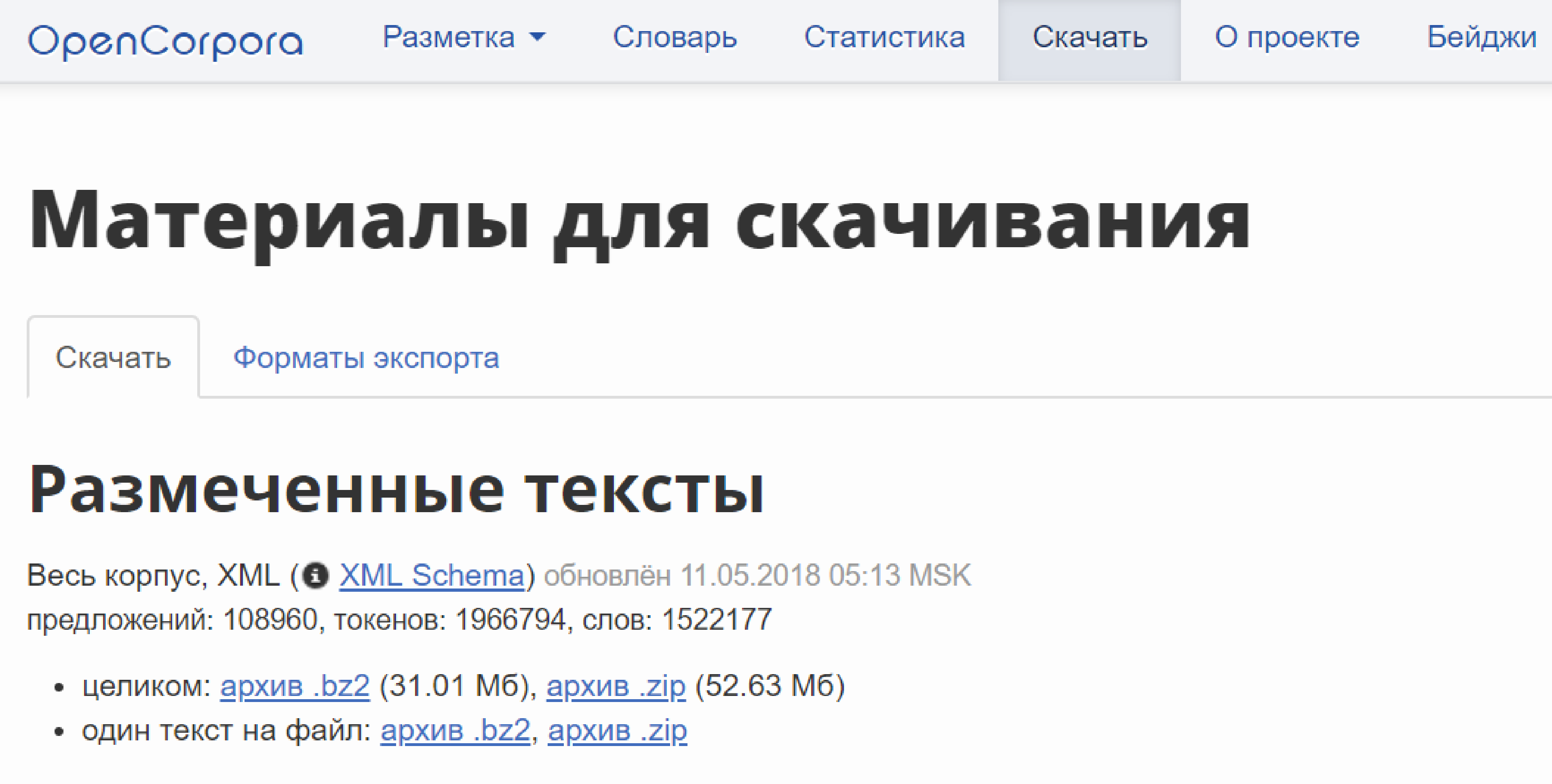
- Параметр a определяет форму графика (= стремительность падения частот)

Улучшения закона Ципфа

- Закон Ципфа — Мандельброта:
- $f(w) = \frac{C}{(r(w)+\beta)^{\alpha}}$
- Что содержательно привносит это изменение в формулу?

Частотный спектр Брауновского корпуса английского языка





Материалы для скачивания

Размеченные тексты

Весь корпус, XML (XML Schema) обновлён 11.05.2018 05:13 MSK
предложений: 108960, токенов: 1966794, слов: 1522177

- целиком: архив .bz2 (31.01 Мб), архив .zip (52.63 Мб)
- один текст на файл: архив .bz2, архив .zip

OpenCorpora

- Общая длина текстов: 1 600 904 слова
- 174 434 слова в словаре
- Самые частотные слова:
  - *в* 59987
  - *и* 53152
  - *на* 24952
- 91 885 слов, которые встретились 1 раз

20 слов, встретившихся 1 раз

- *дешевым*, *прапрадедах*, *лицемерием*, *латиноамериканскую*, *демократичностью*, *непонимающий*, *лицензионная*, *нашагал*, *исправительным*, *ржева*, *бэкграундов*, *межпланетная*, *мьеркуря*, *дуализме*, *помечены*, *айвазовский*, *посещавшей*, *свиных*, *энтомопатогенные*, *западноевропейская*

метана

- На 6364-м–6600-м местах в частотном списке — слово *метана* (29 вхождений, 18,1 ipm)
- В чём проблема?

метана

Строение молекулы **метана** имеет особенно важное значение для всей органической химии, так как оно связано с основными представлениями относительно углеродного атома, а применение новых методов исследования к изучению молекулы **метана** привело к весьма многообещающим результатам.

Метан — бесцветный газ, лишенный запаха; он обнаруживает небольшое уклонение от простых газовых законов и ожижается при —164° С. Он является наиболее важной составной частью природного газа: в некоторых местах природный газ обнаруживает до 99,3% **метана**. Со времени работ Пастера, Вант-Гоффа и Лебеля, в течение более чем 50 лет было общепринято представление о тетраэдрической структуре молекулы **метана**, причем предполагалось, что атом углерода находится в центре тетраэдра, а четыре водородных атома — по вершинам тетраэдра (рис. 1).

метана

- Все 29 вхождений слова *метана* — в тексте «Строение молекулы метана»
- Корпусные позиции: 1336470, 1336519, 1336546, 1336587, 1336608, 1336661, 1336686, 1336738, 1336812, 1336821, 1336854, 1336882, 1336891, 1337287, 1337373, 1337837, 1338047, 1338077, 1338188, 1338192, 1338261, 1338266, 1338297, 1338490, 1338567, 1338886, 1338908, 1338920, 1338933

Частотные словари: проблемы

- Неравная частотность в разных текстах
- Необходимы два вида мер:
  - меры разброса
  - усовершенствованные меры частотности

Меры разброса

- R (*range*) — количество сегментов корпуса, в которых встретилось слово, из n
    - У Ляшевской и Шарова (2009) $n = 100$
    - NB: одновременно и мера частотности

Стандартное отклонение

- Общепринятая статистическая мера разброса значений в выборке — **стандартное отклонение**:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_i - \bar{v})^2}$$

- $\bar{v}$ — среднее арифметическое частот в *n* сегментах
- Размер σ зависит от $\bar{v}$

Мера разброса

- *D* Жуйяна (Juilland’s D) = $100 \times \left(1 - \frac{\sigma}{\bar{v}\sqrt{n}}\right)$

Комбинации R и D

- Высокое R, высокое D — всё хорошо
- Низкое R, высокое D — низкочастотное слово
- Низкое R, низкое D — тематически окрашенное слово
- Высокое R, низкое D — слово есть во всех сегментах, но с очень разными частотами $\Longrightarrow$ общая частотность завышена за счёт выбросов

якорь

- *якорь* (Ляшевская–Шаров):
  25,9 ipm, *R* = 91; *D* = 28
- 6149 вхождений в НКРЯ / 283,4 млн = 21,7 ipm
- Но 1769 из них — в тексте Л. Н. Скрягин, «Книга о якорях» (1973)
- 1769 / 46 179 = 38 300 ipm;
  ср. *и*: 1405 / 46179 = 30 400 ipm
- (6149 – 1 769) / 283,4 млн = 15,5 ipm

Высокое R, низкое D

- Ищем другие слова с такими свойствами (напр., $R \geq 80$, $D \leq 40$)
  - *самец*
  - *самка*
  - *удлинённый*
  - *Крылов*
- Что можно про них сказать?

DP Гриса

1. Оценить, какую долю корпуса занимает каждая его часть (они могут быть разного размера!)
2. Предположить, что доля вхождений слова *x* в каждую часть от общего числа вхождений слова *x* будет соответствовать размеру этой части — **ожидаемые доли, E_i**

DP Гриса

3. Вычислить реальную долю вхождений слова *x* в каждую часть от общего числа вхождений *x* — **наблюдаемые доли**, $\boldsymbol{O_i}$
4. Просуммировать взятые по модулю разности между $\boldsymbol{O_i}$ и $\boldsymbol{E_i}$ и разделить на 2:

$$DP = \frac{\sum_{i=1}^{n} |O_i - E_i|}{2}$$

Усовершенствованные меры частотности

- Одновременно учитывают количество вхождений и разброс
- *R* (range) — тоже мера частотности и разброса, но грубая: слишком мало возможных значений

Усовершенствованные меры частотности

- U Жуйяна $= f \times D$
- Все перечисленные выше меры разброса и частотности основываются на заранее выделенных частях корпуса
- Можно выделять части корпуса динамически (для каждого слова)

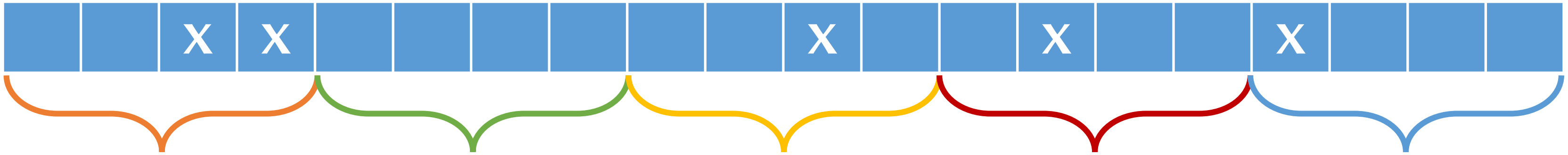
Определения

- Корпус: закольцованный список слов длины L (пронумерованный от 1 до L)
- f — число вхождений в корпус слова x
- n_i — позиция i-го вхождения x
- d_i — расстояние от (i − 1)-го вхождения до i-го вхождения x

Reduced Frequency

- Разделим корпус на f сегментов длины $v = L / f$
  (NB: R у Ляшевской и Шарова делит корпус на равное число сегментов для всех слов)
- Reduced Frequency (RF) — количество сегментов, содержащих слово w хотя бы раз

Reduced Frequency



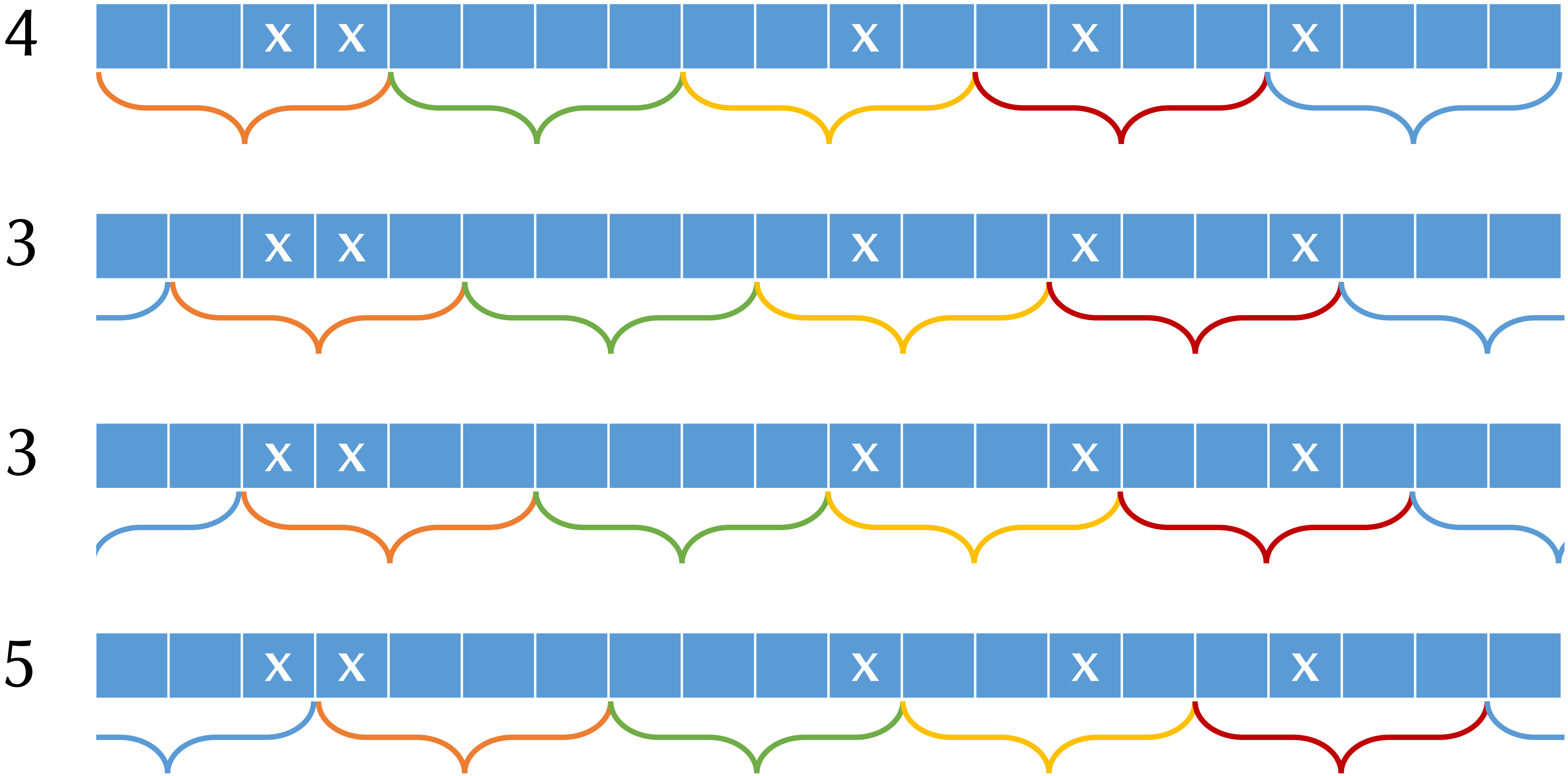
$L = 20, f = 5, v = 4$

$\mathrm{RF} = 4$

Average Reduced Frequency

- Разбиение на сегменты не обязано начинаться с первого слова ⟹ ⟹ надо брать усреднённую RF (Average Reduced Frequency, ARF) по всем разбиениям на сегменты
- Имеет смысл вычислять RF только для сегментов с 1-го до *v*-го, потому что, начав с (*v*+1)-го, получим такое же разбиение, как с 1-го, и т. д.
- $ARF = \frac{1}{v}\sum_{j=1}^{v} RF_j$

Average Reduced Frequency

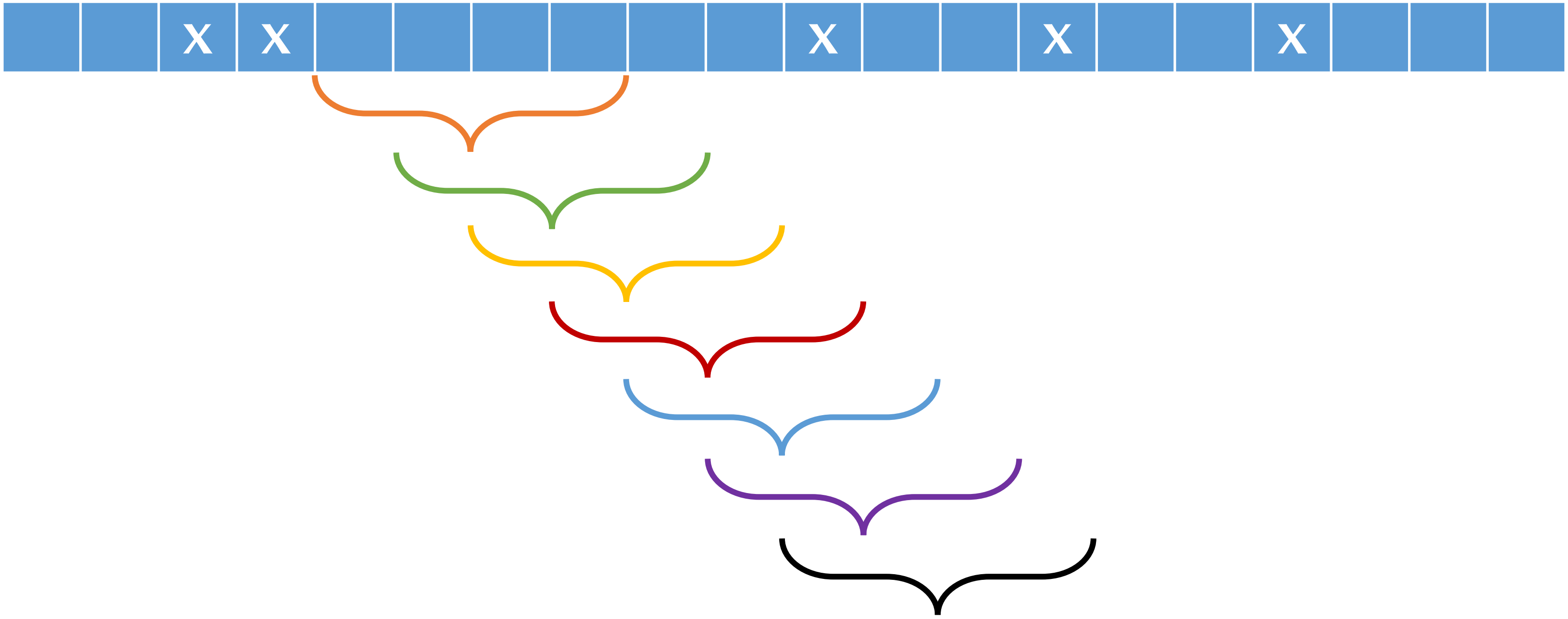


ARF = (4 + 3 + 3 + 5) / 4 = 3,75

Average Reduced Frequency

- В расчёте ARF участвует каждый сегмент от 1-го до L-го
- Можно просто перебрать их все, но очень долго
- Посмотрим на сегменты начиная с $(n_{i-1} + 1)$-го по n_i-й включительно.
- Сколько из них содержат слово x?

Average Reduced Frequency



Не больше v сегментов начиная с $(n_{i-1} + 1)$-го по n_i-й содержат слово x

Average Reduced Frequency

- $ARF = \frac{1}{v}\sum_{j=1}^{v} RF_j = \frac{1}{v}\sum_{i=1}^{f} \min(d_i, v)$

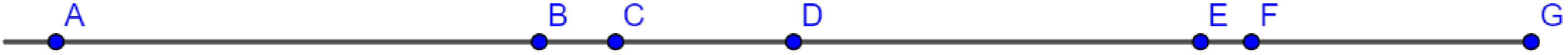
Москва, Голубинская улица, остановка «Универсам»

Москва, Голубинская улица, остановка «Универсам»

- 3 автобуса: 264, 281, 647
- Все идут к метро «Тёплый стан»
- Расписание (будни, 22:52–23:52):
    - 22:54 — 647
    - 23:13 — 281
    - 23:16 — 264
    - 23:23 — 647
    - 23:39 — 281
    - 23:41 — 264
    - 23:52 — 647

Интервалы

- d_i = 2; 19; 3; 7; 16; 2; 11
- Можно ли сказать, что средний интервал — 60 / 7 = 8,6 минут?
- Можно, но не отражает всю правду



Интервалы

- Если автобус ходит раз в n минут, каково математическое ожидание времени ожидания, если я прихожу на остановку в случайный момент?

Интервалы

- Если автобус ходит раз в n минут, каково математическое ожидание времени ожидания, если я прихожу на остановку в случайный момент?
- n / 2

Интервалы

- d_i = 2; 19; 3; 7; 16; 2; 11
- Какова вероятность прийти на остановку в *n*-минутном интервале?

Интервалы

- d_i = 2; 19; 3; 7; 16; 2; 11
- Какова вероятность прийти на остановку в n-минутном интервале?
- n / 60

Интервалы

- Среднее время ожидания t:

$$t = \sum \frac{d_i}{2} \cdot \frac{d_i}{L} = \frac{\sum d_i^2}{2L},$$

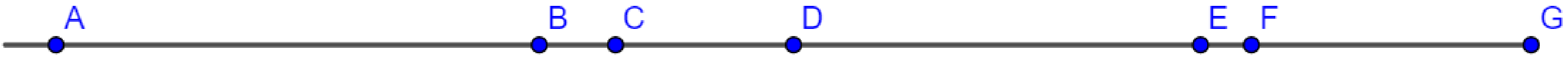
где L — общее время наблюдения

- Взвешенный интервал T и взвешенное количество объектов f_w:

$$T = 2t = \frac{\sum d_i^2}{L}, f_w = \frac{L}{T} = \frac{L^2}{\sum d_i^2}$$

Интервалы

- d_i = 2; 19; 3; 7; 16; 2; 11
- Вычисляем t, T и f_w



- $t = 6{,}7$
- $T = 13{,}4$
- $f_w \approx 4{,}5$
- «На самом деле» — 4,5 автобуса, а не 7

метана

- Абсолютная частота — 29 вхождений (6364-е–6600-е места в частотном списке)
- Взвешенное количество объектов — 1,003086 (72323-е место в частотном списке)
- NB: заодно мы почти избавляемся от разделённых мест

А. С. Пушкин

Я вас любил: любовь еще, быть может,
В душе моей угасла не совсем;
Но пусть она вас больше не тревожит;
Я не хочу печалить вас ничем.
Я вас любил безмолвно, безнадежно,
То робостью, то ревностью томим;
Я вас любил так искренно, так нежно,
Как дай вам бог любимой быть другим.

А. С. Пушкин: частотный словарь

вас	4,21	*безмолвно*	1
я	3,42	*безнадежно*	1
любил	2,57	*бог*	1
не	1,51	*больше*	1
быть	1,31	*в*	1
то	1,08	…	
так	1,08		

- Если посчитать взвешенное количество объектов, то слово *любил* встречается чаще, чем слово *не*, а слово *быть* — чаще, чем слова *то* и *как*!

Ключевые слова в корпусной лингвистике

- Что такое ключевые слова?
- Ключевые слова — это слова, которые встречаются в некотором корпусе с необычно высокой частотой

Автоматическое извлечение ключевых слов

- Взять фокусный корпус (интересный для нас) и референсный корпус (основание для сравнения) и сравнить частоты всех слов
- Для каждого слову вычислить некоторую меру ключёвости
- Отсортировать слова по ключёвости и проинтерпретировать верхнюю часть списка

Простейшая мера ключёвости

- $K(w) = \frac{f_{w,foc}}{f_{w,ref}}$
- На вход подаются относительные частоты
- Измеряется только размер эффекта, но не статистическая значимость

Размер эффекта и статистическая значимость

- **Размер эффекта:** насколько частотнее слово в фокусном корпусе, чем в референсном?
- **Статистическая значимость:** можно ли доверять этому результату?

Усовершенствованная мера ключёвости

$$K(w) = \frac{f_{w,foc} + n}{f_{w,ref} + n}$$

- n — параметр сглаживания
- $n > 0$
- n не даёт делить на 0 и определяет наши предпочтения: размер эффекта (редкие слова) или статистическая значимость (частотные слова)

Влияние *n*: пример

			$n = 0$		$n = 10$		$n = 500$	
	Фокус	Референс	K	Ранг	K	Ранг	K	Ранг
w_1	10	1	10	1	1,818	3	1,018	3
w_2	200	40	5	2	4,200	1	1,296	2
w_3	10000	5000	2	3	1,998	2	1,909	1

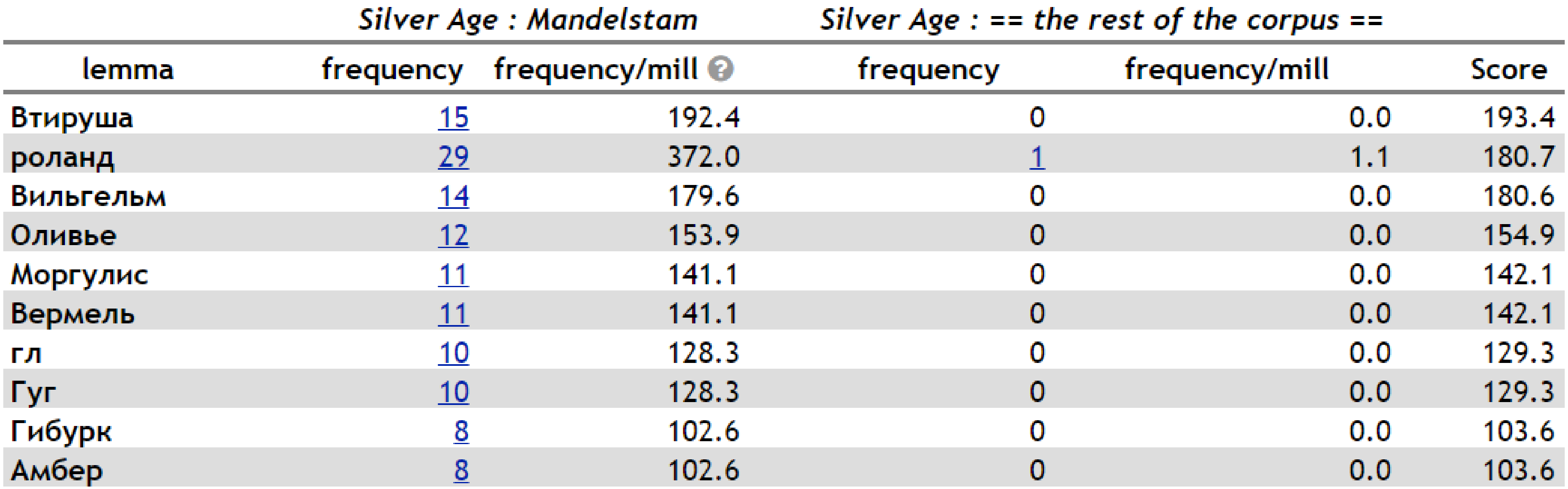
Skecth Engine: автоматическое извлечение ключевых слов

- the.sketchengine.co.uk
- Можно извлекать ключевые слова как из существующих корпусов, так и из пользовательских корпусов

Серебряный век

Поэт	Размер корпуса	Размер корпуса (%)
Анна Ахматова	53 043	7,27
Александр Блок	107 677	14,76
Николай Гумилёв	61 717	8,46
Сергей Есенин	54 881	7,52
Осип Мандельштам	55 860	7,66
Владимир Маяковский	152 209	20,86
Борис Пастернак	72 356	9,92
Марина Цветаева	171 801	23,55
Total	729 544	100,00

Ключевые слова для Мандельштама



	Silver Age : Mandelstam		*Silver Age : == the rest of the corpus ==*		
lemma	frequency	frequency/mill	frequency	frequency/mill	Score
Втируша	15	192.4	0	0.0	193.4
роланд	29	372.0	1	1.1	180.7
Вильгельм	14	179.6	0	0.0	180.6
Оливье	12	153.9	0	0.0	154.9
Моргулис	11	141.1	0	0.0	142.1
Вермель	11	141.1	0	0.0	142.1
гл	10	128.3	0	0.0	129.3
Гуг	10	128.3	0	0.0	129.3
Гибурк	8	102.6	0	0.0	103.6
Амбер	8	102.6	0	0.0	103.6

- При $n = 1$ и без усовершенствованных мер частотности получается плохо

1. **О. Э. Мандельштам. Паломничество Карла великого в Иерусалим и Константинополь (1921-1929)** [омонимия не снята] **Все примеры (15)**

В сводчатом зале в мраморном столбе
В головах у пэров **Втируша** сел,
В скважину за ними всю ночь глядел.
[О. Э. Мандельштам. Паломничество Карла великого в Иерусалим и Константинополь (1921-1929)] [омонимия не снята] ←…
→

«Клянусь Богом, — говорит **Втируша**, — вы могучи и крепко сложены.
[О. Э. Мандельштам. Паломничество Карла великого в Иерусалим и Константинополь (1921-1929)] [омонимия не снята] ←…
→

Извлечение ключевых слов

- $n = 100$
- Используем не обычную частоту, а ARF

Анна Ахматова

словно, *разлука*, *голос*, *Нева*, *липа*, *я*, *уже*, *память*, *казаться*, *А.*, *сад*, *снова*, *оттого*, *горький*, *страшный*, *почти*, *смертный*, *встреча*, *черный*, *слава*, *таинственный*, *зеркало*, *Н.*, *вспоминать*, *душистый*, *даже*, *стать*, *всегда*, *смерть*, *сразу*, *сделаться*, *душный*, *О.*, *отчего*, *тень*, *дом*, *случиться*, *навсегда*, *со*, *прощаться*, *первый*, *последний*, *тайный*, *муза*, *струиться*, *тогда*, *мой*, *бродить*, *иль*, *присниться*

Александр Блок

мечта, *мгла*, *ночной*, *черта*, *вечерний*, *печальный*, *весна*, *темный*, *бледный*, *даль*, *светлый*, *сумрак*, *дальний*, *сон*, *вдали*, *туман*, *страсть*, *безумный*, *холодный*, *дума*, *тишина*, *печаль*, *туманный*, *взор*, *око*, *мрак*, *душа*, *звук*, *огонь*, *пройти*, *смотреть*, *страстный*, *долгий*, *цветок*, *больной*, *ясный*, *путь*, *твой*, *песня*, *там*, *дева*, *глубина*, *близкий*, *тайна*, *сонный*, *луч*, *ночь*, *ты*, *бродить*, *прекрасный*

Николай Гумилёв

взор, пальма, пред, могучий, иль, золотой, бледный, веселый, огненный, неведомый, слон, утес, озеро, девушка, печальный, странный, словно, лев, пустыня, птица, зверь, храм, странно, луна, цветок, он, средь, древний, священный, воин, дракон, но, страна, всегда, мечта, равнина, сладкий, гордый, мечтать, девичий, страшный, дева, рай, море, пылать, светлый, ужас, дивный, дикий, нежный

Осип Мандельштам

воздух, могучий, прозрачный, огромный, холм, князь, нежный, улица, пространство, тяжелый, вода, немного, должный, играть, рыба, сильный, тяжесть, голова, деревянный, крупный, ласковый, старинный, легкий, шуметь, зеленый, хмель, француз, народ, железный, теплый, получить, колючий, ласточка, морской, говорить, язык, дышать, высокий, прекрасный, зрачок, большой, черствый, римский, влажный, хрустеть, когда, каменный, ярость, дремучий, кровь

Владимир Маяковский

мол, работа, очень, газета, марш, хвост, пушка, сидеть, нос, тысяча, человечий, Ленин, пуля, ухо, точка, морда, совет, двадцать, зря, штык, сто, это, переть, работать, чтоб, еле, гражданин, пойти, разный, дыра, купить, труд, ус, дело, я., сразу, мозг, лава, никакой, флаг, черт, этаж, просто, Москва, рост, даже, голод, сегодня, прочий, деньги

Борис Пастернак

дерево, двор, снег, дождь, зима, облако, даль, лед, кровля, ливень, сумерки, стужа, крыша, как, точно, сырой, улица, ветка, рваться, толпа, туча, стекло, поезд, капля, утро, воздух, пред, об, ползти, овраг, пахнуть, детство, со, окно, пруд, они, вне, прибой, что-то, жаркий, ветвь, рассвет, обрывок, ненастье, вьюга, гул, вихрь, кисть, дверь, эхо

Марина Цветаева

уж, лоб, да, царь, вдоль, вздох, сын, сей, спать, коль, уста, грудь, две, плащ, бог, перст, жила, колыбель, целый, мать, вслед, царство, маленький, господь, женский, око, мама, добрый, рука, ровно, два, без, глядеть, бровь, кроме, дитя, брат, грех, Русь, слеза, ж, правый, древо, так, ресница, взмах, левый, муж, эх, царский

Сергей Есенин

Русь, поле, рожь, луна, синий, месяц, голубой, потому, песня, веселый, родимый, грусть, мужик, родной, луг, роща, береза, словно, осенний, теперь, желтый, корова, петь, плетень, изба, синь, село, край, радость, родина, ах, отчий, уж, березка, чувство, звенеть, мать, нужно, светить, метель, недаром, ведь, про, петух, я, степь, туман, заря, клен, овес

	Сущ.	Прил.	Наречия	Глаголы	Союзы	Междом	Частицы	Предл.	Мест.	Им. соб.	Числ.	Σ
Ахматова	12 24%	10 20%	7 14%	9 18%	2 4%	0 0%	3 6%	1 2%	2 4%	4 8%	0 0%	50 100%
Блок	26 52%	17 34%	2 4%	3 6%	0 0%	0 0%	0 0%	0 0%	2 4%	0 0%	0 0%	50 100%
Есенин	28 56%	8 16%	5 10%	3 6%	1 2%	1 2%	1 2%	1 2%	1 2%	1 2%	0 0%	50 100%
Гумилёв	22 44%	18 36%	2 4%	2 4%	3 6%	0 0%	0 0%	2 4%	1 2%	0 0%	0 0%	50 100%
Мандельштам	17 34%	25 50%	1 2%	6 12%	1 2%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%	50 100%
Маяковский	25 50%	3 6%	6 12%	5 10%	1 2%	0 0%	2 4%	0 0%	3 6%	2 4%	3 6%	50 100%
Пастернак	37 74%	2 4%	0 0%	3 6%	2 4%	0 0%	0 0%	4 8%	2 4%	0 0%	0 0%	50 100%
Цветаева	26 52%	7 14%	3 6%	2 4%	1 2%	2 4%	2 4%	3 6%	1 2%	1 2%	2 4%	50 100%

Спасибо за внимание!