Частотные списки. Извлечение ключевых слов

Частотные списки

Частотный словарь служит источником информации о том, какие слова употребимы в языке в большей степени, а какие — в меньшей.

Он содержит списки слов, при которых указывается, с какой частотой они встречаются в текстах. Для того, чтобы этот показатель был более достоверным, частота слова подсчитывается на основе большого корпуса текстов.

Частотные списки

Самый известный **современный** русский частотный словарь — это *словарь Ольги Ляшевской и Сергея Шарова*, который вышел в 2009 году и базируется на *корпусе текстов объемом 92 млн словоформ*.

Этот словарь находится в свободном доступе в интернете, любой может им воспользоваться. http://dict.ruslang.ru/freq.php

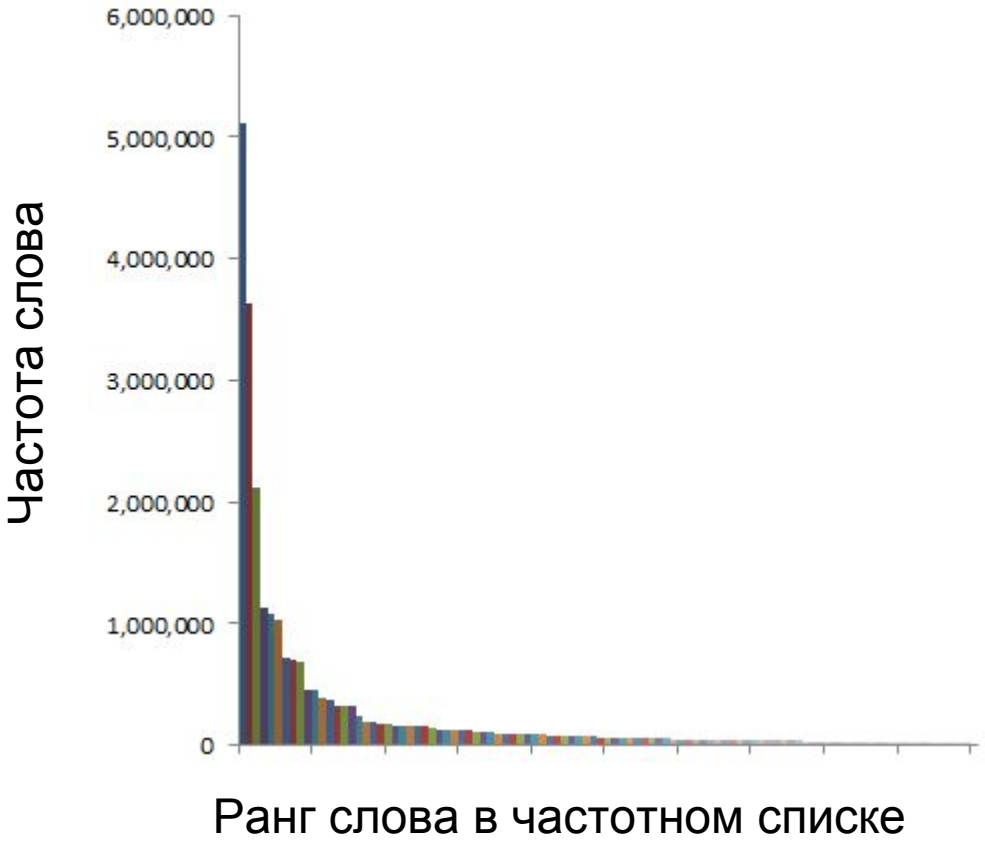
Как устроены частоты в словаре?

Закон Ципфа (Zipf’s Law)

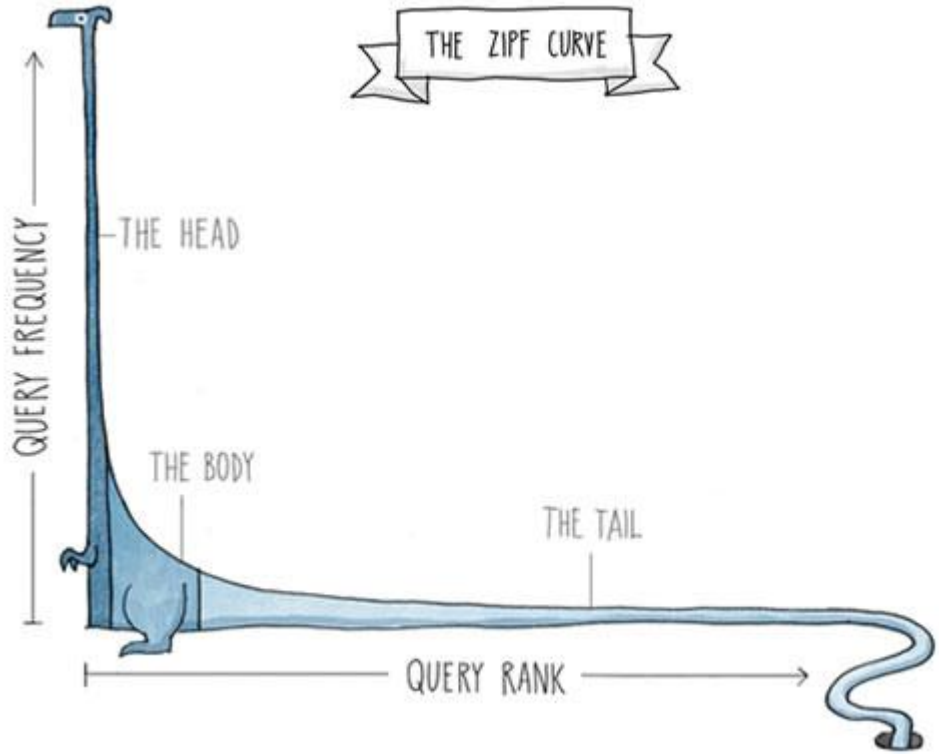
Закон Ципфа гласит следующее: Если все слова языка (или просто относительно длинного текста) упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется *приблизительно* обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова).

Например, второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и так далее.

Закон Ципфа (Zipf’s Law)



Закон Ципфа (Zipf’s Law)



Абсолютная vs. относительная частота

Абсолютная vs. относительная частота

Текст 1
политика 10
экономика 32

Текст 2
политика 50
экономика 35

Абсолютная vs. относительная частота

Текст 1 (500 слов)
политика 10 = 0,02
экономика 32 = 0.064

Текст 2 (500 слов)
политика 50 = 0,1
экономика 35 = 0,07

Абсолютная vs. относительная частота

Текст 1 (500 слов)
политика 10 = 0,02
экономика 32 = **0.064**

Текст 2 (500 слов)
политика 50 = 0,1
экономика 35 = **0,07**

В тексте 2 больше политики, а экономики почти поровну.

Абсолютная vs. относительная частота

Текст 1 (200 слов)
политика 10 = 0,05
экономика 32 = 0,16

Текст 2 (1000 слов)
политика 50 = 0,05
экономика 35 = 0,035

Абсолютная vs. относительная частота

Текст 1 (200 слов)
политика 10 = **0,05**
экономика 32 = 0,16

Текст 2 (1000 слов)
политика 50 = **0,05**
экономика 35 = 0,035

В тексте 1 больше экономики, а политики поровну.

Mepa TF*IDF

Определение

TF-IDF — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов.

Где используется:
в задачах анализа текстов;

в информационном поиске (например, это один из критериев релевантности

документа поисковому запросу);

при расчёте меры близости документов при их классификации

Применение

Три тематики:

Провоз багажа

Возврат билета

Изменение бронирования

В каком столбце какая?

билет, 0.585488	билет, 0.671791	багаж, 0.549997
москва, 0.236778	поменять, 0.224687	килограмм, 0.487409
номер, 0.19588	москва, 0.217878	билет, 0.302772
ноль, 0.18727	дата, 0.1566	сумка, 0.161166
документ, 0.172202	ноль, 0.142983	большой, 0.148648
возврат, 0.16144	сентябрь, 0.136174	москва, 0.143954
справка, 0.150677	сделать, 0.131635	чемодан, 0.129089
больничный, 0.146372	паспорт, 0.118017	место, 0.126742
почта, 0.133457	номер, 0.113478	вопрос, 0.118918
адрес, 0.126999	пассажир, 0.113478	сайт, 0.106401
электронный, 0.120542	добрый, 0.10667	салон, 0.106401
пассажир, 0.109779	день, 0.0998608	ручной, 0.102489
сделать, 0.109779	бронирование, 0.0975913	добрый, 0.098577
рейс, 0.107626	время, 0.0975913	тысяча, 0.0962299
быть, 0.105474	большой, 0.0930521	взять, 0.0891887
деньги, 0.0990164	тысяча, 0.0907826	ребёнок, 0.0868417
паспорт, 0.0947113	телефон, 0.0862434	быть, 0.0813651
бронирование, 0.0904	вылет, 0.0862434	рюкзак, 0.0672827

TF (term frequency)

Параметр TF (term frequency) — это отношение числа раз *k*, которое некоторое слово встретилось в документе, к общему числу слов в документе *n*. Нормализация длиной документа нужна для того, чтобы уравнять в правах короткие и длинные документы.

$TF = k / n$

IDF (inverse document frequency)

Некоторые слова могут **встречаться почти во всех документах коллекции** и, соответственно, оказывать **малое влияние** на принадлежность документа к той или иной категории, а значит, **не быть ключевыми** для этого документа.

Для понижения значимости слов, которые встречаются почти во всех документах, вводят инвертированную частоту термина IDF (inverse document frequency).

IDF (inverse document frequency)

IDF (inverse document frequency) — это натуральный логарифм отношения числа всех документов D к числу документов d, содержащих некоторое слово.

$$IDF = \log \frac{D}{d}$$

Мера TF-IDF

Мера TF*IDF равна произведению TF и IDF, при этом TF играет роль повышающего множителя, IDF — понижающего. Тогда весовыми параметрами векторной модели некоторого документа можно принять значения меры TF*IDF входящих в него слов.

Пример

Пусть коллекция состоит из 3 документов.

1. Мама мыла мылом Машу.

2. Мама мыла, мыла раму.

3. В магазине купила мама мыло.

Задание: посчитать TF*IDF

https://docs.google.com/spreadsheets/d/1r24LB3RFv_lThIIisI6gyLCrhwJZkWr0aMNTdTvxZ8M/edit?usp=sharing

Пример

Слово	Частота в коллекции	IDF
мама	3	0
мыть	2	0.18
мыло	2	0.18
Маша	1	0.47
рама	1	0.47
магазин	1	0.47
купить	1	0.47
в	1	0.47

Коллекция из 3-х документов

1. Мама мыла мылом Машу.
2. Мама мыла, мыла раму.
3. В магазине купила мама мыло.

Алгоритм RAKE

RAKE (Rapid Automatic Keyword Extraction)
Основная идея: среди ключевых слов не бывает стоп-слов.

1. Разбиваем документ по границам слов (пробелы, пунктуация).
   “A scoop of ice cream.” > ["A", "scoop", "of", "ice", "cream"]
2. Убираем по стоп-слова.
   “A scoop of ice cream.” > ["scoop", "ice cream"]
3. Считаем RAKE score: degree(word)/frequency(word)

Алгоритм RAKE

RAKE score: degree(word)/frequency(word)
Frequency — абсолютная частота слова
Degree — встречаемость слова вместе с другими словами (в рамках candidate keywords).

Compatibility – systems – linear constraints – **set** – natural numbers – Criteria – compatibility – system – linear Diophantine equations – strict inequations – nonstrict inequations – Upper bounds – components – **minimal set** – solutions – algorithms – minimal generating sets – solutions – systems – criteria – corresponding algorithms – constructing – **minimal supporting set** – solving – systems – systems

degree("set") = 6 (bold)
degree("natural") = 2 (underlined)