Домашнее задание №1

Алиев Даниял

Задача 1

Соберем статистику по самым частым словам. Чтобы собрать статистику по количеству документов *count*(), где фигурируют эти слова, воспользуемся данными с сайта https://ruscorpora.ru/new/corpora-freq.html.

w	count(w)	частота	google-count(w)	R
человек	35.143	0.2769	575.000.000	2.076.561.938
слово	16.662	0.13	120.000.000	916.030.534
место	25.662	0.2	429.000.000	2.145.000.000
жизнь	30.903	0.24	486.000.000	2.025.000.000
время	59.969	0.47	2.520.000.000	5.361.702.127
хорошо	27.657	0.22	471.000.000	2.140.909.090
сердце	36.560	0.29	168.000.000	579.310.345

частота = $\frac{count(w)}{N}$, N = 126.904 - общее число документов в национальном корпусе РЯ
google-count(w) - число результатов по запросу w
R = $\frac{google-count(w)}{\text{частота}}$
$\overline{R} = \frac{\sum_w R(w)}{num(w)} = 2.177.787.719$

Единственное число проиндексированных страниц в гугле, которое мне удалось найти - 130 триллионов, не уверен в корректности этого числа, но, пожалуй, остановлюсь на нем.

Тогда получаем, $\frac{2.177.787.719}{130.000.000.000.000} = 1.6\%$ - процент интернета, занимаемый Рунетом. Результат не очень похож на правду, но и оценка проводилась достаточно грубо.

Задача 2

По оценкам Google имеет около 130 триллионов проиндексированных страниц. Также считается, что Google покрывает лишь 3% всего интернета. Этим данным не стоит верить на слово, так как для индексации Google использует краулеры, которые работают недетерминировано. Также невозможно дать точную оценку объема Deep/DarkNet и страниц, не поддающимся индексации.

Задача 3

Одним из способов проверки языковой статистики может являться социологический опрос, а также сравнение со статистикой по корпусам, подходящим к данной задаче. Также полезно понимать, какое распределение по времени имеют результаты выдачи поисковика, чтобы правильно подобрать подходящий корпус.

Задача 4

Пользуясь все тем же национальным корпусом РЯ, соберем частоту употребления "ноль"и "нуль"

	основной	газетный	обучающий	поэтический	устный
общее число слов	337 025 184	765 546 444	664 804	12 935 027	13 399 937
нуль / общее число	582	102	3	60	29
ноль / общее число	1 255	5 089	2	97	370
нуль / частота	0.00000173	0.00000013	0.00000045	0.00000464	0.00000216
ноль / частота	0.00000372	0.00000673	0.0000003	0.0000075	0.0000276
ноль/нуль (частота)	2.15	51.7	0.66	1.6	12.8

Анализ:

Видим, что чаще всего нуль употребляется в обучающем и поэтичсеком корпусе, что логично, ведь слово нуль раньше использовалось гораздо чаще, а значит, и в поэзии оно употреблялось чаще. К тому же в технической литературе очень часто предпочитают использовать именно нуль, а не ноль.

В газетном корпусе употребляется более актуальный на сегодняшний день лексикон, потому и частота употребления слова нуль в нем минимальна.

Отношение частот употребления в Google Ngrams за последние года составляется приблизительно 2, что ближе всего к поэтическому корпусу, что логично, учитывая что частота в нем вычисляется по литературным произведениям.

Задача 5

При использовании Google Ngrams нужно понимать этимологию слов - их значение со временем может сильно меняться.

Например, для современных школьников может быть удивительным тот факт, что частота использования "fortnite"в данный момент составляет 0.0000001927, в то время как максимум достигался в 1808 году и составлял 0.0000003789.

Дело в том, что до того, как это название присвоила себе одна из самых популярных видеоигр в данный момент, слово "fortnite"означало промежуток времени - две недели.

Поэтому при подобном анализе лучше быть бдительным, чтобы не прийти к выводу, что в 19 веке эта видеоигра была популярнее, чем сейчас :)