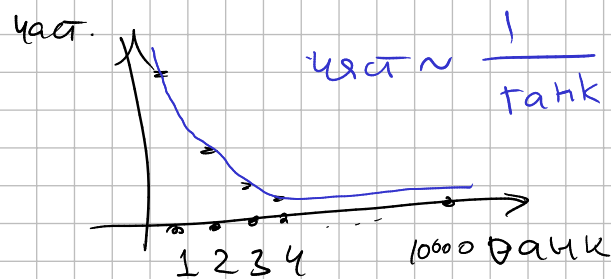


ОС работе ЕЯ.

- всего 3 решения при законе ципера.
- графики

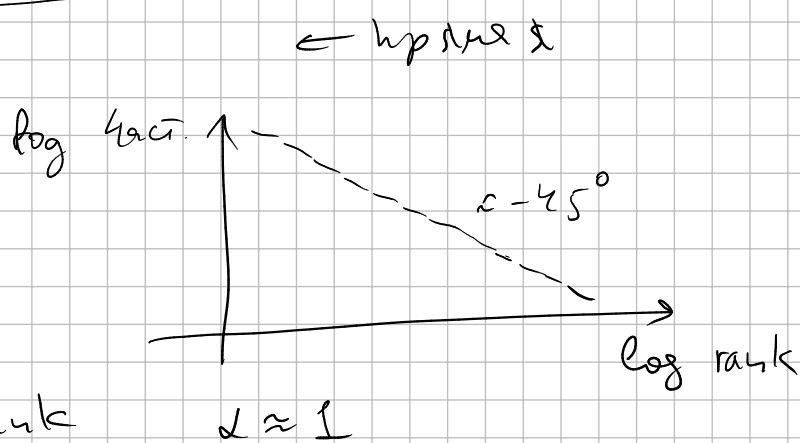
$$\text{частота} \approx \frac{K}{\text{rank}}$$



$$\underbrace{\log \text{частота}}_y \approx \underbrace{\log K - \log \text{rank}}_{k - 1 \cdot x}$$

Уточненный закон

$$\log \text{freq} \approx k - \alpha \log \text{rank}$$



е.х.

$$\text{freq} \approx \frac{K}{\text{rank}^\alpha}$$

иногда берут $\alpha \approx 0.95$

Эон пуниса К задает о законе ципера

- использование numpy.

numpy работает с матрицами (наборами чисел) и

выполняет сложные операции сразу со всеми числами, т.е. векторные операции.

```
import numpy as np  
a = np.array([100, 50, 25])  
# массив
```

- тоже массив, но numpy массив

$b = \text{np.log}(a)$ - поэлементное применение операции.

$\log 100, \log 50, \log 25$

plot умеет работать с numpy массивами

(пример a - numpy массив $1, 2, 3$
 b - $10, 20, 30$
 $a + b$ - $11, 22, 33$)

- использовать pyplot.

т.ч. прописывать раз.

два графика. 1) Исходные формы сплав

2) Начальные формы

" \rightarrow Выход кот кот кот кот \rightarrow

1) $1 \quad 1 \quad 1 \quad 1 \quad 1$

2) $1 \quad 1 \quad 3$

- regression d.

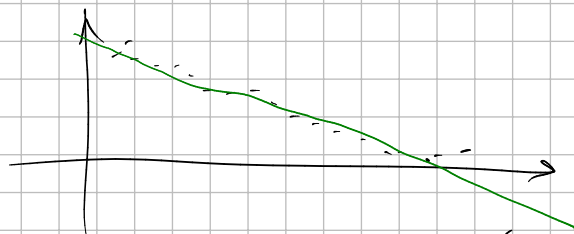


график сам проведёт прямую через (*)

это задача линейной регрессии.

Понятие. Машинное обучение.

пример 1 \rightarrow прогноз 1

пример 2 \rightarrow прогноз 2

пример n \rightarrow прогноз n

пример \rightarrow ? - сама предсказывает

} обучает компьютер

если предсказывает α -т конечного множества
{сам, не сам} {кот, собака, дельфин}
— классификация

если предсказывает число — это
регрессия.

Здесь мы на ранке предсказывает число.

$$freg = \underbrace{\alpha \cdot rank + k}_{\text{лин. ф-ция}} + \text{error}.$$

$$\text{error}_1^2 + \dots + \text{error}_n^2 \rightarrow \min, \text{ где } \text{error}_i$$

— ошибка на примере i

α и k ищуте методом лин. регрессии.

мы используем векторный расчет. (см. отсюда)

Поиск коллокации

Некоторые фразы (группы слов) встречаются
чаще других. New York York без New
Йорк не встречается. фразеологизм: не зря
крепкий чай > сильный чай
— редкость

Надо найти в тексте пары слов, которые
встречаются вместе чаще, чем если бы эти
слова были случайными.

w_1, w_2 — два слова подряд.

образуют ли они коллокацию?

исчислять все пары слов в тексте. w_1, w_2 . Присмотримся

$u_1 \quad w_1 \quad w_2 \quad u_4 \quad w_2 \quad u_6 \quad u_7 \quad w_1 \quad w_2 \quad u_{10}$

	w_2	$\neg w_2$	
w_1	2	0	$\bar{z} = 2$
$\neg w_1$	1	6	
	$z = 3$		$\bar{z} = 9$

9 пар слов

w_1, w_2 - в среднем
1 слово w_1 ,
2 слова w_2

$w_1, \neg w_2$ -
1 слово w_1 ,
2 слова $\neg w_2$

У всех чисел достаточно, чтобы оценить,

является ли w_1, w_2 коллокацией.

Примеры

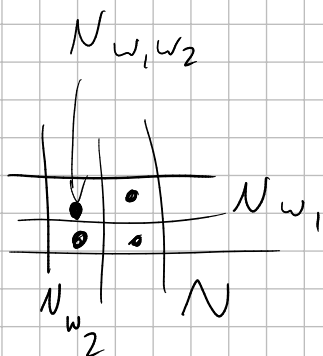
	w_{York}	$\neg w_{\text{York}}$
w_{New}	100	200
$\neg w_{\text{New}}$	0	100000

- колл.

	w_2	$\neg w_2$
w_1	1	20
$\neg w_1$	20	10000

- не колл

$$\frac{0,002 \cdot 0,002^*}{2 \cdot 10^{-6}}$$



составляем оценку, насколько
это похоже на коллокацию

очень много формул для подсчета оценки

пример 1:
$$\frac{2 \cdot N_{w_1, w_2}}{N_{w_1} + N_{w_2}}$$

- MI

↓ мере. ...

Для каждой колонки выбирается мера, и все меры слов оцениваются и выбираются только-то мер с тех мерой.

Как искать колонки.

- токенизируем текст (символы рус. текст)
- приводим или не приводим к нач ф-ме. (2 вар. действий)
- отбрашиваем редкие слова.
- отбрашиваем частые слова (stop-слова)
- выбираем меру Dice, MI ... (3 шт)
- отбираем 100 лучших мер по выбранной мере и записываем в файл вместе с мерой.

файл CSV - comma separated value
кренки; каф; 11.57
смысли; ответ; 10.33
слово; слово; мера.

(pandas в числ. раз)