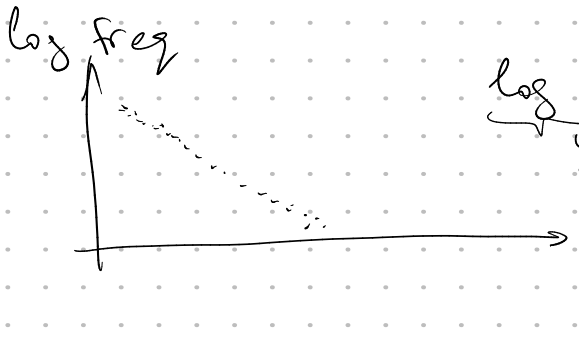
в Zipf



$\underbrace{\log freq}_{y} = C - \underbrace{\log rank}_{x}$

1) попробуйте "нормализовать" слова.

Кот
Кота
котом
... → кот

1 слово

норм. форма

→ лемма (сущ → им, ед.ч
гл → инфин.
...

→ стем (отрезать окончание)

??→ Сталь (сущ)
Стали → стал (гл)

playing → play
player → play
dancing → dance (e)
...

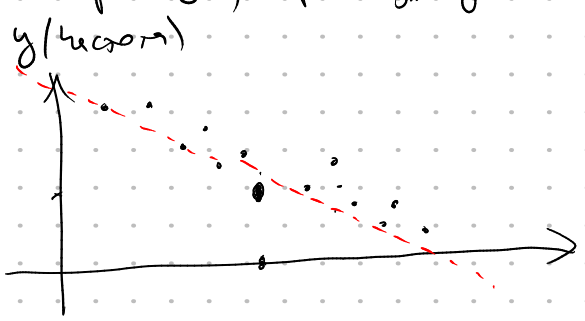
Англ. Porter Stemmer
– алг, правила отрезания окончаний.

Доп задание I
Нормализуем через морф. анализатор
- берём первый в-т анализа.

pymorphy 2
– для слова выдаёт варианты нач. форм, если не знает – угадывает.

Доп. задание II

Попытаемся подобрать прямую, проходящую ближе всего к нашим точкам.



Задача: Давайте пытаться предсказывать частоту по рангу

Есть примеры (обучающие)

	ранк	часов
много	100	10
	20	50
	120	?

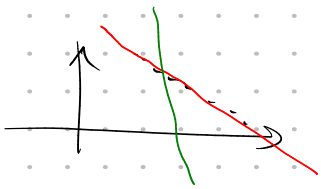
← предсказать

Задача предсказания непрерывной величины — задача регрессии

(см. классификация)

Линейная регрессия предполагает:

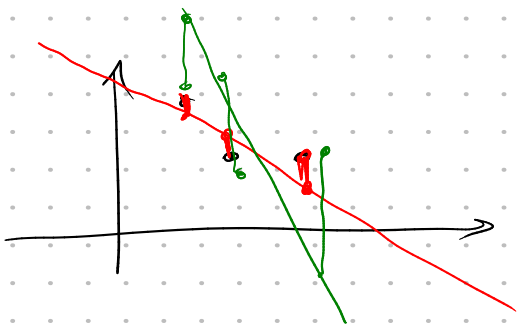
$$\underbrace{Y}_{\text{надо предсказать}} = b + k\underbrace{X}_{\text{дано}} + Err, \quad Err \sim N(0, \sigma)$$



Научиться предсказывать = подобрать b и k.

При подборе делаем так, чтоб $\sum_{i=1}^{N} err_i^2 \to min$

сумма по примерам.



оч. зел >> оч. красн

Как подобрать прямую?

I. numpy

Данные с числами лучше хранить в numpy-матрицах.

numpy — создание и работа с матрицами из чисел.

значительно эффективнее и по памяти, и по времени, по сравнению со списками и циклами.

```
import numpy as np

x = np.array([10, 20, 30])
y = np.array([[10, 20], [30, 40]])
```

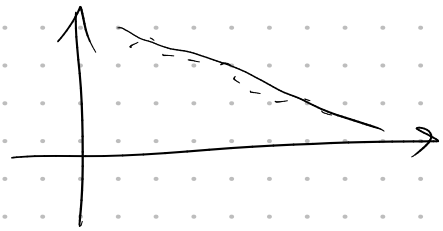
10	20	30

10	20
30	40

обучаем модель

X	y
0 7	.
1 2	.
1.5 1	.
2 3	.

в Zipf



проверяем, что получилось ≈ -1

$$\log freq = C - \log rank$$

$$\log freq = b + k \cdot \log rank$$

$$k = -1$$

$$freq = \frac{C}{rank^{a=1}}$$

N - грамм модели.

Строим модель языка.

] язык - это множество предложений.

Например, в русском "я иду снег". $\in$ Русский

"я иду насос". $\notin$ Русский

"мы стали более лучшее одеваться" $\notin$ Русский

Лучше иначе:

$P(\text{предложение}) \longrightarrow$ Число от 0 до 1 $[0, 1]$

вероятность встретить предлож.

$P(\text{"я иду снег"}) >>> P(\text{"я иду насос"})$

Для вероятностной модели важно $\sum_{\text{предл}} P(\text{предл}) = 1$

→ это вероятностная модель языка.

Модель языка (P : предл → числа)

зачем

— при генерации текста, выбор нужного варианта.

I sit by the table → я сижу у стола (1)
я сижу у таблицы (2)

$P(1) > P(2)$

выберем 1ый вариант.

— поиск ошибок. я глажу кита (1) — было
я глажу кота (2) — н.б.

$P(1) < P(2)$ ⇒ н.б. вы имели в виду "кот"

У нас будет N-грамм модель.

$P(w_1\ w_2\ w_3\ w_4 \dots w_n) =$ Гипотеза: след слово зависит от N-1 предыдущих.

слово

$N=2$

$= P($ начало $\langle s\rangle\ w_1\ w_2\ w_3 \dots w_4\ \langle/s\rangle$ конец $) :=$

$= \underbrace{P(w_1 \mid \langle s\rangle)}_{\text{в слове } w_1 \text{ после } \langle s\rangle} \cdot P(w_2 \mid w_1) \cdot P(w_3 \mid w_2) \cdot \dots \cdot P(\langle/s\rangle \mid w_4)$

вероятность слова w_2 после w_1

$P(A|B) := \frac{P(AB)}{P(B)}$ усл. в-ть

кубик $P(\text{чет} \mid \geq 3) = \frac{P(\text{чет} \geq 3)}{P(\geq 3)} = \frac{2/6}{4/6} = 1/2$

1 2 3 4 5 6
1/6 1/6 1/6 1/6 1/6 1/6

$$P(\text{нечет} \mid \geq 4) = \frac{P(\text{нечет} \geq 4)}{P(\geq 4)} = \frac{1/6}{3/6} = \frac{1}{3}$$ 1 2 3 4 5 6

$$P(\text{нечет}) = 3/6 = \frac{1}{2}$$

$N = 3$

$$P(\langle s\rangle\langle s\rangle\ w_1\ w_2\ w_3 \dots\ w_n\ \langle/s\rangle) =$$

$$= P(w_1 \mid \langle s\rangle\langle s\rangle) \cdot P(w_2 \mid \langle s\rangle w_1) \cdot P(w_3 \mid w_1 w_2) \cdots$$

$$\cdot P(w_i \mid w_{i-2} w_{i-1}) \cdot \dots P(\langle/s\rangle \mid w_{n-1} w_n).$$

Как оценить $P(u \mid w)$? слово u после слова w?

Метод максимального правдоподобия (MLE) — maximum likelihood estimation

Дан корпус, где много текстов / предложений / слов.

можно посчитать правдоподобие

$$P(\text{корпуса}) = P(\text{предл } 1) \cdot P(\text{предл } 2) \cdot \dots \cdot P(\text{предл } n) =$$

$$= P(w_1 \mid \langle s\rangle) \cdot \dots \cdot \dots \cdot \dots \rightarrow \max.$$

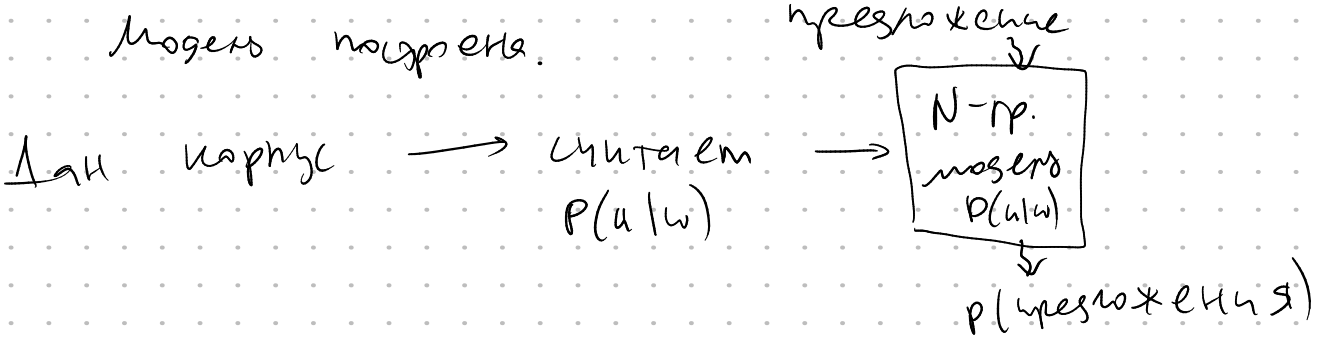
Max достигается при

$$P(u \mid w) = \frac{c(wu)}{c(w)}$$

$c(wu)$ — сколько раз в корпусе было написано слово w, потом u

$c(w)$ — сколько слов w в корпусе.

Модель построена.



$P(\text{я иду спать}) = P(\text{я}|<s>) \cdot P(\text{иду}|\text{я}) \cdot P(\text{спать}|\text{иду}) \cdot P(</s>|\text{спать})$

– $P(u|w) = 0$, если в корпусе нет подряд w и.

$P(\ldots w u \ldots) = 0$. нездобно иметь 0 вероятность, они будут часто встречаться.

Сглаживание – $P(u|w) =$ чуть исправить значение, чтобы всегда $\neq 0$.

P должна остаться вероятностью.

$$\sum_{u \in V} P(u|w) = 1$$

V – весь словарь.

например $P(\text{я}|<s>) + P(\text{он}|<s>) + P(\text{они}|<s>) + P(\text{стол}|<s>) + \ldots + P(\text{ядер}|<s>) = 1$

все слова.

– как сравнить качество разных сглаживаний?
в маш. обучении – подбор гиперпараметров

– оценить качество модели

перерыв до 10:45

Пример: Корпус $N=2$

<s> я выхожу стол </s>

<s> я выхожу стул </s>

<s> я выхожу спать </s>

<s> я ем спать </s>

<s> я стол </s>

Результат обучения

$$\frac{C(\langle s\rangle я)}{C(\langle s\rangle)} = \frac{C(я\ выхожу) = 3}{C(я) = 5}$$

w \ u	<s>	я	выхожу	ем	стол	стул	спать	</s>
<s>	0	1	0	0	0	0	0	0
я	0	0	3/5	1/5	1/5	0	0	0
выхожу	0	0	0	0	1/3	1/3	1/3	0
ем	0						1	
стол	0							1
стул	0							1
спать	0							1
</s>	0	0	0	0	0	0	0	0

$\Sigma = 1 \longrightarrow$

$$\frac{C(выхожу\ стол) = 1}{C(выхожу) \quad 3}$$

$$P(я\ выхожу) = P(я|\langle s\rangle) \cdot P(выхожу|я) \cdot P(\langle/s\rangle|выхожу) =$$

$$= 1 \cdot \tfrac{3}{5} \cdot 0 = 0.$$

гиперпараметр

1 0.1 0.01

Избавление от $P(u|w) = 0$

$$P(u|w) = \frac{C(wu) + \lambda}{C(w) + \lambda|V|}$$

$|V|$ – весь словарь

$\lambda = 1$

w \ u	<s>	я	выхожу	ем	стол	стул	спать	</s>	UNKN
<s>	0	1	0	0	0	0	0	0	
я	1/13	1/13	4/13	2/13	2/13	1/13	1/13	1/13	$\Sigma = 1$
выхожу	0	0	0	0	1/3	1/3	1/3	0	
ем	0						1		
стол	0							1	
стул	0							1	
спать	0							1	
</s>	0	0	0	0	0	0	0	0	
UNKN									

$\frac{0}{5} \to \frac{1}{5+1\cdot 8}$

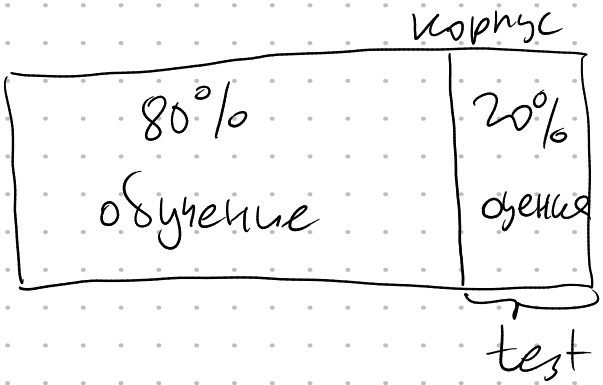
$\frac{1}{5} \to \frac{2}{5+1\cdot 8}$

$\frac{3}{5} \to \frac{4}{5+1\cdot 8}$

При сглаживании все неизвестные слова (не из корпуса) заменяются на UNKN.

Качество модели. Оценивается на корпусе.

обычно делим исходный корпус на части



$P(\text{тест. корпуса}) =$

$= P(\text{предл 1}) \cdot P(\text{предл 2}) \cdot P(\text{предл 3}) \cdot \ldots \to \max$

perplexity — примерно как D, но $\to \min$