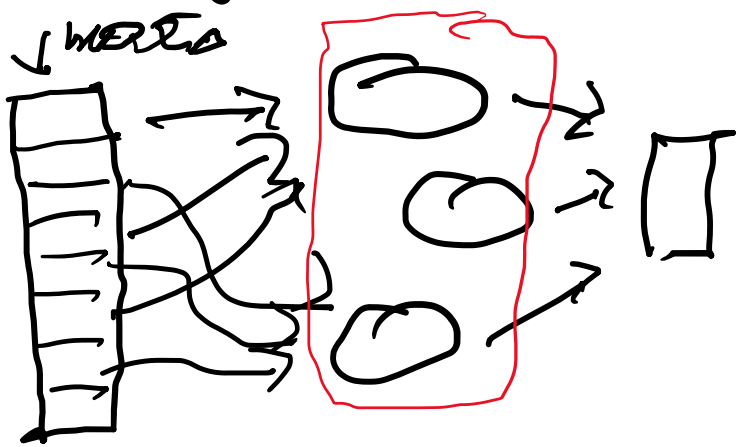
Big Data

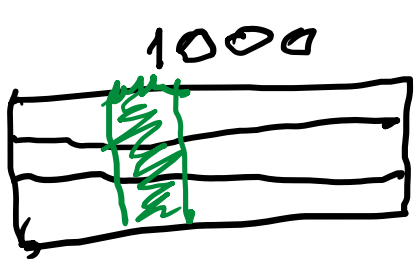
① $10^{12} \approx n$



map-reduce

streaming

② 10^4

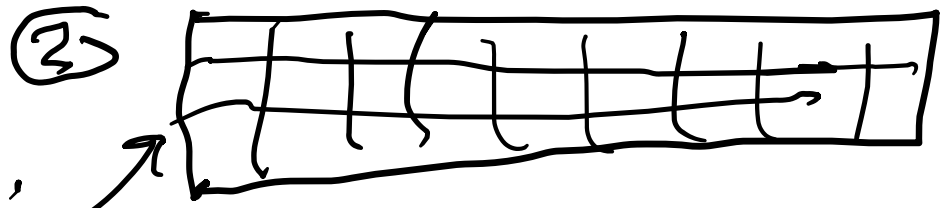


niedy dane

Koszyk zakupów

2^{1000} - podzb

③ 10^6 $\mathbb{R}^n$



n - wiedy dane

$A \subseteq \mathbb{R}^m$, m - duże

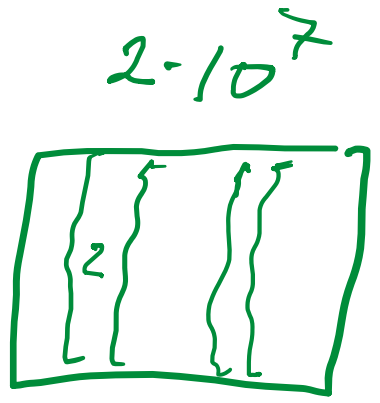
dimension reduction

Koincydencje

① tynek pracy : $20 \cdot 10^6$ – stanowisk

długość pracy : 10 lat

ile ludzi zmieni pracę rocznie $\frac{20 \cdot 10^6}{10} =$



$= 2 \cdot 10^6$

dla jednego urzędnika : 10^6 rok

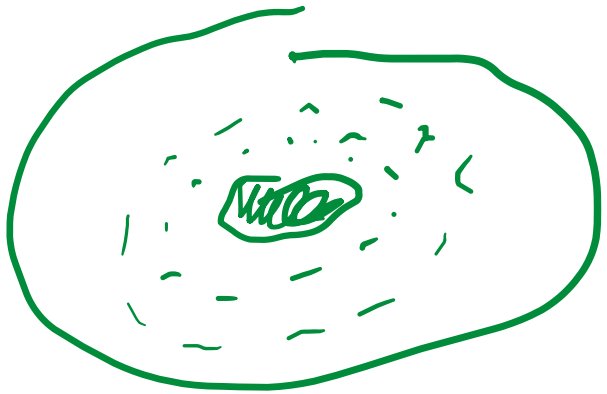
$p = \frac{25 \cdot 10^2}{20 \cdot 10^6} \simeq \frac{1}{10^4}$ $\quad \frac{10^6}{4 \cdot 10^2} = \frac{1}{4} 10^4 = 0.25 \cdot 10^4$ dwóch

q = pr. zobaczenia czerwonego testu $\simeq \frac{1}{10^2}$

$$P[\text{kot} \wedge \text{ciężarna}] = \frac{1}{10^4} \cdot \frac{1}{10^2} = \frac{1}{10^6}$$

Ile osób dookoła? $20 \cdot 10^6 \cdot \frac{1}{10^6} \simeq 20$ osób.

Ile wampirów: $20 \cdot 30 \simeq 600$



1950-70 : paraps. Rhine (J.B.)

10 kart

próba odgłosu paraps. sekcji

- osoby : $\approx 10^4$
- efekt : 10 osób bez ID dupch
  - podejrzany eksp : 0 osób

① $P[\text{sukces}] = \left(\frac{1}{2}\right)^{10} \approx \frac{1}{1000}$

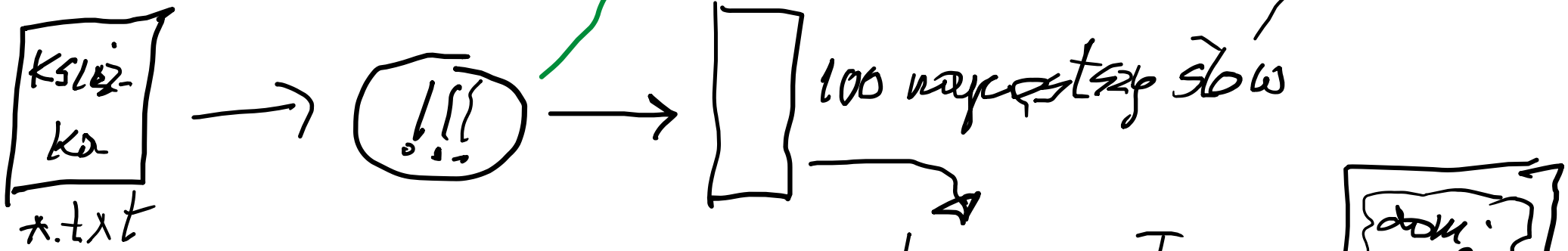
$E[L_1] = 10^4 \cdot 10^{-3} = 10$

$E[L_2] = 10 \cdot 10^{-3} = \frac{1}{100}$.

PIERWSZY PROGRAM

Hello world dla BD:

Analiza tekstu:



① kodowanie → UTF-8

② upper-case

③ eliminacja "STOP-WORDS"

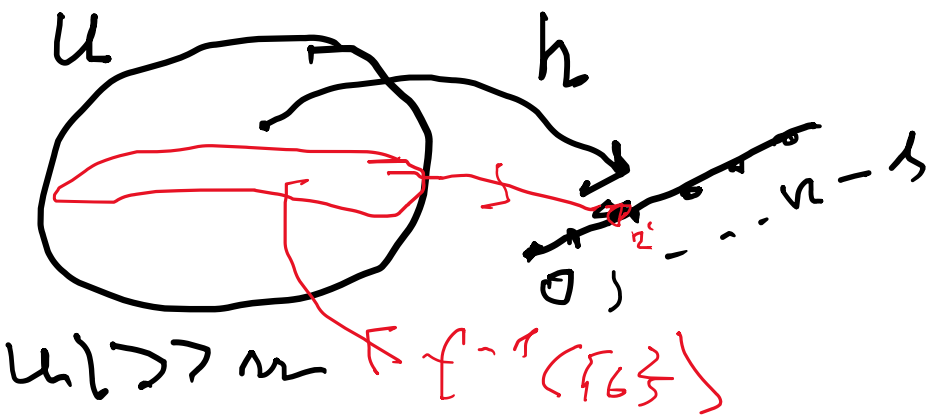
④ tokenizacja słów: ! , ; ? .

- (słowo, pole, słowo, pies, słowo) → groupBy [[słowo, 3], [pole, 1], ...]

filterNot

sortBy

Hash - functions



u - duża liczba

$a_0 a_1 a_2 \dots a_m$ - ciąg znaków

$\downarrow$ 1-1

$\sum_{i=1}^{m} a_i \cdot 256^i$

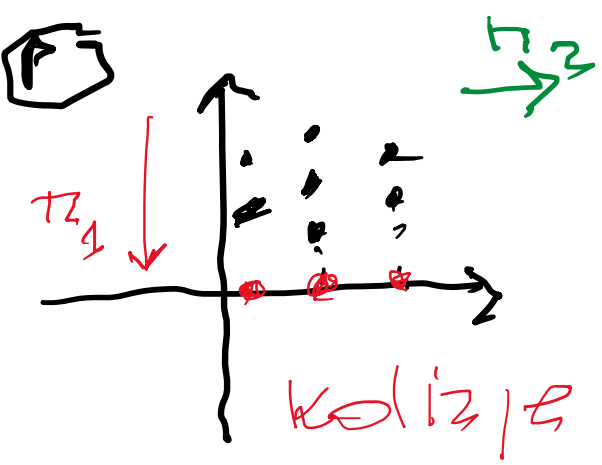
- $h : U \longrightarrow [n]$

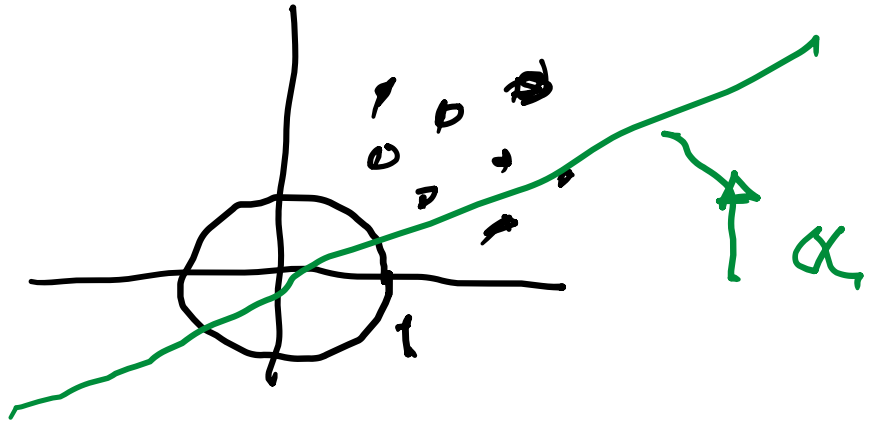
$\rightarrow \exists i \; |h^{-1}(\{i\})| \geq \frac{|U|}{n}$

$U = \bigcup_{i=0}^{n-1} h^{-1}(\{i\})$; gdyby $|h^{-1}(\{i\})| < \frac{|U|}{n}$

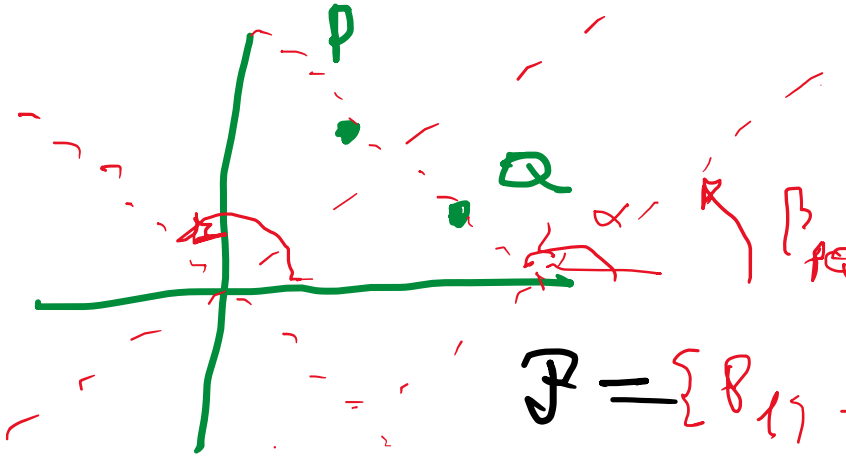
$|U| = \sum_{i=0}^{n-1} |h^{-1}(\{i\})| < \frac{|U|}{n} \cdot n = |U|$ ⚡

$x \neq y$ a $h(x) = h(y)$ ← KOLIZJA





$\alpha \in [0, \pi)$



$\pi_\alpha(x,y) = \cos\alpha \cdot x + \sin\alpha \cdot y$

β_{QP} = "kąt"

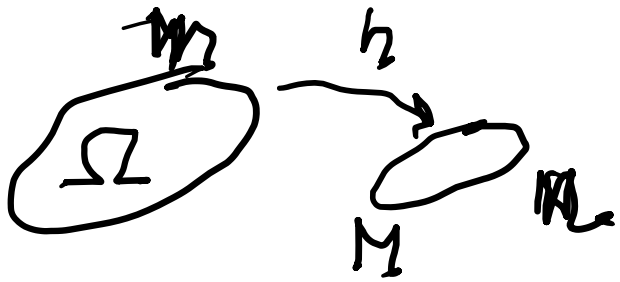
$\mathcal{P} = \{\beta_{1,1} \dots \beta_{n,n}\} : 0 < i < j \leq m$

$\{\beta_{P,P} : 0 < i < j \leq n\} : \binom{n}{2}$

$P[\pi_\alpha$ ma kolizję dla $\beta] = 0$

WYGENERUJ LOSOWO $\alpha \in [0, \pi)$; weź π_α.

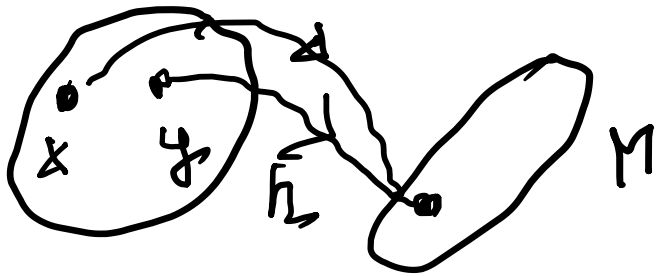
UNIVERSAL HASHING



$\mathcal{H} = M^{\Omega}$; pr. prob.

$P(\omega) = \frac{1}{|M|^{|\Omega|}}$

$\omega \in \mathcal{H}$



$P_h[h(x)=h(y)] = \frac{|\{h \in \mathcal{H} : h(x)=h(y)\}|}{|M|^{|\Omega|}} \leq$

FIX x, y ; LOSUJEMY $h \in \mathcal{H}$

$x \neq y$

$= \frac{|M|^{|\Omega|-1}}{|M|^{|\Omega|}} = \frac{1}{|M|}$ ∎

$\{h : h(x)=h(y)\} = \bigcup_{a \in M} \{h : h(x)=h(y)=a\}$

$\leq \sum_{a \in M} |M|^{|\Omega|-2} = |M|^{|\Omega|-1}$

DEF. $\mathcal{H} \subseteq M^{\Omega}$ jest univ. haszująca

$\mathcal{H}$

$$(\forall x,y \in \Omega)\left(x \neq y \longrightarrow \Pr_{h}[h(x) = h(y)] \leq \frac{1}{|M|}\right)$$