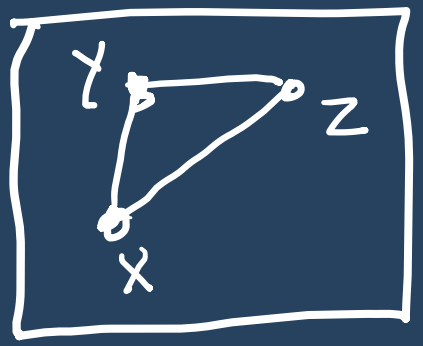
METODA MIN-HASH

① Prz. metryczna : (X, d), $d: X \times X \to [0, \infty)$

a. $(d(x,y) = 0) \equiv (x = y)$

b. $d(x,y) = d(y,x)$

c. $d(x,z) \leq d(x,y) + d(y,z)$



② $\mathbb{R}^n$; $d_e(x,y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}$

od. euklidesowa

③ $C_n = \{0,1\}^n$: $d_H(x,y) = \sum_{i=1}^{n} |x_i - y_i|$ d_H <- odl. Hamminga

Zadanie : d_H jest odl.

$$d_H(x,y) = \sum_{i=1}^{n} |x_i - y_i| = |\{i : x_i = 1 \wedge y_i = 0\}| +$$

$x, y \in \{0,1\}^n$ ($|x_i - y_i| \in \{0,1\}$)

$$|\{i : x_i = 0 \wedge y_i = 1\}|$$

$X = \{i : x_i = 1\}$

$Y = \{i : y_i = 1\}$

$$= |X \setminus Y| + |Y \setminus X|$$

$$= |X \triangle Y|$$

bonus ek:

$(\mathcal{P}(\Omega), \hat{d})$: $\hat{d}(X,Y) = |X \triangle Y|$.

Ω — zb. skończ

przestrzeń metryczna

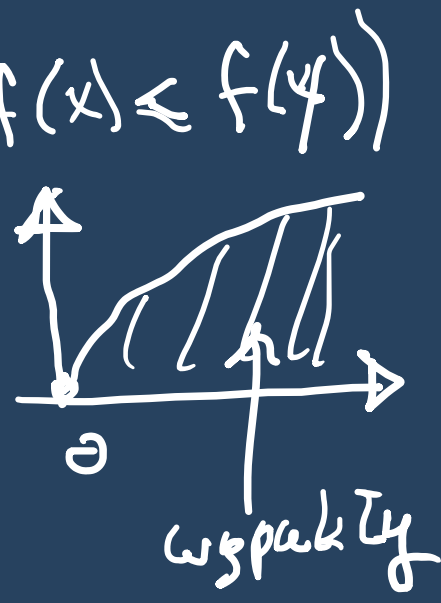
- modyfikacje metryki : (X,d) - fix.
  - $\alpha > 0$: $d^*(x,y) = \alpha \cdot d(x,y)$

    $(X, \alpha \cdot d)$ - p. metr.
  - Tw. Zał. że (X,d) - p. metr. Niech $f: [0,\infty) \to [0,\infty)$ będzie taka, że
    - $f(0) = 0$
    - f jest niemalejąca ($x \le y \to f(x) \le f(y)$)
    - f jest WKLĘSŁA.

    niech $\rho(x,y) = f(d(x,y))$.

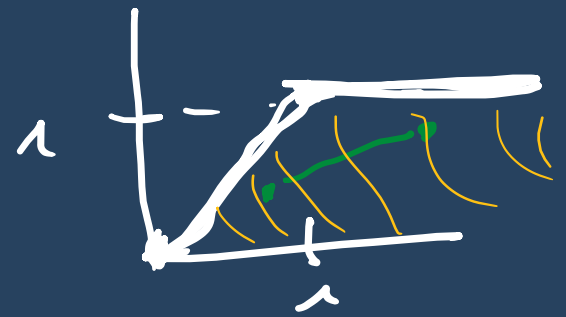
    wtedy (X,ρ) jest p. metryczną.

Zadanie z listy.



Wniosek: (X, d) – prz. metryczna

$$f(x) = \min\{x, 1\}$$



$$\tilde{d}(x,y) = f(d(x,y)) = \min\{d(x,y), 1\}$$

↑ metryka

(PP) $g(x) = \sqrt{x}$

$g'(x) = \left(x^{\frac{1}{2}}\right)' = \frac{1}{2} x^{-\frac{1}{2}} > 0$ dla $x > 0$

$g''(x) = -\frac{1}{4} x^{-\frac{3}{2}} < 0$

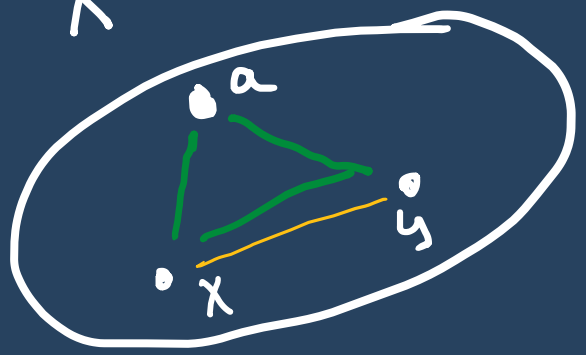
d – metryka

$\tilde{d}(x,y) = \sqrt{d(x,y)}$ – to też jest metryka

▶ Tw (Steinhaus) Zał. że (X, d) jest p. metr. Niech $a \in X$. Na $X \setminus \{a\}$ określamy funkcję

$$\rho(x,y) = \frac{2 \cdot d(x,y)}{d(x,a) + d(y,a) + d(x,y)}$$

wtedy ρ też jest p. metryką (na $X \setminus \{a\}$).



Uwaga:

$$d(x,a) + d(a,y) + d(x,y) \geq d(x,y) + d(x,y) = 2 \cdot d(x,y)$$

Zatem $\rho(x,y) \leq \frac{2 \cdot d(x,y)}{2 \cdot d(x,y)} = 1$.

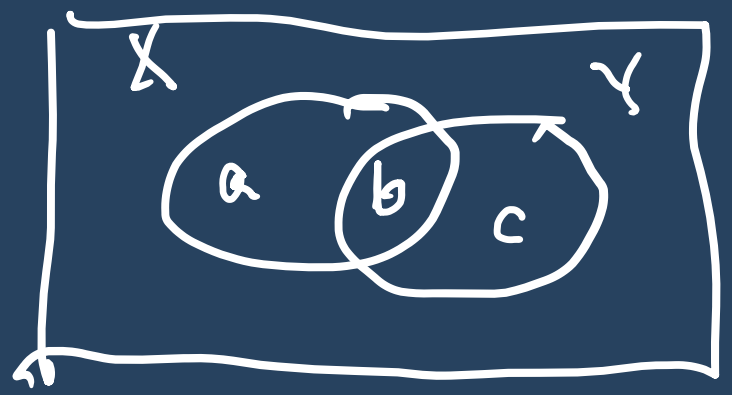
D-d ← zadanie z listy.

Zast. tw. Steinhausa

Ustalamy Ω. Dla $X, Y \subseteq \Omega$: $d(X,Y) = |X \triangle Y|$.

W tw Steih. za a podstawiamy ϕ.

$$\rho(X,Y) = \frac{2 \cdot |X \triangle Y|}{|X \triangle \phi| + |Y \triangle \phi| + |X \triangle Y|} = \frac{2 \cdot |X \triangle Y|}{|X| + |Y| + |X \triangle Y|}$$



$a = |X \setminus Y|$
$b = |X \cap Y|$
$c = |Y \setminus X|$

$$= \frac{2 \cdot |X \triangle Y|}{(a+b) + (b+c) + (a+c)} =$$

$$= \frac{2 \cdot |X \triangle Y|}{2 \cdot (a+b+c)} = \frac{|X \triangle Y|}{|X \cup Y|}$$

CZYLI: na $\{P(\Omega) \setminus \emptyset$ mamy określoną odległość

$$d_J(X,Y) = \frac{|X \triangle Y|}{|X \cup Y|}$$

ODLEGŁOŚĆ JACCARD'a.

2. Podobieństwo obiektów. // similarity

(X, s) $\quad s: X \times X \longrightarrow [0,1]$

- $s(X,X) = 1$
- $s(X,Y) = s(Y,X)$

Idea: jeśli $s(X,Y) \approx 1$ to X jest bardzo podobne do Y.

Dobre podobieństwo:

takie s, że $1-s$ jest metryką

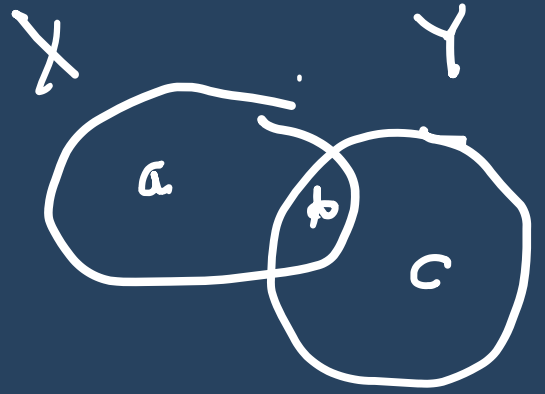
(ograniczone przez 1)

$d(x,y) = 1 - s(x,y)$ ← metryka

(P) $(P^+(\Omega), d_J)$

$s(X,Y) = 1 - d_J(X,Y)$ ← podobieństwo

$$s(X,Y) = 1 - \frac{|X \triangle Y|}{|X \cup Y|} = \frac{|X \cup Y| - |X \triangle Y|}{|X \cup Y|}$$



$$= \frac{(a+b+c) - (a+c)}{|X \cup Y|} = \frac{b}{|X \cup Y|}$$

$$s_J(X,Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

podobieństwo JACCARD'a .

(P) Ω - zbiór słów., D_1, D_2 - dokumenty

D_1^*, D_2^* ← słowa występujące w D_i.

$$J(D_1^*, D_2^*) = \frac{|D_1^* \cap D_2^*|}{|D_1^* \cup D_2^*|}.$$

- $D \subseteq (\Sigma)^*$ $D = (i_1, i_2, i_3, \ldots)$)

Σ - ascii k-shingles (k - zakładki)

D = 

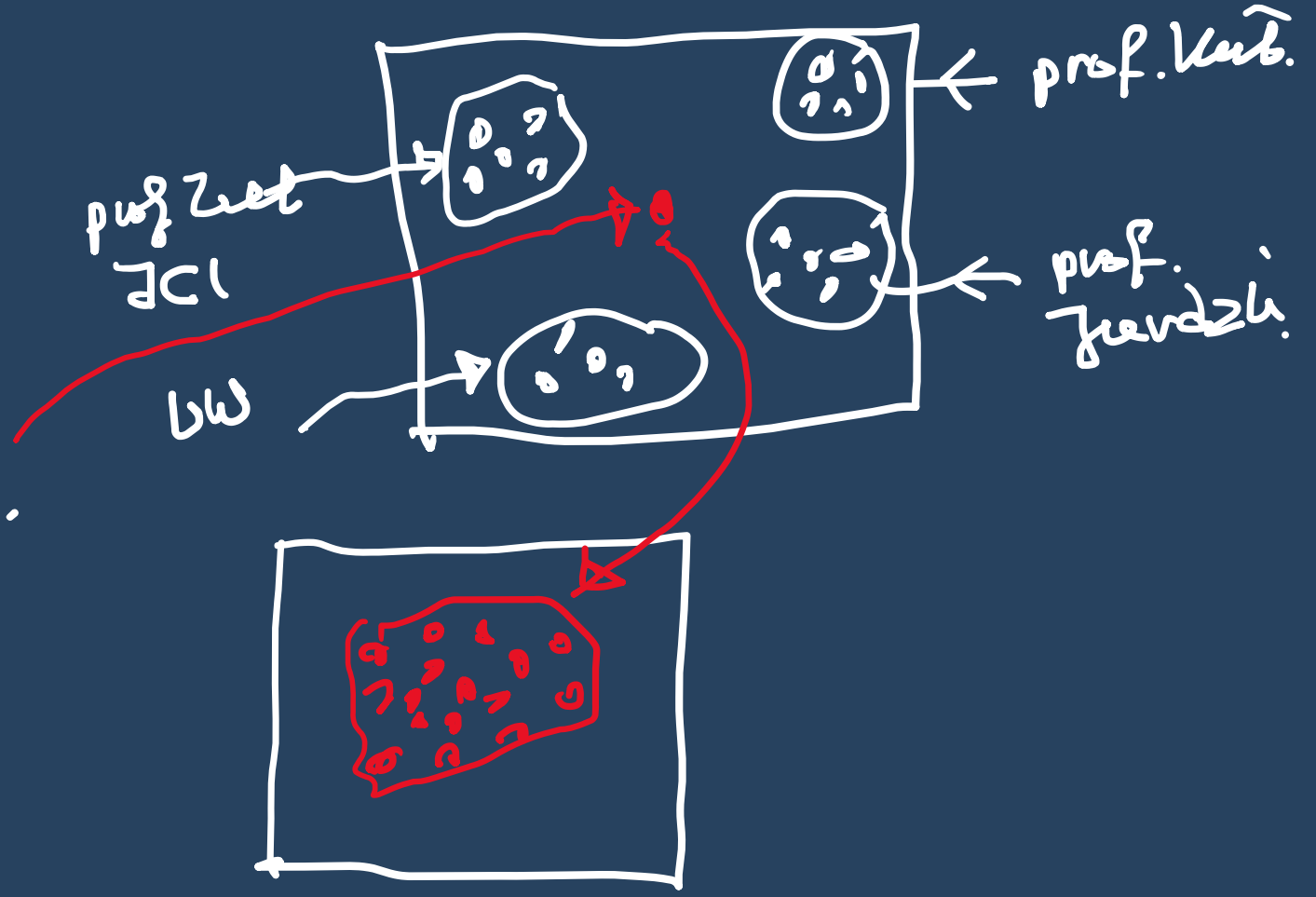
$$3\text{-sh}(D) = \{i_1 i_2 i_3, i_2 i_3 i_4, i_3 i_4 i_5, \ldots\}$$

Podob: $J(3\text{-sh}(X), 3\text{-sh}(Y))$ ← DUŻA SIŁA KLASYFIK.

PRZYKŁAD - $\mathcal{D} = \{D_1, \ldots, D_{500}\}$ - prace w pdf 2 informatyce

$\gamma(D_i^*, D_j^*)$ $i \neq j$

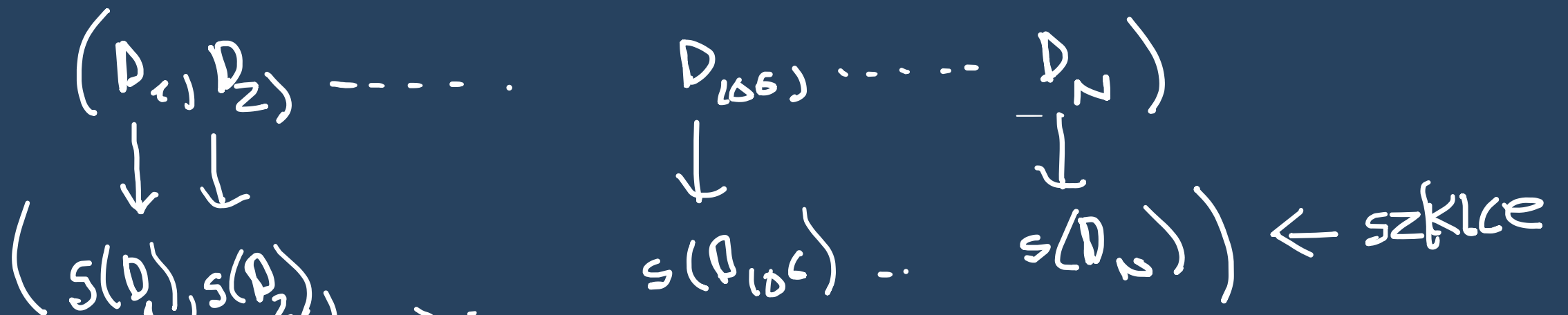
$\Rightarrow$ klasteryzacja



$\mathcal{D}$ u kolekcja
obiekt
nauki dośw.

100

JAK WYZNACZAĆ EFEKTYWNIE $J(X,Y)$?



- mogą to być wektory $[n_1, \dots, n_k]$
  k - niezbyt duże (np. 100, ... 1000)

Mgślimy, że $D \subseteq \{\omega_1, \omega_2, \dots, \omega_N\}$
$\subseteq \{1, 2, \dots, N\} = \mathrm{Ex}.\Omega.$

- Dla $\sigma \in \mathrm{Sym}(\Omega)$ definiujemy

$$f_\sigma(D) = \min\{k : \sigma(k) \in D\}$$

- Na $\mathrm{Sym}(\Omega)$ rozważamy rozkład jednostajny:

$$A \subseteq \mathrm{Sym}(\Omega): \quad \Pr[A] = \frac{|A|}{|\Omega|!}.$$

Ustalamy $D_1, D_2 \subseteq \Omega$, $D_1, D_2 \neq \emptyset$.

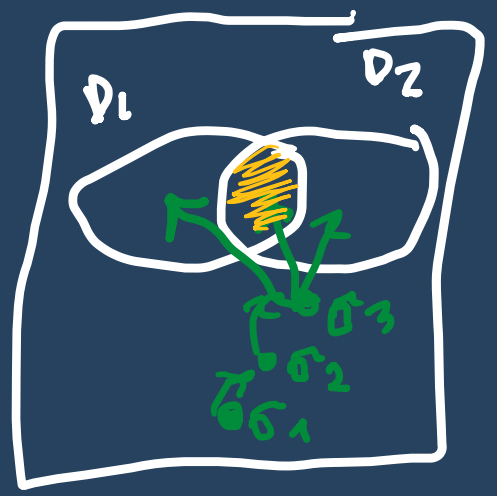
$$\Pr\left[h_\sigma(D_1) = h_\sigma(D_2)\right] =$$

$$= \sum_{i=1}^{N} \Pr\left[h_\sigma(D_1) = h_\sigma(D_2) \mid \min\{k : \sigma(k) \in D_1 \cup D_2\} = i\right] \cdot \Pr\left[\min\{k : \ldots\} = i\right]$$

$$= \sum_{i=1}^{N} \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} \cdot \Pr\left[\min\{k : \ldots\} = i\right]$$

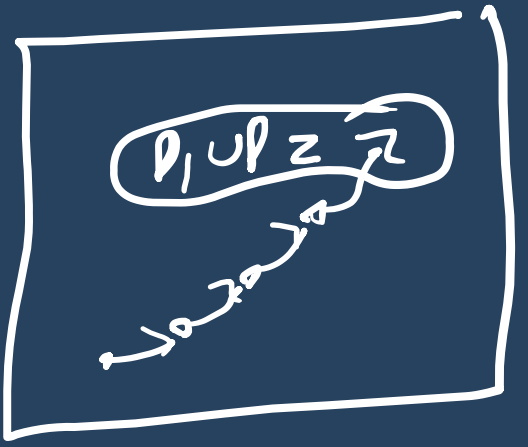
$N = |\Omega|$

$h_\sigma(D) = \min\{k : \sigma(k) \in D\}$



$$= \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|} \sum_{i=1}^{k} \Pr\left[\min\{k : \sigma(k) \in D_1 \cup D_2\} = i\right]$$

$$= J(D_1, D_2)$$



$$\Pr\left[h_\sigma(D_1) = h_\sigma(D_2)\right] = J(D_1, D_2)$$

Q: jak wylosować $\sigma \in Sym(\ell)$?

! czym to możemy zastąpić !

PORZĄDNE FUNKCJE HASZUJĄCE.