

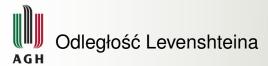
AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

## Przetwarzanie Języka Naturalnego Lab 2

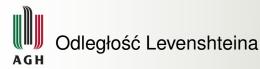
Wojciech Korczyński wojciech.korczynski@agh.edu.pl

Wydział IEiT Katedra Informatyki

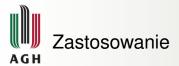
18.03.2015



- inaczej: odległość edycyjna, redakcyjna
- 🔀 metryka w przestrzeni ciągów znaków
- miara podobieństwa dwóch napisów
- uogólnienie odległości Hamminga (uwzględnienie napisów o różnych długościach)



- najmniejsza ilość działań prostych, przekształcających jeden napis na drugi
- działanie proste:
  - dodanie nowego znaku
  - usunięcie znaku
  - zamiana znaku na inny znak



- korekta błędów
- ★ rozpoznawanie mowy
- \* analiza łańcuchów DNA
- wykrywanie plagiatów



- $\bigstar LD(kot, kot) = 0$
- $\bigstar LD(kot, kod) = 1$
- $\blacktriangleright$  *LD*(telefon, telegraf) = 4



$$LD(a,b) = lev_{a,b}(|a|,|b|)$$

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} & min(i,j) \neq 0 \end{cases}$$

$$1_{a_i \neq b_j} = \begin{cases} 0 & a_i = b_j \\ 1 & a_i \neq b_i \end{cases}$$



		В	ı	U	R		
	0	1	2	3	4	5	6
Ρ	1						
I	2						
I Ó R O	3						
R	4 5						
0	5						



-		В	I	U	R	K	0
— Р І	0	1	2	3	4	5	6
Р	1	1					
1	2						
I Ó R O	3						
R	4						
0	5						

$$min(1+1,1+1,0+1)$$



					R		
	0	1	2	3	4	5	6
Ρ		1	2				
1	2						
I Ó R O	3						
R	4						
0	5						

$$min(1+1,2+1,1+1)$$



		В	1	U	R	K	0
	0	1	2	3	4	5	6
Ρ	1	1	2	3	4	5	6
I	2	2					
Ó	3	3			4 4		
R	4	4					
0	5	5					



		В	I	U	R	K	0
	0	1	2	3	4	5	6
Ρ	1	1	2	3	4	5	6
1	2	2	1				
Ó	3	3					
R	4	4					
0	5	5			4 4		

$$min(2+1,2+1,1+0)$$



		В	1	U	R	K	0
	0	1	2	3	4	5	6
Р	1	1	2	3	4	5	6
ı	2	2	1	2	3	4	5
Ó	3	3	2	2	4 4 3 3 <b>2</b> 3	4	5
R	4	4	3	3	2	3	4
0	5	5	4	4	3	3	3



- Napisać program wyliczający odległość Levenshteina między dwoma wprowadzonymi słowami (0.5 pkt)
- Dokonać modyfikacji funkcji wyliczającej odległość Levenshteina, aby uwzględniała ona:
  - błędy ortograficzne w języku polskim (0.5 pkt)
  - znaki diakrytyczne w języku polskim (0.5 pkt)
  - tzw. "czeskie błędy" (0.5 pkt)
- Korzystając z listy form występujących w języku polskim, dokonać korekty wprowadzonych wyrazów (1 pkt)

## Formy:

http://home.agh.edu.pl/~wojtek/pjn2015/lab2.tar.gz

W. Korczyński (KI AGH) PJN 2 2015 13 / 13