



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

$$\text{✚ } LD(kot, kot) = 0$$

$$\text{✚ } LD(kot, kod) = 1$$

$$\text{✚ } LD(telefon, telegraf) = 4$$

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

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

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

		B	I	U	R	K	O
	0	1	2	3	4	5	6
P	1						
I	2						
Ó	3						
R	4						
O	5						

Przykład

		B	I	U	R	K	O
	0	1	2	3	4	5	6
P	1	1					
I	2						
Ó	3						
R	4						
O	5						

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

Przykład

		B	I	U	R	K	O
	0	1	2	3	4	5	6
P	1	1	2				
I	2						
Ó	3						
R	4						
O	5						

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

		B	I	U	R	K	O
	0	1	2	3	4	5	6
P	1	1	2	3	4	5	6
I	2	2					
Ó	3	3					
R	4	4					
O	5	5					

Przykład

		B	I	U	R	K	O
	0	1	2	3	4	5	6
P	1	1	2	3	4	5	6
I	2	2	1				
Ó	3	3					
R	4	4					
O	5	5					

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

Przykład

		B	I	U	R	K	O
	0	1	2	3	4	5	6
P	1	1	2	3	4	5	6
I	2	2	1	2	3	4	5
Ó	3	3	2	2	3	4	5
R	4	4	3	3	2	3	4
O	5	5	4	4	3	3	3

- ❶ Napisać program wyliczający odległość Levenshteina między dwoma wprowadzonymi słowami (0.5 pkt)
- ❷ 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)
- ❸ 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`