



AKADEMIA GÓRNICZO-HUTNICZA  
IM. STANISŁAWA STASZICA W KRAKOWIE

# Przetwarzanie Języka Naturalnego

## Lab 3 – Metryki w przestrzeni napisów

Zbigniew Kaleta  
`zkaleta@agh.edu.pl`

Wydział IEiT  
Katedra Informatyki

22.03.2019

- ✚ Levenshteina (edycyjna)
- ✚ n-gramowe
- ✚ Longest Common Substring

często  
nie w  
możliwe  
potrzebie  
taki problem

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

- ✚ najmniejsza liczba działań prostych, przekształcających jeden napis na drugi
- ✚ działanie proste:
  - ▶ dodanie nowego znaku
  - ▶ usunięcie znaku
  - ▶ zamiana znaku na inny

- ✚  $LD(kot, kot) = 0$
- ✚  $LD(kot, kod) = 1$
- ✚  $LD(telefon, telegraf) = 4$

Chocemy  
wykrywać  
literówki

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

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

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

		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	<b>1</b>					
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
	<b>0</b>	1	2	3	4	5	6
P	1	<b>1</b>	2	3	4	5	6
I	2	2	<b>1</b>	2	3	4	5
Ó	3	3	2	<b>2</b>	3	4	5
R	4	4	3	3	<b>2</b>	<b>3</b>	4
O	5	5	4	4	3	3	<b>3</b>

- ✚ korekta błędów
- ✚ rozpoznawanie mowy
- ✚ analiza łańcuchów DNA
- ✚ wykrywanie plagiatów

nie  
wizualizacji  
mody

✦  $x, y$  - napisy

✦ Dice's coefficient:  $DICE(x, y) = 1 - \frac{2 \times |Ngrams(x) \cap Ngrams(y)|}{|Ngrams(x)| + |Ngrams(y)|}$   
( $Ngrams(x)$  – zbiór wszystkich n-gramów występujących w  $x$ )

✦ “metryka” Dice’a nie spełnia warunku trójkąta

✦ Metryka cosinusowa:  $COSINE(x, y) = 1 - \frac{Ngrams(x) \cdot Ngrams(y)}{|Ngrams(x)| |Ngrams(y)|}$   
( $Ngrams(x)$  – statystyka n-gramów w postaci wektora)

# Metryka LCS (Longest Common Substring)

- ✚  $x, y$  - napisy
- ✚  $f(x, y)$  - najdłuższy wspólny podciąg napisów  $x$  i  $y$
- ✚  $LCS(x, y) = 1 - \frac{|f(x, y)|}{\max(|x|, |y|)}$

to  
liczenie  
reprodukcje

## Preprocessing: Stoplista

wyłączone  
w  
analizie

- ✦ lista słów (lub innych jednostek) bez znaczenia dla całego przetwarzania
- ✦ generowana automatycznie (na podstawie częstotliwości występowania), ręcznie lub hybrydowo
- ✦ na początku przetwarzania należy odfiltrować (usunąć) wszystkie wystąpienia tokenów znajdujących się w stopliście



- ✚ SOUNDEX (1918)
- ✚ Metaphone (1990)
- ✚ Double Metaphone (2000)
- ✚ są to algorytmy stratne

*bonoliter  
specjalizacja  
terminy*

- ✦ Micro-average precision: należy zsumować TP, FP i FN policzone dla każdej klasy osobno i obliczyć precyzję wg. normalnego wzoru
- ✦ Macro-average precision: należy obliczyć precyzję dla każdej klasy osobno, a następnie obliczyć średnią arytmetyczną
- ✦ F1 jest zawsze średnią harmoniczną precision i recall

## ✚ Davies-Bouldin index:

$$DB = \frac{1}{n} \sum_{i=1}^n \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)}$$

$c_x$  to centroid klastra, a  $\sigma_x$  to średnia odległość między elementami klastra  $x$

## ✚ Dunn index:

$$D = \frac{\min_{1 \leq i < j \leq n} d(i, j)}{\max_{1 \leq k \leq n} d'(k)}$$

$d$  to odległość pomiędzy klastrami, a  $d'$  to rozmiar klastra

- 1 Napisać program klasteryzujący nazwy firm z pliku *lines.txt*:
  - ▶ Wykonać potrzebny preprocessing (**stworzyć stoplistę**, etc.) (1 pkt)
  - ▶ Dokonać klasteryzacji przy pomocy dwóch wybranych metryk (1 pkt)
  - ▶ Ocenić jakość klasteryzacji, porównać wyniki (1 pkt)

Materiały:

<http://home.agh.edu.pl/~zkaleta/pjn/lab3.tar.gz>