

AKADEMIA GÓRNICZO-HUTNICZA IM. STANISŁAWA STASZICA W KRAKOWIE

Przetwarzanie Języka Naturalnego Lab 3 – Metryki w przestrzeni napisów

Zbigniew Kaleta zkaleta@agh.edu.pl

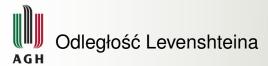
Wydział IEiT Katedra Informatyki

22.03.2019



Metryki w przestrzeni napisów

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



- 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



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

chouse the service



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

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

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

Z. Kaleta (KI AGH) PJN 3 2019 6 / 20



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



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

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

Z. Kaleta (KI AGH) PJN 3 2019 8 / 20



-					R	K	0
	0	1	2	3	4	5	6
	1	1	2				
1	2						
I Ó R O	3						
R	4						
0	5						

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

Z. Kaleta (KI AGH) PJN 3 2019 9 / 20



		В		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					

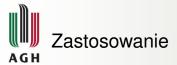


		В	I	U	R	K	0
	0	1	2	3	4	5	6
Р	1	1	2	3	4	5	6
1	2	2	1				
Ó	3	3					
R	4	4					
0	5	5			4 4		

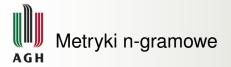
$$min(2+1,2+1,1+0)$$



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



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





- \times 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)

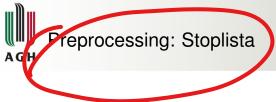


Metryka LCS (Longest Common Substring)

- \times x, y napisy
- $\maltese f(x,y)$ najdłuższy wspólny podciąg napisów x i y

$$LCS(x, y) = 1 - \frac{|f(x,y)|}{max(|x|,|y|)}$$





winterne

- generowana automatycznie (na podstawie częstaliwości występowania), ręcznie lub hybrydowo
- na początku przetwarzania należy odfiltrować (usunąć) wszystkie wystąpienia tokenów znajdujących się w stopliście



Preprocessing: Algorytmy fonetyczne

- **★** SOUNDEX (1918)
- Metaphone (1990)
- ★ Double Metaphone (2000)
- ¥ są to algorytmy stratne

Z. Kaleta (KI AGH) PJN 3 2019 17 / 20

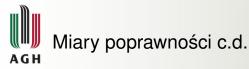


Miary poprawności c.d.

borologo TP EP i EN poliozono

- 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 średnia harmoniczna precision i recall

Z. Kaleta (KI AGH) PJN 3 2019 18 / 20



▼ 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 σ_x to średnia odległość między elementami klastra x

■ Dunn index:

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

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



- 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)

Materialy:

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