



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Algorytmy Tekstowe

Lab 5 – Metryki w przestrzeni napisów

Zbigniew Kaleta
`zkaleta@agh.edu.pl`

Wydział IEiT
Instytut Informatyki

05.05.2021

- ✚ Levenshteina (edycyjna)
- ✚ n-gramowe (Dice, cosinusowa, euklidesowa)
- ✚ Longest Common Substring

Metryka LCS (Longest Common Substring)

✚ x, y - napisy

✚ $f(x, y)$ - najdłuższy wspólny podnapis napisów x i y

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

	A	B	C	D	E	F
	0	0	0	0	0	0
B	0	0	1	0	0	0
C	0					
D	0					
F	0					

	A	B	C	D	E	F
	0	0	0	0	0	0
B	0	0	1	0	0	0
C	0	0	0	2	0	0
D	0					
F	0					

	A	B	C	D	E	F
	0	0	0	0	0	0
B	0	0	1	0	0	0
C	0	0	0	2	0	0
D	0	0	0	0	3	0
F	0	0	0	0	0	1

Czy da się szybciej niż $O(n*m)$?

- ✚ budujemy tablicę sufiksów i LCP – $O(n+m)$
- ✚ korzystamy z okna przesuwnego (minimum sliding range query problem) – $O(N)$, albo w łatwiejszej wersji $O(N\log N)$

<https://www.youtube.com/watch?v=Ic80xQFWevc>

- ✦ n-gramem nazywamy **każdą** sekwencję n kolejnych składowych
- ✦ sekwencje się zazębiają
- ✦ w przypadku analizy języka składowymi mogą być litery, sylaby lub słowa

Słowo: algorytm

digramy: al, lg, go, or, ry, yt, tm

trigramy: alg, lgo, gor, ory, ryt, ytm

Zdanie: Mężny bądź, chroń pułk twój i sześć flag.

digramy: Mężny bądź, bądź chroń, chroń pułk, pułk twój, twój i, i sześć, sześć flag

✚ 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)

x = "abcabc"

digramy: ab, bc, ca, ab, bc

$$Ngrams(x) = (n_{aa}, n_{ab}, n_{ac}, \dots, n_{zz}) = (0, 2, 0 \dots 0)$$

albo

$$Ngrams(x) = (f_{aa}, f_{ab}, f_{ac}, \dots, f_{zz}) = (0, 0.4, 0 \dots 0)$$

- ✚ lista słów (lub innych jednostek) bez znaczenia dla dalszego 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

✦ 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:

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

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