Rozdział 1

Metryki na przestrzeni ciągów znaków

1.1 Podstawowe definicje

Definicja 1.1.1. Napisem *nazywamy skończone złączenie symboli (znaków) ze skończonego* alfabetu, *oznaczonego przez* Σ. *Produkt kartezjański rzędu* q, $\Sigma \times \ldots \times \Sigma$ *oznaczamy przez* Σ^q, *natomiast zbiór wszystkich skończonych napisów, które można utworzyć ze znaków z* Σ *oznaczamy przez* Σ^*. Pusty napis, *oznaczany* ε, *również należy do* Σ^*. *Napisy zwyczajowo będziemy oznaczać przez* s, t *oraz* u, *a ich* długość, *czyli liczbę znaków w napisie, przez* $|s|$.

Przykład 1.1.1. *Niech* Σ *będzie alfabetem złożonym z* 26 *małych liter alfabetu łacińskiego oraz niech* $s =' ala'$. *Wówczas mamy* $|s| = 3$, $s \in \Sigma^3$ *oraz* $s \in \Sigma$. *Pojedyncze znaki oznaczamy przez indeks dolny, stąd mamy* $s_1 =' a'$, $s_2 =' l'$, $s_3 =' a'$. *[1]*

Definicja 1.1.2. *Funkcję* d *nazywamy* metryką *na* Σ^*, *jeśli ma poniższe własności:*

- $d(s,t) \geq 0$
- $d(s,t) = 0$ *wtedy i tylko wtedy, gdy* $s = t$
- $d(s,t) = d(t,s)$
- $d(s,u) \leq d(s,t) + d(t,u)$,

gdzie s, t, u *są napisami.*

Metryki na napisach można podzielić na trzy grupy:

- oparte na operacjach edycyjnych (*edit operations*),
- oparte na q-gramach,
- miary heurystyczne.

1.2 Odległości na napisach oparte na operacjach edycyjnych

Metryki oparte na operacjach edycyjnych zliczają liczbę opercji potrzebnych do przetworzenia jednego napisu w drugi. Najczęściej wymieniamymi operacjami są:

- zamiana znaku, np. $'ala' \to' ela'$
- usunięcie znaku, np. $'ala' \to' aa'$
- wstawienie znaku, np. $'ala' \to' alka'$
- transpozycja dwóch przylegających znaków, np. $'ala' \to' laa'$

Przykładowe odległości: Hamminga, najdłuższego wspólnego podnapisu (*longest common substring*), Levenshteina, optymalnego dopasowania napisów (*optimal string alignment*), Damareu-Levenshteina. Nie wszystkie z ww. odległości są metrykami.

Metryka **najdłuższego wspólnego podnapisu**, ozn. d_{lcs}, zlicza liczbę usunięć i wstawień, potrzebnych do przetworzenia jednego napisu w drugi. Np. $d_{lsc}('leia',' leela') = 3$, bo $leela \xrightarrow{us.\ e} lela \xrightarrow{us.\ l} lea \xrightarrow{wst.\ i} leia$.

Uogólniona **odległość Levenshteina**, ozn. d_{lv} zlicza ważoną sumę usunięć, wstawień oraz zamian znaków, potrzebnych do przetworzenia jednego napisu w drugi.

Gdy za wagi przyjmuje się 1 mamy do czynienia ze zwykłą odległością Levenshteina, np.

$d_{lv}('leia','leela') = 2$, bo $leela \xrightarrow{us.\ e} lela \xrightarrow{zm.\ l\ na\ i} leia$.

Gdy za wagi przyjmiemy np. $(0.1, 1, 1)$,
$d_{lv}('leia','leela') = 1.1$, bo $leela \xrightarrow[0.1]{us.\ e} lela \xrightarrow[1]{zm.\ l\ na\ i} leia$

Bibliografia

[1] Mark P. J. van der Loo. The stringdist Package for Approximate String Matching. *The R Journal*, 6:111–122, 2014.

Warszawa, dnia .....

Oświadczenie

Oświadczam, że pracę licencjacką pod tytułem: „Automatyczna kategoryzacja tematyczna tekstów przy użyciu metryk w przestrzeni ciągów znaków”, której promotorem jest dr Marek Gągolewski, wykonałem/am samodzielnie, co poświadczam własnoręcznym podpisem.

.....