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* $|s|$. *Poprzez* s_i *rozumiemy* i*-ty znak z napisu* s, *dla każdego* $i \in \{1, \ldots, |s|\}$ *[3].*

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* $s_1 =' a'$, $s_2 =' l'$, $s_3 =' a'$. *[3]*

Odległość $d(s, t)$ pomiędzy dwoma napisami s i t to minimalny koszt ciągu operacji potrzebnego do przetransformowania s w t (i ∞, gdy taki ciąg nie istnieje). Koszt ciągu operacji jest sumą kosztów pojedynczych operacji. Przez operacje rozumiemy skończoną liczbę reguł w formie $\delta(x, y) = a$, gdzie x i y to różne podnapisy, a a to nieujemna liczba rzeczywista. Kiedy już, przy pomocy operacji, podnapis x zostanie przekształcony w napis y, żadne dalsze operacje nie mogą być wykonywane na y [2].

Zauważmy w szczególności ostatnie ograniczenie, które nie pozwala wielokrot-

nie przekształcać tego samego podnapisu. Gdyby pominąć to założenie, każdy system przekształcający napisy spełniałby definicję i stąd odległość między dwoma napisami nie byłaby, w ogólności, możliwa do policzenia [2].

Jeśli dla każdej operacji $\delta(x, y)$, istnieje odpowiednia operacja $\delta(y, x)$ o takim samym koszcie, to odległość jest symetryczna (tj. $d(s, t) = d(t, s)$). Zauważmy również, że:

- $d(s, t) \geq 0$ dla wszystkich napisów s, t,
- $d(s, s) = 0$,
- $d(s, u) \leq d(s, t) + d(t, u)$.

Stąd, jeśli odległość jest symetryczna, przestrzeń napisów tworzy przestrzeń metryczną [2].

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 z Σ^*.

Odległości 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 minimalną liczbę opercji potrzebnych do przetworzenia jednego napisu w drugi. Najczęściej wymieniamymi operacjami są:

- usunięcie znaku: $\delta(l, \varepsilon)$, tj. usunięcie litery $'l'$, np. $'ala' \rightarrow' aa'$
- wstawienie znaku: $\delta(\varepsilon, l)$, tj. wstawienie litery $'l'$, np. $'ala' \rightarrow' alka'$
- zamiana znaku: $\delta(a, e)$, tj. zamiana litery $'a'$ na $'e'$, np. $'ala' \rightarrow' ela'$
- transpozycja: $\delta(al, la)$, tj. przestawienie dwóch przylegających liter $'a'$ i $'l'$, np. $'ala' \rightarrow' laa'$

Dla wszystkich odległości, które dopuszczają więcej niż jedną operację edycyjną, może być znaczące nadanie wag poszczególnym operacjom, na przykład dając transpozycji mniejszy koszt niż wstawienie znaku. Odległości, dla których takie wagi zostają nadane są zazwyczaj nazywane *uogólnionymi* odległościami [1].

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.

Definicja 1.2.1. Odległością Hamminga *na* Σ^* *nazywamy:*

$$d_{hamming}(s,t) = \begin{cases} \sum_{i=1}^{|s|}[1-\delta(s_i,t_i)], & \textit{gdy } |s| = |t|, \\ \infty, & \textit{w przeciwnym przypadku}, \end{cases}$$

gdzie

$$\delta(s_i,t_i) = \begin{cases} 1, & \textit{gdy } s_i = t_i, \\ 0, & \textit{w przeciwnym przypadku}. \end{cases}$$

Odległość Hamminga dopuszcza jedynie zamianę znaku, stąd jest zdefiniowana tylko dla napisów o równej długości.

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 $leela \xrightarrow[0.1]{us.\ e} lela \xrightarrow[1]{zm.\ l\ na\ i} leia$

Metryka **optymalnego dopasowania napisów**, ozn. d_{osa}, zlicza liczbę usunięć, wstawień, zamian oraz transpozycji przylegających znaków, potrzebnych do przetworzenia jednego napisu w drugi. Np. $d_{osa}('leia',' leela') = 2$, bo $leela \xrightarrow{us.\ e} lela \xrightarrow{zm.\ l\ na\ i} leia$.

Metryka ta nie spełnia nierówności trójkąta: $2 = d_{osa}('ba',' ab') + d_{osa}('ab',' acb') \leq d_{osa}('ba',' acb') = 3$

Bibliografia

[1] Leonid Boytsov. Indexing methods for approximate dictionary searching: Comparative analysis. *J. Exp. Algorithmics*, 16:1.1:1.1–1.1:1.91, May 2011.

[2] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[3] 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.

.....