



Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych



Automatyczna kategoryzacja tematyczna tekstów przy użyciu metryk w przestrzeni ciągów znaków

Natalia Potocka
Warszawa, 21.04.2014

- Cel pracy
- O metrykach słów kilka
- Postęp prac
- Co dalej?

Celem pracy jest skategoryzowanie tekstów z polskiej Wikipedii pod względem tematu na podstawie liczności słów występujących w tekście. Można się spodziewać, że jeśli w dwóch tekstach występuje dużo podobnych do siebie słów, to pochodzą one z tej samej kategorii tematycznej.

Celem pracy jest skategoryzowanie tekstów z polskiej Wikipedii pod względem tematu na podstawie liczności słów występujących w tekście. Można się spodziewać, że jeśli w dwóch tekstach występuje dużo podobnych do siebie słów, to pochodzą one z tej samej kategorii tematycznej.

A		B		C		D	
całka	10	całka	5	niewłaściwy	3	ułamek	4
po pochodna	5	po pochodna	15	powieść	7	mianownik	5
niewłaściwa	4	granica	7	granica	15	niewłaściwy	6

Co ze słowami podobnymi? Przykładowo słowa *niewłaściwy* i *niewłaściwa* mają ten sam temat, różnią się tylko rodzajem (męski / żeński). W tekstach mogą też występować błędy ortograficzne, błędy spowodowane brakami znaków diaktrycznych (ą, ę, ł, ...) itd. Takie słowa również chcielibyśmy traktować jak "podobne". W celu określenia jak bardzo dwa słowa są do siebie podobne, posłużą *metryki określone na napisach*.

DEFINICJA

Napisem nazywamy skończone złączenie symboli (znaków) ze skończonego *alfabetu*, oznaczonego przez Σ . Produkt kartezjański rzędu q , $\Sigma \times \dots \times \Sigma$ 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 $|s|$.

DEFINICJA

Napisem nazywamy skończone złączenie symboli (znaków) ze skończonego *alfabetu*, oznaczonego przez Σ . Produkt kartezjański rzędu q , $\Sigma \times \dots \times \Sigma$ 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. 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 $s_1 = 'a'$, $s_2 = 'l'$, $s_3 = 'a'$. Podnapis oznaczamy przez $m : n$ w indeksie dolnym, np. $s_{1:2} = 'al'$. Jeśli $n < m$, to $s_{m:n} = \varepsilon$, czyli napis pusty.

DEFINICJA

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

- $d(s, t) \geq 0$
- $d(s, t) = 0$ wtw $s = t$
- $d(s, t) = d(t, s)$
- $d(s, u) \leq d(s, t) + d(t, u)$,

gdzie s, t, u są napisami.

DEFINICJA

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

- $d(s, t) \geq 0$
- $d(s, t) = 0$ wtw $s = t$
- $d(s, t) = d(t, s)$
- $d(s, u) \leq d(s, t) + d(t, u)$,

gdzie s, t, u są napisami.

Nie wszystkie metryki na napisach posiadają wszystkie z wyżej wymienionych własności.

DEFINICJA

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

- $d(s, t) \geq 0$
- $d(s, t) = 0$ wtw $s = t$
- $d(s, t) = d(t, s)$
- $d(s, u) \leq d(s, t) + d(t, u)$,

gdzie s, t, u są napisami.

Nie wszystkie metryki na napisach posiadają wszystkie z wyżej wymienionych własności.

Metryki na napisach można podzielić na trzy grupy:

- oparte na operacjach edytowania (*edit operations*)
- oparte na q -gramach
- miary heurystyczne

DEFINICJA

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

- $d(s, t) \geq 0$
- $d(s, t) = 0$ wtw $s = t$
- $d(s, t) = d(t, s)$
- $d(s, u) \leq d(s, t) + d(t, u)$,

gdzie s, t, u są napisami.

Nie wszystkie metryki na napisach posiadają wszystkie z wyżej wymienionych własności.

Metryki na napisach można podzielić na trzy grupy:

- **oparte na operacjach edytowania** (*edit operations*)
- oparte na q -gramach
- miary heurystyczne

Metryki oparte na operacjach edytowania zliczają liczbę operacji potrzebnych do przetworzenia jednego napisu w drugi. Najczęściej wymienianymi operacjami są:

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

Przykładowe metryki: Hamminga, najdłuższego wspólnego podnapisu (*longest common substring*), Levenshteina, optymalnego dopasowania napisów (*optimal string alignment*), Damareu-Levenshteina.

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 $leela \xrightarrow{us. e} lela \xrightarrow{us. l} lea \xrightarrow{wst. i} leia$.

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.

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 $leela \xrightarrow{us. e} lela \xrightarrow{zm. l \text{ na } i} leia$.

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 przyjmujemy 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.

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

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:

$$2 = d_{osa}('ba', 'ab') + d_{osa}('ab', 'acb') \leq d_{osa}('ba', 'acb') = 3$$

Dziękuję za uwagę.