Politechnika Warszawska
Wydział Matematyki i Nauk Informacyjnych

Automatyczna kategoryzacja tematyczna tekstów przy użyciu metryk w przestrzeni ciągów znaków

Natalia Potocka

Warszawa, 25.01.2016

Cel pracy

Celem pracy jest skategoryzowanie tekstów z polskiej Wikipedii pod względem tematu na podstawie liczności słów występujących w tekście.

Cel pracy

Celem pracy jest skategoryzowanie tekstów z polskiej Wikipedii pod względem tematu na podstawie liczności słów występujących w tekście.
Problemy:

- duże wymiary danych (1 075 568 $\times$ 2 806 765),
- dane bardzo rzadkie (ponad $99{,}99\%$),
- duża złożoność obliczeniowa i pamięciowa.

Cel pracy

Celem pracy jest skategoryzowanie tekstów z polskiej Wikipedii pod względem tematu na podstawie liczności słów występujących w tekście.
Problemy:

- duże wymiary danych (1 075 568 $\times$ 2 806 765),
- dane bardzo rzadkie (ponad $99{,}99\%$),
- duża złożoność obliczeniowa i pamięciowa.

Remedium to redukcja liczba słów przy użyciu *stemmingu* oraz *odległości na przetrzeni ciągów znaków*.

Stemming

TABLICA : Przykładowe skupienia uzyskane przy pomocy *stemmingu*.

działalność	niemiecki	odkryty	okres	postać
działalność	niemiecki	odkryta	okres	postaci
działalności	niemieckiej	odkryte	okresu	postacie
działalnością	niemieckiego	odkryty	okresach	postać
działalnościach	niemieckich	odkrytych	okresem	postacią
działalnościami	niemieckim	odkrytym	okresy	postaciami
	niemiecką	odkrytą	okresów	postaciach
	niemiecka	odkrytego	okresami	postaciom
	niemieccy	odkrytej	okresom	postał
	niemieckimi	odkrytymi		postała
	niemiecku	nieodkrytych		postania
	niemieckiemu	nieodkryte		postało
	nieniemieckich	odkrytemu		postały
	nieniemieckiej	odkryci		postaniu

Operacje edytowania

Odegłości oparte na operacjach edytowania zliczają liczbę opercji potrzebnych do przetworzenia jednego napisu w drugi. Najczęściej wymieniamymi operacjami są [1]:

- zamiana znaku, np. $'ela' \rightarrow' ala'$
- usunięcie znaku, np. $'ela' \rightarrow' ea'$
- wstawienie znaku, np. $'ela' \rightarrow' elka'$
- transpozycja dwóch przylegających znaków, np. $'ela' \rightarrow' lea'$

Operacje edytowania

Odegłości oparte na operacjach edytowania zliczają liczbę opercji potrzebnych do przetworzenia jednego napisu w drugi. Najczęściej wymieniamymi operacjami są [1]:

- zamiana znaku, np. $'ela' \rightarrow' ala'$
- usunięcie znaku, np. $'ela' \rightarrow' ea'$
- wstawienie znaku, np. $'ela' \rightarrow' elka'$
- transpozycja dwóch przylegających znaków, np. $'ela' \rightarrow' lea'$

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

Odległości oparte na q-gramach

Definicja

Podnapis złożony z kolejnych, przylegających do siebie znaków, o ustalonej długości $q \geq 1$ jest nazywany q-*gramem*.

Definicja

Niech $\mathcal{Q}(s, q)$ oznacza zbiór unikalnych q-gramów występujących w napisie s. Wówczas *odległość Jaccarda*, d_{jac}, między napisami s i t definiuje się jako:

$$d_{\text{jac}}(s, t, q) = 1 - \frac{|\mathcal{Q}(s, q) \cap \mathcal{Q}(t, q)|}{|\mathcal{Q}(s, q) \cup \mathcal{Q}(t, q)|},$$

gdzie $|\cdot|$ oznacza liczność zbioru.

Miary heurystyczne

Niech s i t będą napisami. Niech m oznacza liczbę wspólnych znaków z s i t, przy czym zakładając, że $s_i = t_j$, to znak ten jest *wspólny* dla obu napisów, jeśli $|i - j| < \lfloor \frac{max\{|s|,|t|\}}{2} \rfloor$ i każdy znak z s może być wspólny ze znakiem z t tylko raz. W końcu, jeśli s' i t' są podnapisami utworzonymi z s i t poprzez usunięcie znaków, które nie są wspólne dla obu napisów, to T jest liczbą transpozycji potrzebnych to otrzymania t' z s'. Transpozycje znaków nieprzylegających są dozwolone.

Definicja

Odległość Jaro definiuje się jako [2]:

$$d_{\text{jaro}}(s,t) = \begin{cases} 0, & \text{gdy } s = t = \varepsilon, \\ 1, & \text{gdy } m = 0 \text{ i } |s| + |t| > 0, \\ 1 - \frac{1}{3}(\frac{m}{|s|} + \frac{m}{|t|} + \frac{m-T}{m}) & \text{w przeciwnym przypadku.} \end{cases}$$

Utworzenie skupień słów

Zaproponowano trzy algorytmy opierające się na wybranych odległościach:

1. Dołączeniu do skupień słów jeszcze niepogrupowanych.
2. Dołączeniu do skupień zawierających pięć i więcej elementów, podzbiorów o mniejszej liczności.
3. Zastosowaniu najpierw punktu 1, a następnie punktu 2.

Utworzenie skupień słów

Zaproponowano trzy algorytmy opierające się na wybranych odległościach:

1. Dołączeniu do skupień słów jeszcze niepogrupowanych.
2. Dołączeniu do skupień zawierających pięć i więcej elementów, podzbiorów o mniejszej liczności.
3. Zastosowaniu najpierw punktu 1, a następnie punktu 2.

W ten sposób otrzymano 16 różnych reprezentacji tekstów, odpowiadających różnym grupom słów, otrzymanych przy użyciu różnych odległości i powyższych algorytmów.

Postępy prac

Postępy prac

Co dalej?

Bibliografia

[1] Leonid Boytsov. Indexing methods for approximate dictionary searching: Comparative analysis. *Journal of Experimental Algorithmics*, 16:1–91, 2011.

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