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 zbadanie wpływu doboru odległości na przestrzeni napisów na jakość automatycznej kategoryzacji tematycznej tekstów.

Cel pracy

Celem pracy jest zbadanie wpływu doboru odległości na przestrzeni napisów na jakość automatycznej kategoryzacji tematycznej tekstów.
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 zbadanie wpływu doboru odległości na przestrzeni napisów na jakość automatycznej kategoryzacji tematycznej tekstów.
Problemy:

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

Grupujemy więc słowa przy użyciu *stemmingu* oraz *odległości na przestrzeni 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 [2]:

- 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 [2]:

- 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 [4]:

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

Utworzenie skupień słów

Zaproponowano trzy algorytmy opierające się na wybranych odległościach:

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

Utworzenie skupień słów

Zaproponowano trzy algorytmy opierające się na wybranych odległościach:

1. dołączenie do skupień słów jeszcze niepogrupowanych,
2. dołączenie do skupień zawierających pięć i więcej elementów, podzbiorów o mniejszej liczności,
3. zastosowanie 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.

Algorytm k-średnich

W metodzie k-średnich minimalizujemy

$$\sum_i d(\mathbf{x}_i, \mathbf{m}_{C(i)}),$$

gdzie $\mathbf{x}_i$ to wektor cech, $C(i) \in \{1, \ldots, k\}$ to identyfikator skupienia, $\mathbf{m}_1, \ldots, \mathbf{m}_k$ to środek skupienia ($\mathbf{m}_l = \frac{1}{n_l} \sum_{C(i)=l} \mathbf{x}_i$, gdzie n_l to liczność l-tego skupienia), a d to odległość Euklidesowa.

Algorytm k-średnich

W metodzie k-średnich minimalizujemy

$$\sum_i d(\mathbf{x}_i, \mathbf{m}_{C(i)}),$$

gdzie $\mathbf{x}_i$ to wektor cech, $C(i) \in \{1, \ldots, k\}$ to identyfikator skupienia, $\mathbf{m}_1, \ldots, \mathbf{m}_k$ to środek skupienia ($\mathbf{m}_l = \frac{1}{n_l} \sum_{C(i)=l} \mathbf{x}_i$, gdzie n_l to liczność l-tego skupienia), a d to odległość Euklidesowa.

W metodach najszybszego spadku [1]

$$\mathbf{m}_l^{(t+1)} = \mathbf{m}_l^{(t)} + \begin{cases} \frac{1}{n_l}(\mathbf{x}_i - \mathbf{m}_l^{(t)}), & \text{gdy } l = C(i), \\ 0, & \text{wpp.} \end{cases} \quad (1)$$

Indeks Fowlkesa-Mallowsa

Definicja

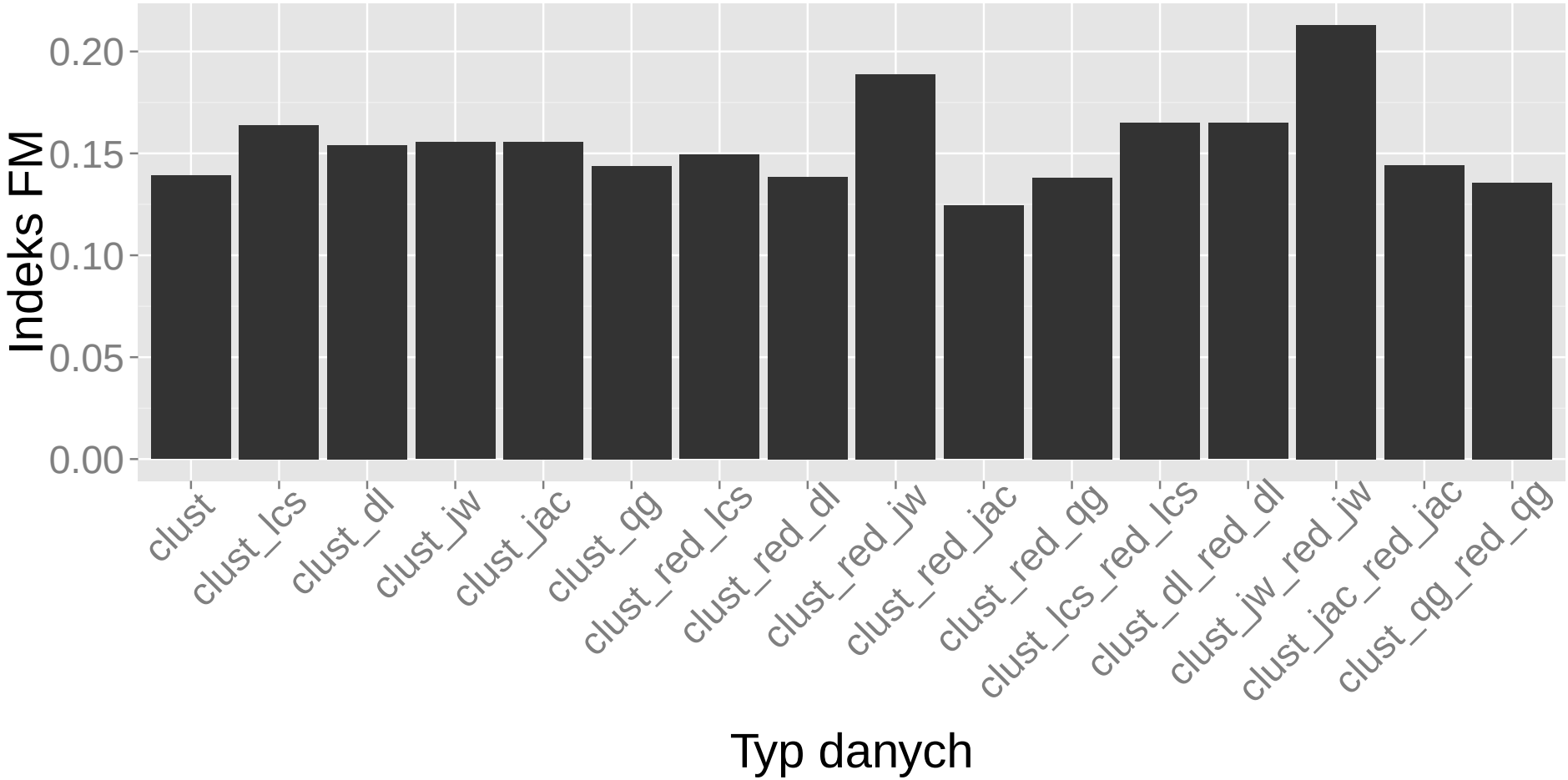
Niech macierz $M = [m_{ij}]$, $i, j = 1, \ldots, k$ oznacza liczbę elementów z próby, które należą do i-tego skupienia w K i j-tej klasy w C, a n to liczba obserwacji. Możemy wówczas zdefiniować *indeks Fowlkesa-Mallowsa*, ozn. indeks FM [3]:

$$\mathrm{FM} = \frac{T}{\sqrt{P \cdot Q}},$$

gdzie

$$T = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - n, \quad P = \sum_{i=1}^{k}\left(\sum_{j=1}^{k} m_{ij}\right)^2 - n, \quad Q = \sum_{j=1}^{k}\left(\sum_{i=1}^{k} m_{ij}\right)^2 - n.$$

Uzyskane wyniki



Rysunek : Indeks Fowlkesa-Mallowsa.

Uzyskane wyniki

clust: azja,kumak górski, lew, ptaki, salamandra plamista, traszka karpacka, wiatr, antarktyda, kultura łużycka, erytrocyt, ptolemeusz xii neos dionizos (auletes), foyer, medinet habu, ...

Uzyskane wyniki

clust: azja,kumak górski, lew, ptaki, salamandra plamista, traszka karpacka, wiatr, antarktyda, kultura łużycka, erytrocyt, ptolemeusz xii neos dionizos (auletes), foyer, medinet habu, ...
clust_qg: antoni gorecki, biblioteka, czesław miłosz, ignacy krasicki, jan kochanowski, literatura polska, literatura polska – romantyzm, literatura polska – średniowiecze, polska literatura współczesna, mikołaj rej, poezja, wisława szymborska, ...

Uzyskane wyniki

clust: azja,kumak górski, lew, ptaki, salamandra plamista, traszka karpacka, wiatr, antarktyda, kultura łużycka, erytrocyt, ptolemeusz xii neos dionizos (auletes), foyer, medinet habu, ...
clust_qg: antoni gorecki, biblioteka, czesław miłosz, ignacy krasicki, jan kochanowski, literatura polska, literatura polska – romantyzm, literatura polska – średniowiecze, polska literatura współczesna, mikołaj rej, poezja, wisława szymborska, ...
clust_jw_red_jw: armia czerwona, bitwa pod lenino, bitwa warszawska 1920, hagana, kampania wrześniowa, narodowe siły zbrojne, powstanie warszawskie, powstanie wielkopolskie, reichswehra, wehrmacht, waffen-ss, wojciech jaruzelski, ...

Wnioski

- Użycie odległości na przestrzeni ciągów znaków ma pozytywny wpływ na kategoryzację tematyczną tekstów.
- Najlepsze rezultaty uzyskano przy użyciu odległości Jaro.
- Algorytm 3 miał najlepsze wyniki.

Bibliografia

[1] Léon Bottou and Yoshua Bengio. Convergence properties of the k-means algorithms. In *Advances in Neural Information Processing Systems 7*, pages 585–592. MIT Press, 1995.

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

[3] Ethelbert B. Fowlkes and Colin L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

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