
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 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 |

## Definicja

*Odległością Levenshteina* [2] nazywamy:

$$d_{lv}(s,t) = \begin{cases} 0, & \text{gdy } s = t = \varepsilon, \\ min\{ & \\ \quad d_{lv}(s, t_{1:|t|-1}), & \\ \quad d_{lv}(s_{1:|s|-1}, t), & \\ \quad d_{\mathrm{lv}}(s_{1:|s|-1}, t_{1:|t|-1}) + \delta(s_{|s|}, t_{|t|}) & \\ \}, & \text{w przeciwnym przypadku,} \end{cases}$$

gdzie $s, t$ to napisy, a $\delta(s_i, t_j) = 0$, gdy $s_i = t_j$ i 1 w przeciwnym przypadku.

# 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.

## 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]

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

# Wnioski

- Użycie odległości na przestrzeni ciągów znaków ma pozytywny wpływ na kategoryzację tematyczną tekstów.
- Dla dwóch odległości zaobserwowano lepsze wyniki niż w przypadku pozostałych.
- Najlepsze rezultaty zostały otrzymane przy użyciu algorytmu 3.

# 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] Vladimir Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17, 1965.