

Data oddania: _____

Ocena: _____

Justyna Hubert 210200

Karol Podlewski 210294

Zadanie 2: Podsumowania lingwistyczne*

1. Cel

Celem zadania było aplikacji desktopowej, która posiada charakter doradczy, generujący pewną ilość podsumowań lingwistycznych dla podanej bazy, a następnie przedstawia użytkownikowi wybrane - według zastosowanych miar jakości wyniki, czyli podsumowania lingwistyczne.

2. Wprowadzenie

Zagadnieniem jakim zajmowaliśmy się w ramach projektu była analiza działania lingwistycznych podsumowań baz danych na zbiorach rozmytych. Zbiór rozmyty jest podstawowym pojęciem wykorzystywanym przy naszym zadaniu, zatem przytoczmy jego definicję:

Definicja 1 (Zbiór rozmyty typu I). Niech \mathcal{X} będzie zbiorem, którego elementy interesują nas w sposób bezpośredni (czyli „zbiorem zwykłym” znanym z teorii mnogości). Wówczas *zbiorem rozmytym na uniwersum \mathcal{X}* nazywamy każdy zbiór A postaci:

$$A = \bigcup_{x \in \mathcal{X}} \{(x, \mu_A(x))\},$$

gdzie $\mu_A(x) : \mathcal{X} \rightarrow [0, 1]$ nazywamy *funkcją przynależności do zbioru rozmytego A* .

* GitHub: <https://github.com/hubjust/KSR>

W celu określenia, czy element należy do zbioru wykorzystaliśmy funkcję przynależności. Określa ona stopień przynależności elementu do zbioru. W naszym projekcie skorzystaliśmy z funkcji przynależności trójkątnej oraz prostokątnej. Przytoczmy ich definicje:

Definicja 2 (Zbiór rozmyty o prostokątnej funkcji przynależności). Zbiór rozmyty A typu I na uniwersum \mathbb{R} jest *liczbą rozmytą prostokątną o parametrach a, b* wtedy i tylko wtedy, gdy $a \leq b$ oraz:

$$\mu_A(x) = \begin{cases} 0 & \text{gdy } x \in (-\infty, a], \\ 1 & \text{gdy } x \in (a, b), \\ 0 & \text{gdy } x \in [b, +\infty). \end{cases}$$

Definicja 3 (Zbiór rozmyty o trójkątnej funkcji przynależności). Zbiór rozmyty A typu I na uniwersum \mathbb{R} jest *liczbą rozmytą trójkątną o parametrach a, b, c* wtedy i tylko wtedy, gdy $a \leq b \leq c$ oraz:

$$\mu_A(x) = \begin{cases} 0 & \text{gdy } x \in (-\infty, a], \\ (x-a)/(b-a) & \text{gdy } x \in (a, b), \\ 1 & \text{gdy } x = b, \\ (c-x)/(c-b) & \text{gdy } x \in (b, c), \\ 0 & \text{gdy } x \in [c, +\infty). \end{cases}$$

Wyjaśnijmy także, czym jest lingwistyczne podsumowanie. Niech \mathcal{D} będzie bazą danych, czyli zbiorem m krotek opisujących poszczególne rekordy (każdą i -tą krotkę dla $i = 1 \dots m$ będziemy oznaczać jako d_i). Każda kolumna opisuje cechę pewnego typu. Cechę krotki możemy nazwać *zmienną lingwistyczną*, a jej wartością może być konkretna liczba, ale może być to także rozmyta klasa (np. mało/trochę/dużo/sporo). Niech P będzie zbiorem rozmytym na uniwersum krotek bazy danych \mathcal{D} , który zawiera w sobie krotki posiadające pewną własność. Taki zbiór może nam posłużyć jako podmiot podsumowania lingwistycznego (np. mężczyźni, kobiety, samochody, zawodnicy). Bardzo ważnym elementem, wykorzystywanym we wszystkich rodzajach podsumowań lingwistycznych, jest kwantyfikator oznaczany jako Q . Przykładami kwantyfikatorów mogą być: "około 10", "ponad 70" (kwantyfikatory absolutne - zbiory rozmyte na uniwersum \mathbb{R}) lub "większość", "znikoma część" (kwantyfikatory relatywne - zbiory rozmyte na uniwersum $[0, 1]$). Istotny dla nas będzie stopień przynależności P do Q . Zdefiniujmy także sumaryzator S_j . Jest to zbiór rozmyty na zbiorze wartości przyjmowanych przez j -tą kolumnę bazy danych. Np. gdyby krotki dotyczyły różnych osób, a jedną ze zmiennych lingwistycznych były ich zarobki, to sumaryzatory mogłyby mieć postać "zarabia za mało", "zarabia ponad 6400 zł" itp.

Wykorzystując powyższe elementy można skonstruować **lingwistyczne podsumowanie bazy danych wg. Yagera**, czyli:

$$Q \text{ } P \text{ jest/są } S_j \text{ } [T] \text{ ,}$$

gdzie T to stopień prawdziwości podsumowania.

W celu rozszerzenia podsumowania lingwistycznego należy skorzystać ze złożonego sumaryzatora. Sumę sumaryzatorów można w podsumowaniu lingwistycznym zapisać za pomocą słowa "lub", zaś iloczyn za pomocą słowa "i". W rezultacie **podsumowanie ze złożonym sumaryzatorem** może mieć postać:

$$Q \text{ } P \text{ jest/są } S_1 \text{ i/lub } S_2 \text{ i/lub } \dots \text{ i/lub } S_n \text{ } [T] \text{ .}$$

Innym sposobem rozszerzenia pojęcia podsumowań jest zastosowanie kwalifikatora. Kwalifikator W jest zbiorem rozmytym na \mathcal{D} , który opisuje w jakim stopniu krotki posiadają pewną własność. Typowe przykłady to "[osoby] które są bezrobotne", "[osoby] które są dziećmi". **Podsumowanie z kwalifikatorem** ma postać:

$$Q \text{ } P \text{ mających własność } W \text{ ma własność } S_j \text{ } [T] \text{ .}$$

Aby określić jakość naszych podsumowań zaimplementowaliśmy poniższe miary jakości:

2.1. T_1 – stopień prawdziwości

Najbardziej naturalną miarą jakości podsumowania lingwistycznego jest wprowadzony przez Yagera stopień prawdziwości. Abstrahując od tego, czy sumaryzator S_j jest wynikiem złożenia wielu sumaryzatorów czy nie, możemy obliczyć sumę przynależności wszystkich rozważanych krotek do niego:

$$r = \sum_{i=1}^m \mu_{ce(S_j)}(d_i) \text{ ,}$$

gdzie $ce(S_j)$ jest rozszerzeniem cylindrycznym sumaryzatora S_j , a m liczba wszystkich krotek. Zatem dla kwantyfikatorów relatywnych stopień prawdziwości możemy zapisać jako $T_1 = \mu_Q(\frac{r}{m})$, zaś dla kwantyfikatorów absolutnych jako $T_1 = \mu_Q(r)$.

2.2. T_2 – stopień nieprecyzyjności

Dla podsumowania z n sumaryzatorami $S_1 \dots S_n$ (w szczególności dla $n = 1$) możemy określić miarę jakości podsumowania określaną jako stopień nieprecyzyjności, definiowaną następującym wzorem:

$$T_2 = 1 - \left(\prod_{j=1}^n \text{in}(S_j) \right)^{1/n} \text{ .}$$

Wyrażenie $\left(\prod_{j=1}^n \text{in}(S_j) \right)^{1/n}$ to po prostu średnia geometryczna ze stopni rozmycia wykorzystanych sumaryzatorów.

2.3. T_3 – stopień pokrycia

Stopień pokrycia T_3 jest zdefiniowany dla podsumowań z kwalifikatorami, czyli tymi opartymi na drugiej formie ilościowych wyrażeń lingwistycznych. Stopień pokrycia T_3 Dla każdego $i = 1 \dots m$ (związanego z krotką d_i z bazy danych) możemy zdefiniować:

$$t_i = \begin{cases} 1 & \text{gdy } \mu_{ce(S_j)}(d_i) > 0 \wedge \mu_W(d_i) > 0 \\ 0 & \text{w przeciwnym wypadku.} \end{cases}$$
$$h_i = \begin{cases} 1 & \text{gdy } \mu_W(d_i) > 0 \\ 0 & \text{w przeciwnym wypadku.} \end{cases}$$

Przy powyższych oznaczeniach:

$$T_3 = \frac{\sum_{i=1}^m t_i}{\sum_{i=1}^m h_i} .$$

2.4. T_4 – stopień stosowności

Dla podsumowania z n sumaryzatorami $S_1 \dots S_n$ oraz m krotkami w bazie danych możemy wprowadzić oznaczenia:

$$g_{ij} = \begin{cases} 1 & \text{gdy } \mu_{ce(S_j)}(d_i) > 0 \\ 0 & \text{w przeciwnym wypadku.} \end{cases}$$

oraz

$$r_j = \frac{\sum_{i=1}^m g_{ij}}{m} .$$

Wówczas możemy zapisać:

$$T_4 = \left| \left(\prod_{j=1}^n r_j \right) - T_3 \right| .$$

2.5. T_5 – długość podsumowania

Dla podsumowania z n sumaryzatorami $S_1 \dots S_n$ miarę długości podsumowania definiujemy jako:

$$T_5 = \left(\frac{1}{2} \right)^{n-1} .$$

2.6. T_6 – stopień nieprecyzyjności kwantyfikatora

Nazwy miar T_6 – T_{10} można określić jako samoopisujące się, więc przedstawienie wzoru wyczerpuje zagadnienie w stopniu wystarczającym dla niniejszego sprawozdania:

$$T_6 = 1 - \ln(Q) .$$

2.7. T_7 – stopień liczności kwantyfikatora

W przeciwieństwie do T_6 , zamiast zliczać elementy z nośnika Q , policzymy moc zbioru rozmytego:

$$T_7 = 1 - \frac{(Q)}{(\mathcal{X}_Q)} .$$

2.8. T_8 – stopień liczności sumaryzatora

Możemy mieć do czynienia z sumaryzatorem złożonym; wtedy, podobnie jak przy poprzednich miarach, stosujemy średnią geometryczną. Dla podsumowania z n sumaryzatorami $S_1 \dots S_n$:

$$T_8 = 1 - \left(\prod_{j=1}^n \frac{(S_j)}{(\mathcal{X}_j)} \right) .$$

2.9. T_9 – stopień nieprecyzyjności kwalifikatora

Stopień precyzji kwalifikatora T_9 jest oparty na drugiej formie podsumowań tzn.: Q obiektów będących/mających W jest/ma S , gdzie W jest reprezentowane przez zbiór rozmyty i jest kwalifikatorem. Definicja tej miary jest następująca:

$$T_9 = 1 - \text{in}(W) .$$

2.10. T_{10} – stopień liczności kwalifikatora

Stopień kardynalności kwalifikatora T_{10} definiujemy jako:

$$T_{10} = 1 - \frac{(W)}{(\mathcal{X}_g)} .$$

gdzie miara $()$ jest dana tak jak we wcześniejszych sekcjach.

2.11. T_{11} – długość kwalifikatora

Długość kwalifikatora T_{11} definiujemy następująco:

$$T_{11} = 2 \left(\frac{1}{2} \right)^{(W)} .$$

gdzie (W) jest ilością zbiorów rozmytych, z których kwalifikator jest skomponowany. Im bardziej złożony kwalifikator tym jakość podsumowania gorsza.

3. Opis implementacji

Program został stworzony w języku C#. Graficzny interfejs użytkownika został stworzony przy wykorzystaniu Windows Presentation Foundation. Logika aplikacji została odseparowana od GUI, w zgodzie ze wzorcem projektowym Model-view-viewmodel (MVVM), poprzez implementację trzech projektów (Logic, ViewModel i GUI).

3.1. Logic

4. Materiały i metody

Do przeprowadzenia badań i generowania konkretnych podsumowań wykorzystaliśmy bazę danych dotyczącą przechowującą statystyki piłkarzy z gry Fifa 2019. Składa się ona z 15397 krotek znajdujących się w tabeli z 20 różnymi kolumnami - w ramach naszego projektu skorzystaliśmy tylko z 13. Przedstawiamy je poniżej:

- Wiek
- Wzrost
- Waga
- Tempo
- Przyspieszenie
- Prędkość
- Dribbling
- Zręczność
- Balans
- Reakcje
- Kontrola piłki
- Opanowanie
- Precyzja
- Ustawienie się

Każda z ww. kolumn jest typem całkowitym.

5. Wyniki

Poniższej umieszczone tabele oraz wykresy są wynikami przeprowadzonych przez nas eksperymentów.

5.1. Wpływ liczby k sąsiadów oraz wyboru metryki na klasyfikację

k	places [%]	topics [%]	authors [%]
2	74.4	53.7	43.9
3	78.5	52.2	43.9
5	80.2	52.2	36.6
7	81.0	53.7	26.8
10	81.5	60.4	24.4
15	81.6	62.7	29.3
20	81.4	61.2	31.7

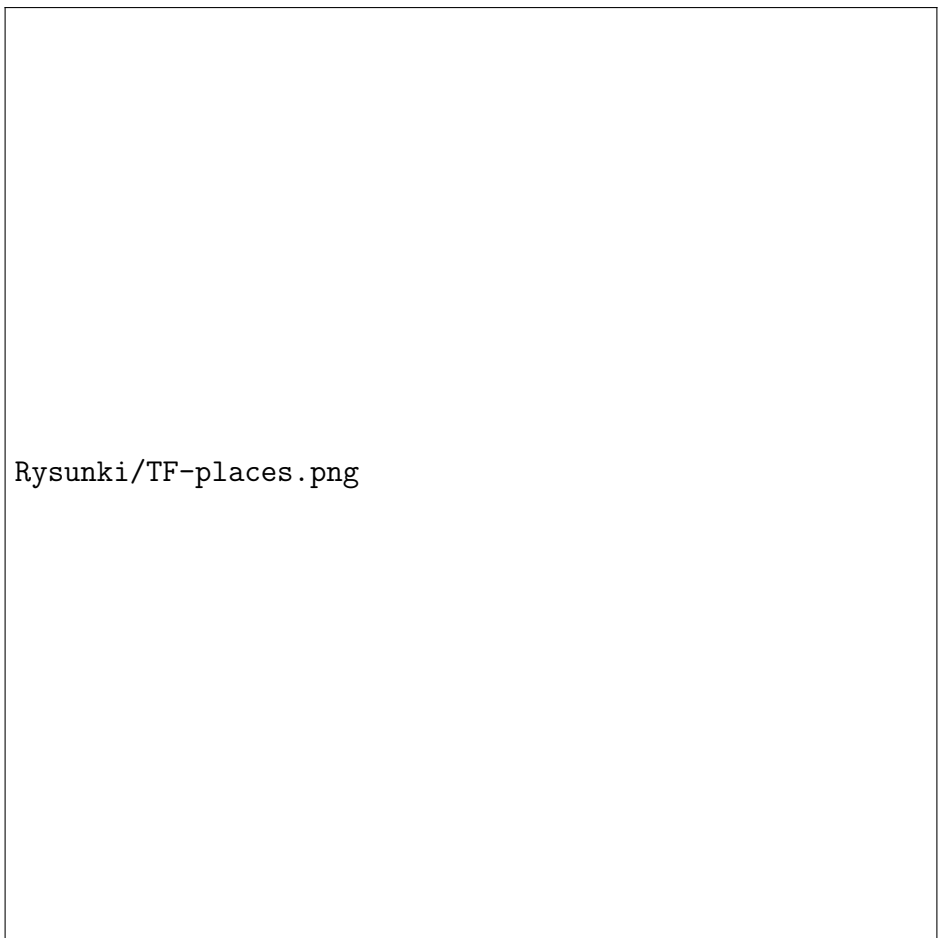
Tabela 1. Skuteczność klasyfikacji dla metryki Euklidesowej dla pierwszego sposobu ekstrakcji

k	places [%]	topics [%]	authors [%]
2	75.4	56.7	36.6
3	79.4	56.7	39.0
5	81.0	61.2	36.6
7	81.3	59.0	31.7
10	81.9	64.9	24.4
15	82.0	64.9	29.3
20	82.1	63.4	29.3

Tabela 2. Skuteczność klasyfikacji dla metryki ulicznej dla pierwszego sposobu ekstrakcji

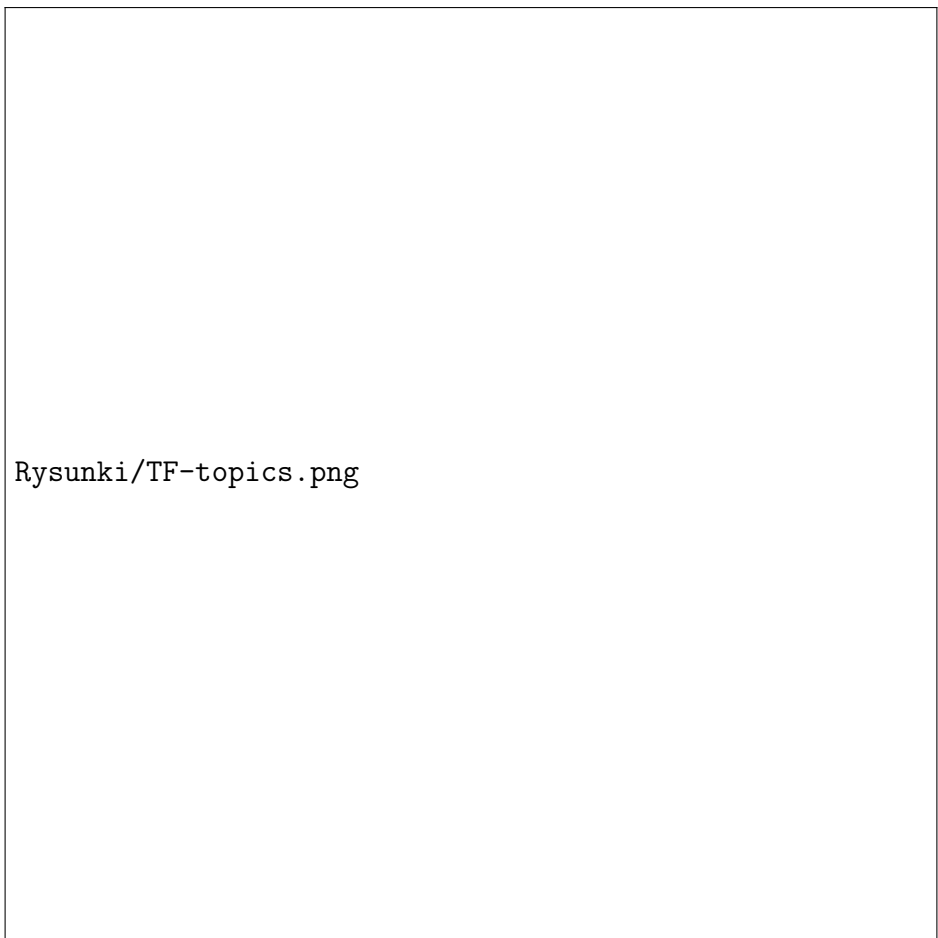
k	places [%]	topics [%]	authors [%]
2	80.3	14.9	17.1
3	80.3	44.0	17.1
5	80.3	44.0	17.1
7	80.3	44.0	17.1
10	80.3	44.0	17.1
15	80.3	44.0	17.1
20	80.3	44.0	17.1

Tabela 3. Skuteczność klasyfikacji dla metryki Czebyszewa dla pierwszego sposobu ekstrakcji



Rysunki/TF-places.png

Rysunek 1. Dane z Tabel 1-3 dla kategorii places



Rysunki/TF-topics.png

Rysunek 2. Dane z Tabel 1-3 dla kategorii topics

Rysunki/TF-authors.png

Rysunek 3. Dane z Tabel 1-3 dla kategorii authors (własne teksty)

k	places [%]	topics [%]	authors [%]
2	79.0	63.4	22.0
3	82.0	64.2	19.5
5	82.1	59.0	29.3
7	83.3	62.1	22.0
10	82.0	64.9	26.8
15	81.9	67.9	24.4
20	81.1	67.1	17.1

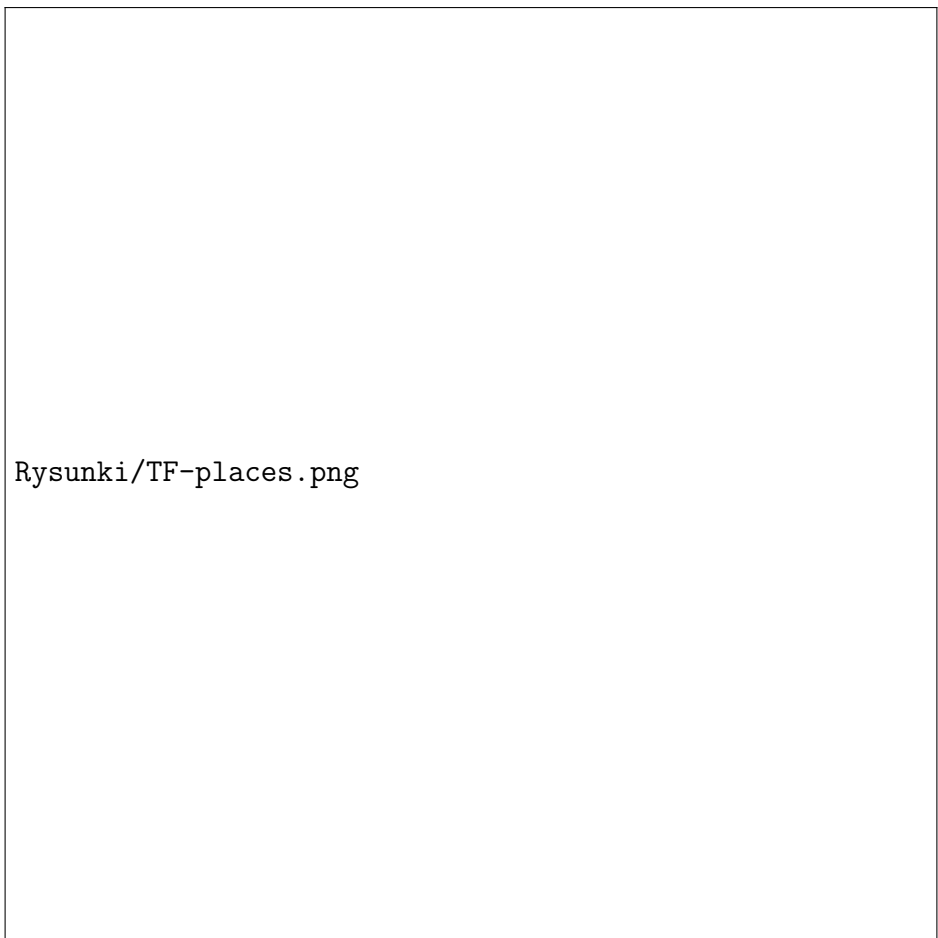
Tabela 4. Skuteczność klasyfikacji dla metryki Euklidesowej dla drugiego sposobu ekstrakcji

k	places [%]	topics [%]	authors [%]
2	80.2	59.7	22.0
3	82.4	65.7	19.5
5	82.6	67.2	29.3
7	83.3	67.2	22.0
10	82.6	67.2	26.8
15	82.1	67.2	24.4
20	81.6	67.9	17.1

Tabela 5. Skuteczność klasyfikacji dla metryki ulicznej dla drugiego sposobu ekstrakcji

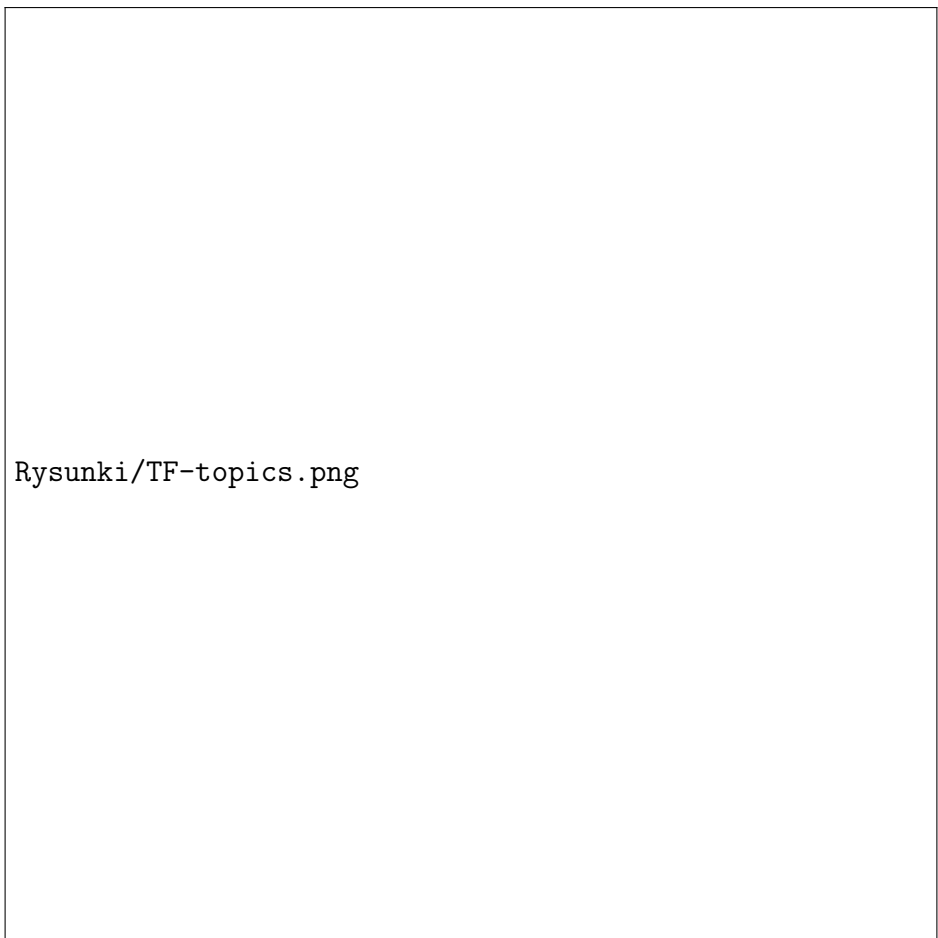
k	places [%]	topics [%]	authors [%]
2	77.0	14.9	17.1
3	77.0	44.0	17.1
5	77.0	44.0	17.1
7	77.0	44.0	17.1
10	77.0	44.0	17.1
15	77.0	44.0	17.1
20	77.0	44.0	17.1

Tabela 6. Skuteczność klasyfikacji dla metryki Czebyszewa dla drugiego sposobu ekstrakcji



Rysunki/TF-places.png

Rysunek 4. Dane z Tabel 4-6 dla kategorii places



Rysunki/TF-topics.png

Rysunek 5. Dane z Tabel 4-6 dla kategorii topics

Rysunki/TF-authors.png

Rysunek 6. Dane z Tabel 4-6 dla kategorii authors (własne teksty)

k	places [%]	topics [%]	authors [%]
2	69.5	47.8	44.8
3	75.3	53	53.7
5	78.3	47	48.8
7	79.4	48.5	63.4
10	80.2	49.3	63.4
15	80.5	47.8	58.5
20	80.7	44.8	53.7

Tabela 7. Skuteczność klasyfikacji dla metryki Euklidesowej dla trzeciego sposobu ekstrakcji

k	places [%]	topics [%]	authors [%]
2	69	49.3	56.1
3	75.1	47	56.1
5	78.2	47	48.8
7	79.3	46.3	58.5
10	80	51.5	61
15	80.5	45.5	58.5
20	80.7	45.5	56.1

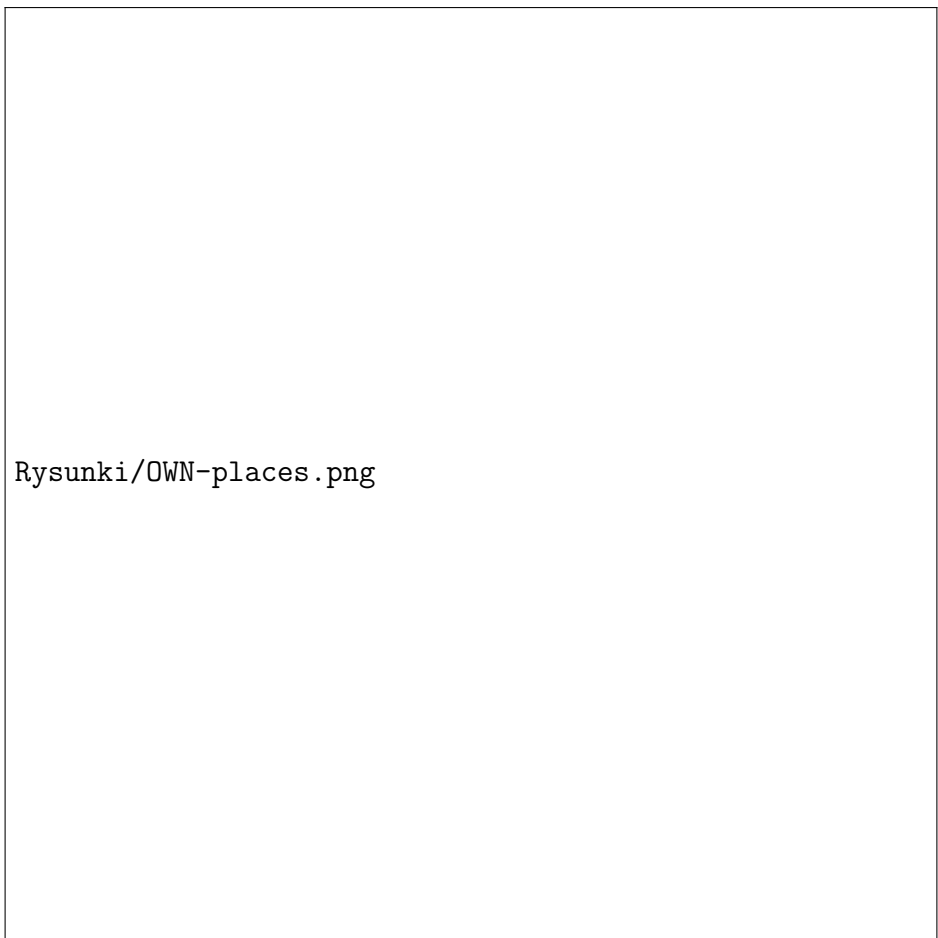
Tabela 8. Skuteczność klasyfikacji dla metryki ulicznej dla trzeciego sposobu ekstrakcji

k	places [%]	topics [%]	authors [%]
2	80.3	14.9	17.1
3	80.4	44	17.1
5	80.5	44	17.1
7	80.6	44	17.1
10	80.7	44	17.1
15	80.8	44	17.1
20	80.9	44	17.1

Możliwość wniesienia i

instalacji sprzętu

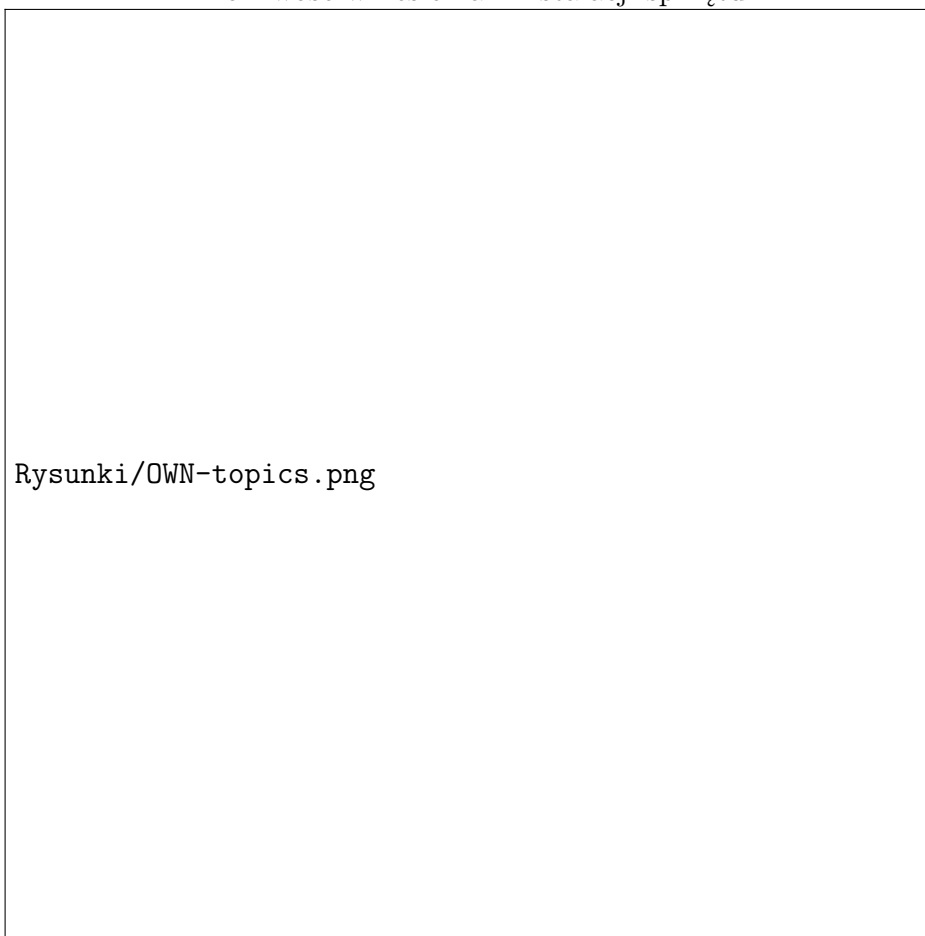
Tabela 9. Skuteczność klasyfikacji dla metryki Czebyszewa dla trzeciego sposobu ekstrakcji



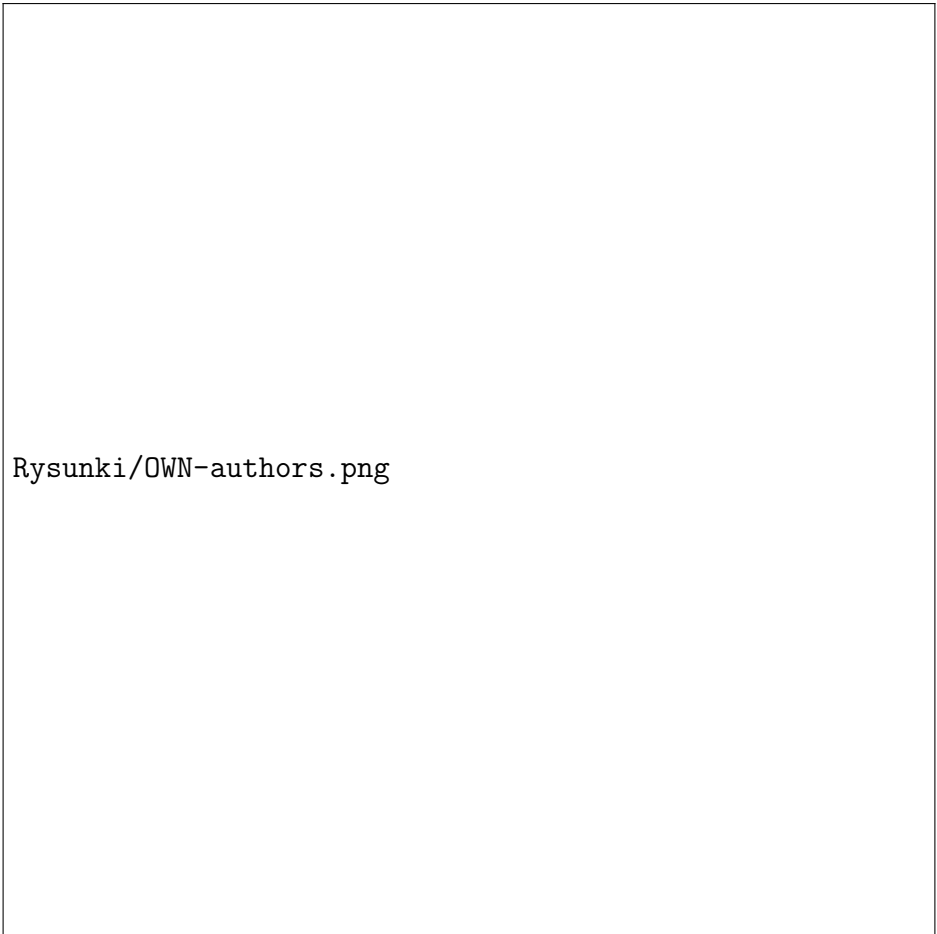
Rysunki/OWN-places.png

Rysunek 7. Dane z Tabel 7-9 dla kategorii places

Możliwość wniesienia i instalacji sprzętu



Rysunek 8. Dane z Tabel 7-9 dla kategorii topics



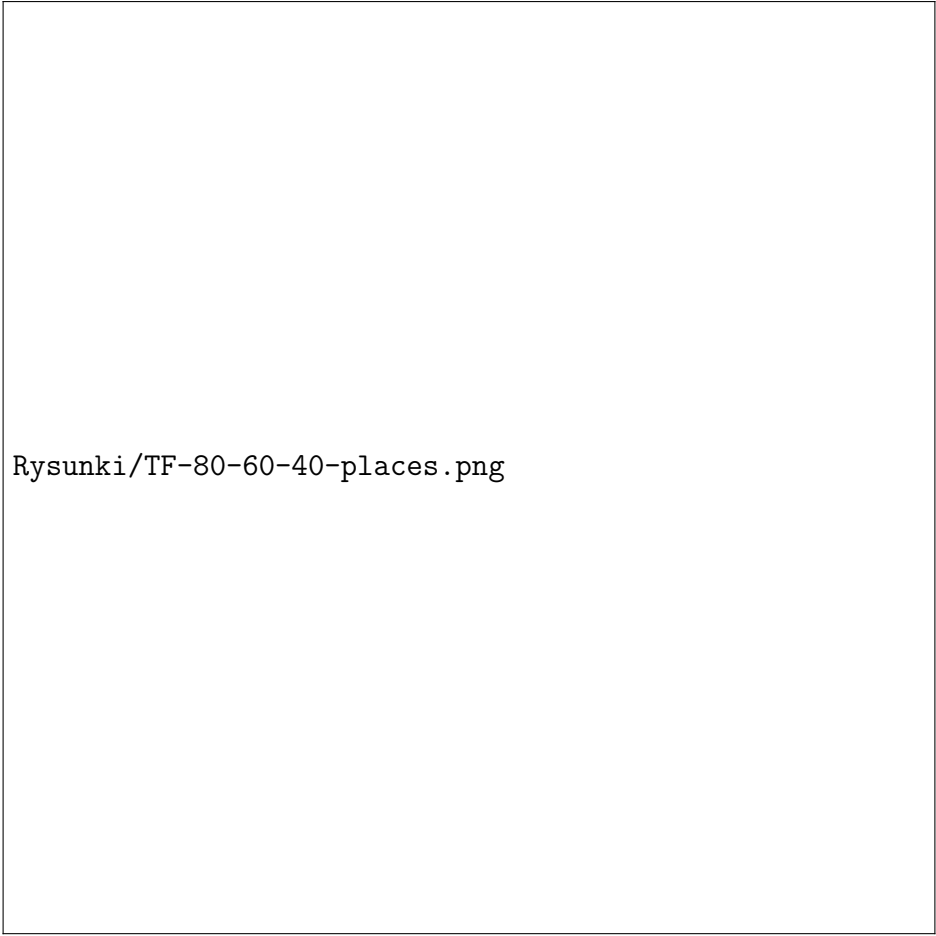
Rysunki/OWN-authors.png

Rysunek 9. Dane z Tabel 7-9 dla kategorii authors (własne teksty)

5.2. Wpływ podziału tekstów na zbiory treningowe i testowe na klasyfikację

k	80%	60%	40%
5	82.7	80.2	78.4
7	83.3	81.0	79.9
10	84.2	81.5	80.4
15	83.9	81.6	80.4

Tabela 10. Skuteczność klasyfikacji dla pierwszego sposobu ekstrakcji, dla kategorii places

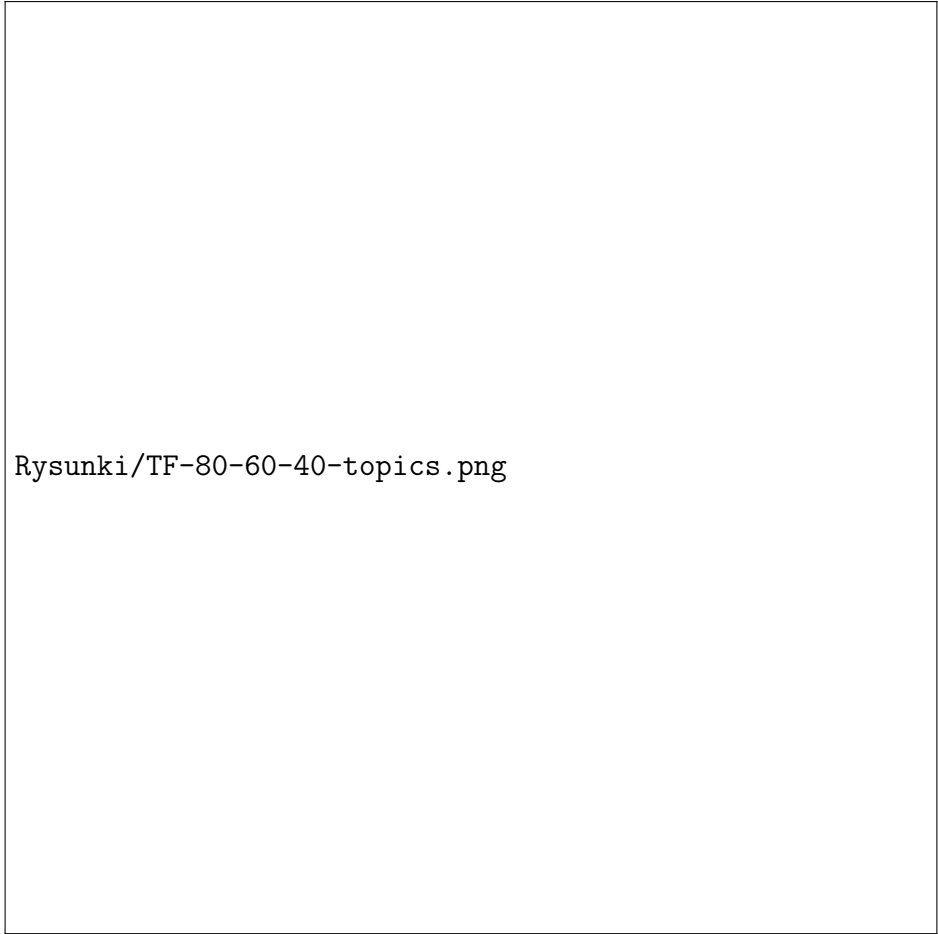


Rysunki/TF-80-60-40-places.png

Rysunek 10. Skuteczność klasyfikacji dla pierwszego sposobu ekstrakcji, dla kategorii places

k	80%	60%	40%
7	56.7	53.7	62.7
10	59.7	60.4	62.2
15	58.2	62.7	64.7
20	62.7	61.2	64.7

Tabela 11. Skuteczność klasyfikacji dla pierwszego sposobu ekstrakcji, dla kategorii topics

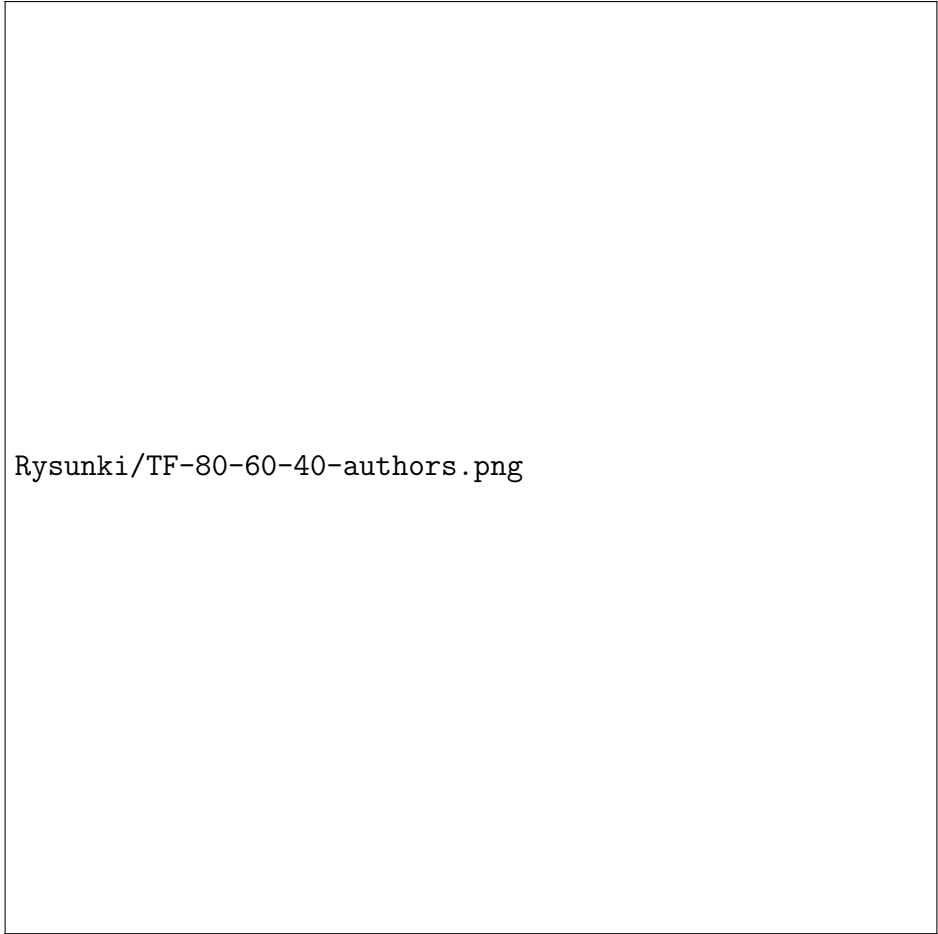


Rysunki/TF-80-60-40-topics.png

Rysunek 11. Skuteczność klasyfikacji dla pierwszego sposobu ekstrakcji, dla kategorii topics

k	80%	60%	40%
2	19.0	43.9	38.7
3	19.0	43.9	37.1
5	23.8	36.6	29.0
7	14.3	26.8	32.3

Tabela 12. Skuteczność klasyfikacji dla pierwszego sposobu ekstrakcji, dla kategorii authors

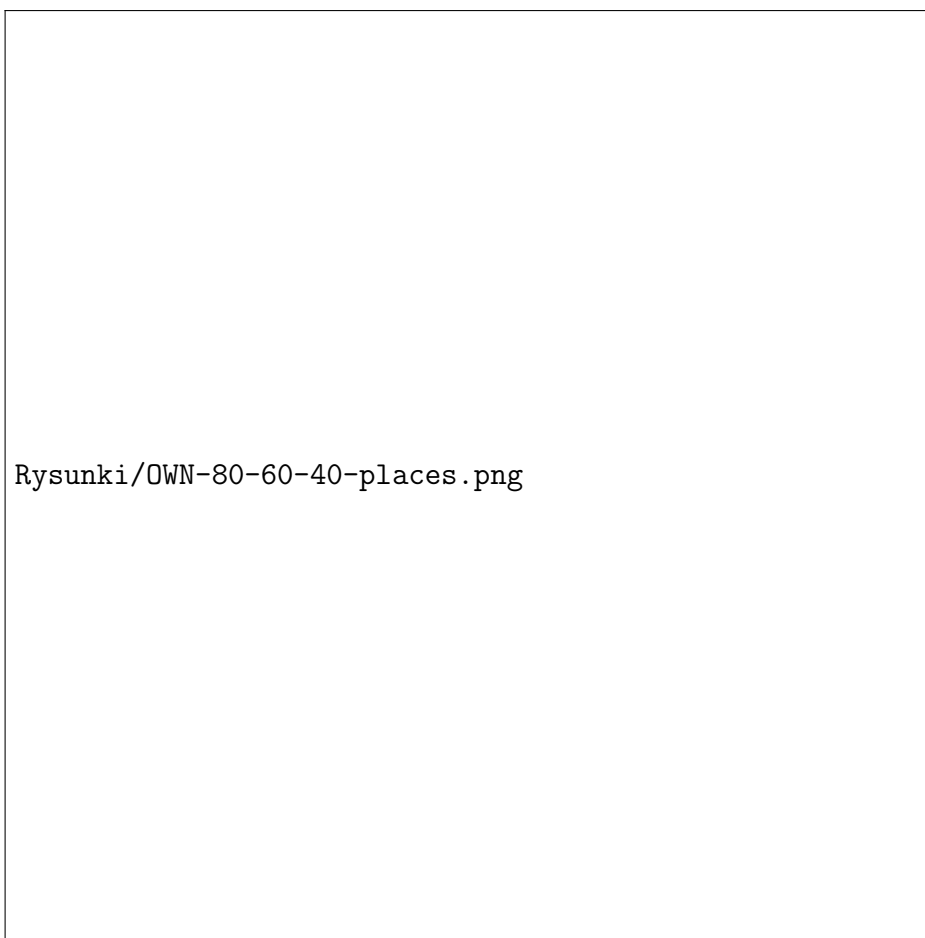


Rysunki/TF-80-60-40-authors.png

Rysunek 12. Skuteczność klasyfikacji dla pierwszego sposobu ekstrakcji, dla kategorii authors

k	80%	60%	40%
7	82.6	79.4	78.5
10	83.1	80.2	79.2
15	83.5	80.5	79.6
20	83.3	80.7	79.7

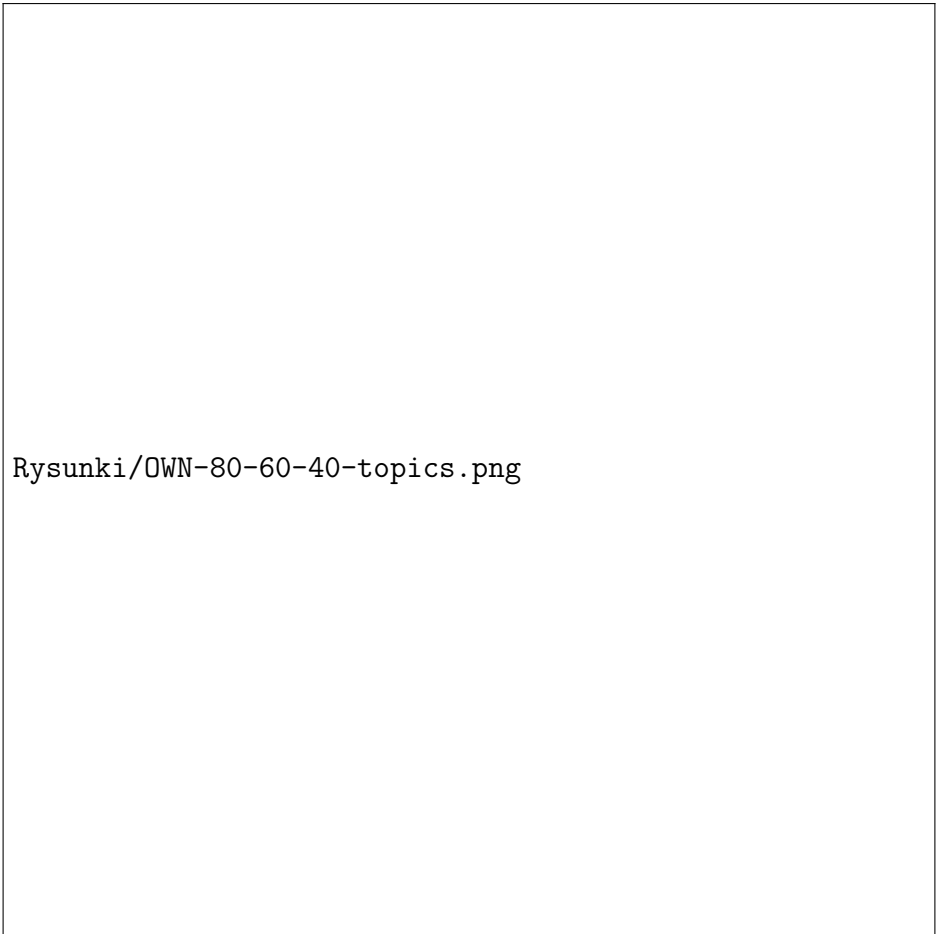
Tabela 13. Skuteczność klasyfikacji dla drugiego sposobu ekstrakcji, dla kategorii places



Rysunek 13. Skuteczność klasyfikacji dla drugiego sposobu ekstrakcji, dla kategorii places

k	80%	60%	40%
3	55.2	53.0	43.3
5	55.2	47.0	43.8
7	55.2	48.5	42.3
10	55.2	49.3	40.8

Tabela 14. Skuteczność klasyfikacji dla trzeciego sposobu ekstrakcji, dla kategorii topics

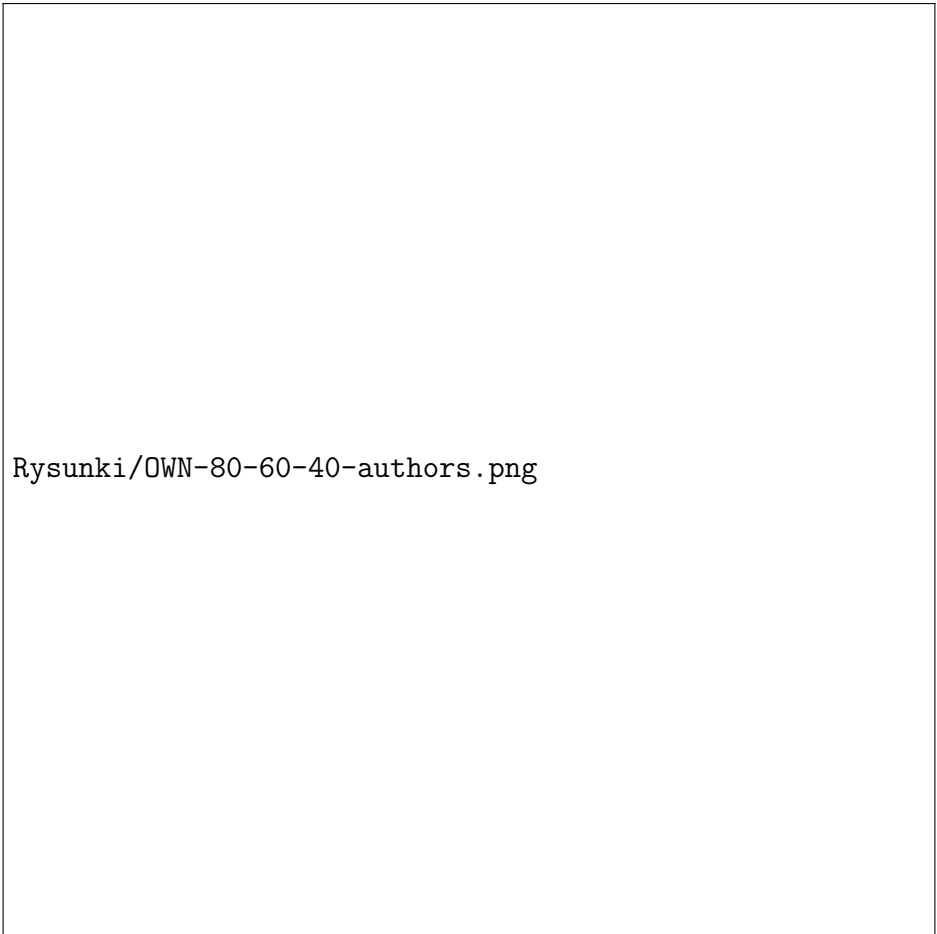


Rysunki/OWN-80-60-40-topics.png

Rysunek 14. Skuteczność klasyfikacji dla trzeciego sposobu ekstrakcji, dla kategorii topics

k	80%	60%	40%
7	71.4	63.4	46.8
10	66.7	63.4	48.4
15	71.4	58.5	50.0
20	71.4	53.7	33.9

Tabela 15. Skuteczność klasyfikacji dla trzeciego sposobu ekstrakcji, dla kategorii authors



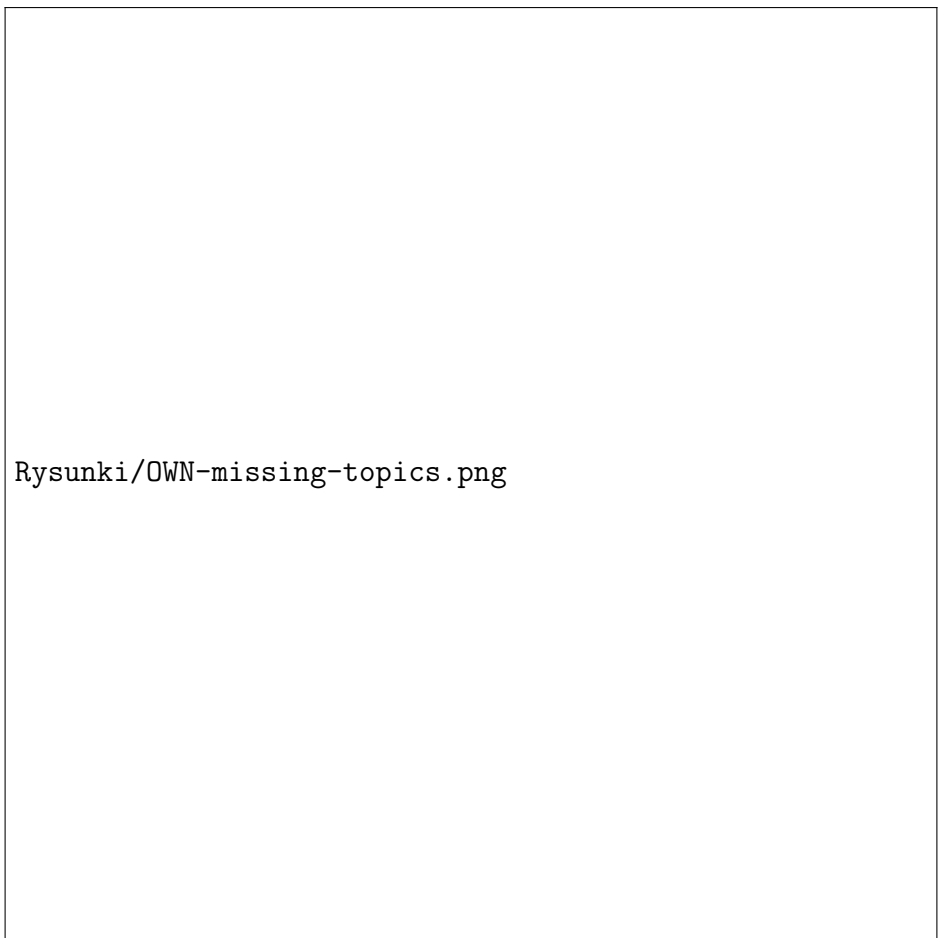
Rysunki/OWN-80-60-40-authors.png

Rysunek 15. Skuteczność klasyfikacji dla trzeciego sposobu ekstrakcji, dla kategorii authors

5.3. Wpływ konkretnych cech na klasyfikację

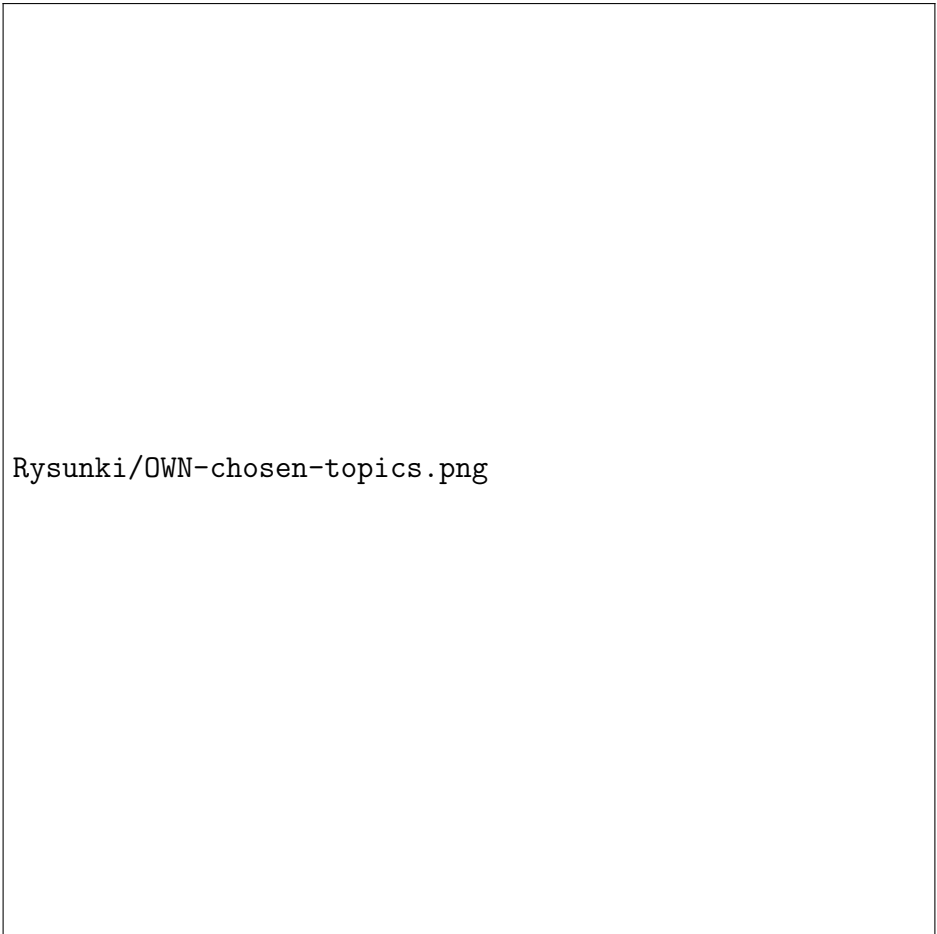
Na wykresach widoczne są następujące oznaczenia:

- c_1) Liczba słów,
- c_2) Liczba słów, których długość nie przekracza 3 znaków,
- c_3) Liczba słów, których długość zawiera się w zakresie 4-7 znaków,
- c_4) Liczba słów, których długość przekracza 8 znaków,
- c_5) Liczba unikalnych słów,
- c_6) Liczba słów napisanych wielką literą,
- c_7) liczba słów rozpoczynających się wielką literą.



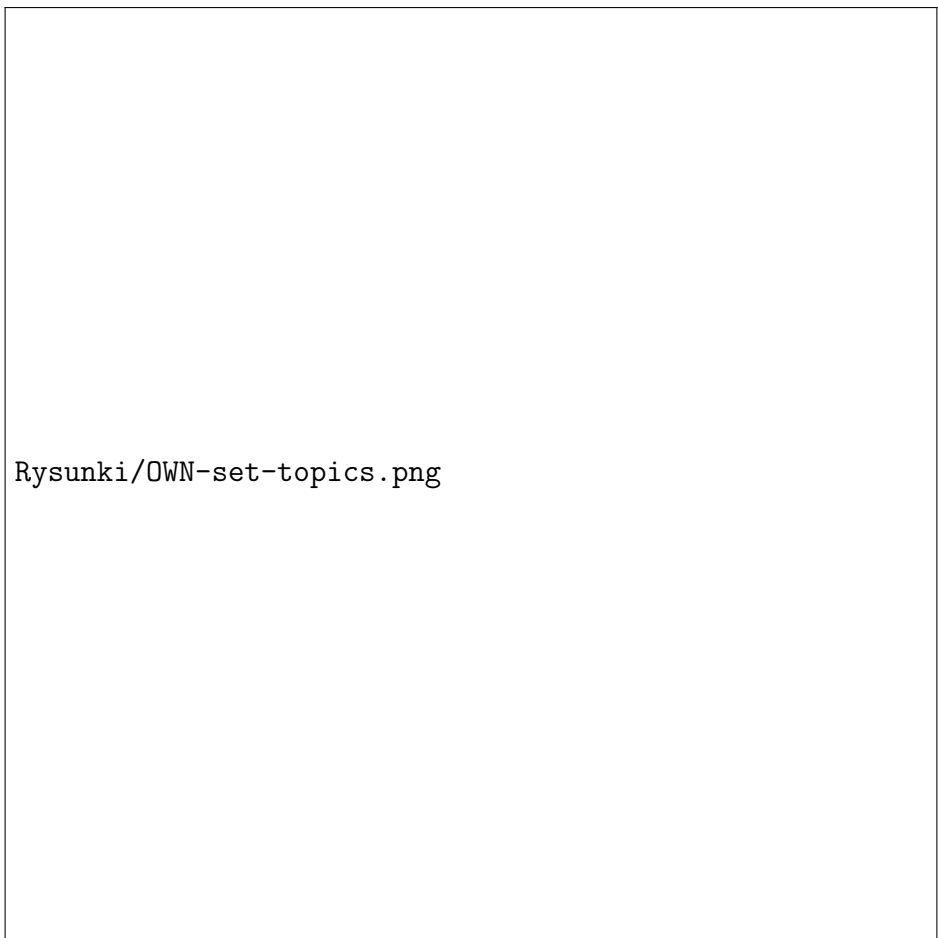
Rysunki/OWN-missing-topics.png

Rysunek 16. Skuteczność klasyfikacji dla brakujących cech, dla kategorii topics



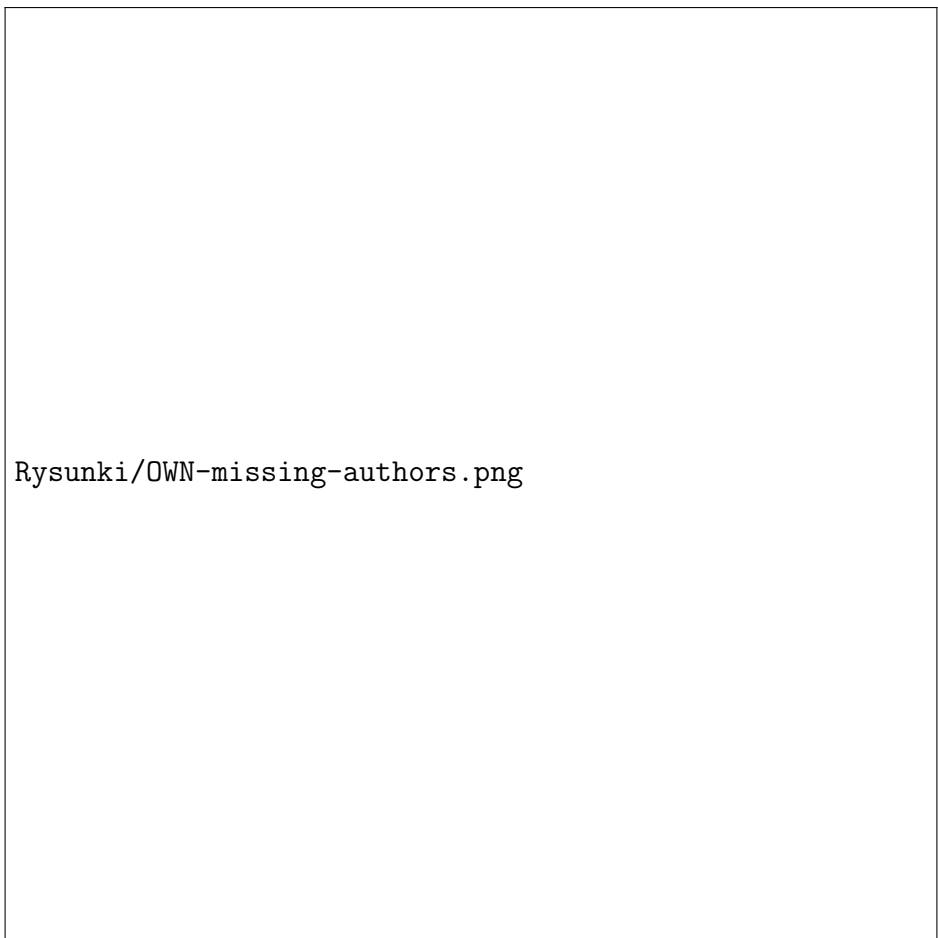
Rysunki/OWN-chosen-topics.png

Rysunek 17. Skuteczność klasyfikacji dla wybranych cech, dla kategorii topics



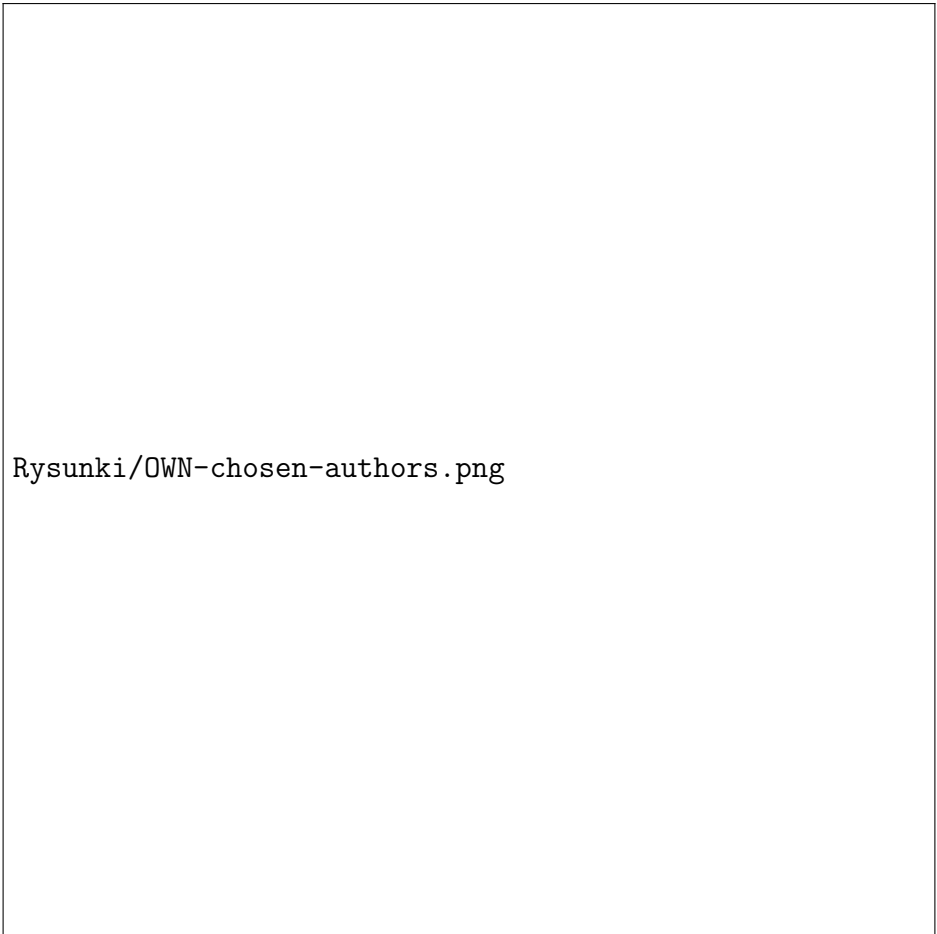
Rysunki/OWN-set-topics.png

Rysunek 18. Skuteczność klasyfikacji dla zestawu cech, dla kategorii topics



Rysunki/OWN-missing-authors.png

Rysunek 19. Skuteczność klasyfikacji dla brakujących cech, dla kategorii authors



Rysunki/OWN-chosen-authors.png

Rysunek 20. Skuteczność klasyfikacji dla wybranych cech, dla kategorii authors

Rysunki/OWN-set-authors.png

Rysunek 21. Skuteczność klasyfikacji dla zestawu cech, dla kategorii authors

5.4. Najlepsze wyniki

Kategoria	Skuteczność	Metryka	Ekstrakcja	k
Places	83.3%	Euklidesowa	IDF	7
Places	83.3%	Uliczna	IDF	7
Topics	67.9%	Euklidesowa	IDF	15
Topics	67.9%	Uliczna	IDF	20
Authors	43.9%	Euklidesowa	TF	2, 3

Tabela 16. Tabela przedstawiająca najlepsze wyniki z pierwszego eksperymentu (4.1)

Kategoria	Skuteczność	Zb. treningowy	Ekstrakcja	k
Topics	64.73%	40%	TF	15, 20
Authors	43.9%	60%	TF	2, 3
Topics	55.2%	80%	Własne cechy	3-10
Authors	71.4%	80%	Własne cechy	7, 15, 20

Tabela 17. Tabela przedstawiająca najlepsze wyniki z drugiego eksperymentu (4.2)

Kategoria	Skuteczność	Wykorzystane cechy
Topics	62,7%	c_3, c_4, c_5, c_6, c_7
Authors	76,2%	c_3, c_4, c_5, c_6, c_7

Tabela 18. Tabela przedstawiająca najlepsze wyniki z trzeciego eksperymentu (4.3)

6. Dyskusja

6.1. Wpływ liczby k sąsiadów oraz wyboru metryki na klasyfikację

W przypadku wszystkich trzech sposobów ekstrakcji, metryka Euklidesowa oraz metryka uliczna osiągają bardzo podobne wyniki i nie jesteśmy w stanie stwierdzić, która z nich wykazuje lepszą skuteczność. Metryka Czebyszewa charakteryzuje się zdecydowanie słabszą zdolnością do klasyfikacji. Osiąga niższe wyniki, niż dwie wcześniej wspomniane metryki.

W przypadku pierwszego i drugiego sposobu ekstrakcji cech dla kategorii topics i places, zauważyliśmy, że wraz ze wzrostem liczby k sąsiadów zwiększa się także skuteczność. Najsłabsze wyniki osiągane były dla k równego 2. Jeśli zaś chodzi o kategorię authors, najwyższa skuteczność wykazywała mała liczba k sąsiadów (od 2 do 3). Wyraźny spadek wyników zaobserwowaliśmy, gdy k równało się 10. Podczas eksperymentu trzeciego sposobu ekstrakcji cech zauważyliśmy bardzo zmienną skuteczność w przypadku zmiany liczby k sąsiadów w zależności od wybranych kategorii. Kategoria places osiąga najslabsze wyniki przy małej liczbie sąsiadów, z kolei kategoria topics najlepsze. Zauważyliśmy, że najwyższe wyniki w kategorii authors osiągane są przy liczbie sąsiadów równej 7 oraz 10.

6.2. Wpływ podziału tekstów na zbiory treningowe i testowe na klasyfikację

W przeważającej większości najwyższe wyniki osiągane były przy 80% zbioru treningowego. Tylko w jednym przypadku użycie 40% zbioru treningowego pozwoliło osiągnąć najwyższą skuteczności (pierwszy sposób ekstrakcji, kategoria topics). Zazwyczaj jednak ten dobór procentowy okazywał się być najslabszym ze względu na niedouczenie.

6.3. Wpływ konkretnych cech na klasyfikację

Podczas klasyfikacji dla kategorii topics, zauważyliśmy, że liczba słów oraz liczba słów, których długość nie przekracza 3 znaków mają negatywny wpływ na osiąganą skuteczność. Świadczyć może o tym fakt, iż bez ww. cech osiągnęliśmy najwyższą skuteczność. Dużo ważniejsze okazały się cechy związane z unikalnością słów oraz wielkimi literami.

Podczas klasyfikacji dla kategorii authors najważniejsza okazała się cecha odpowiadająca za liczbę unikalnych słów. Bez niej skuteczność spadła z 71% na 47%. Podobnie jak w przypadku kategorii topics, cechy sprawdzające liczbę słów oraz liczbę krótkich słów osłabiały nasze wyniki - dzięki wyłączeniu ich, uzyskaliśmy wyższe wyniki niż w przypadku wszystkich cech.

7. Wnioski

- Liczba k sąsiadów ma spory wpływ na skuteczność klasyfikacji, jednak nie ma jednej, optymalnej wartości - zmiana metryki, podziału zbiorów czy klasyfikowanych kategorii może spowodować obniżenie wyników dla stałego k .
- Dla mniejszych zbiorów tekstowych lepiej sprawdzają się mniejsze wartości k sąsiadów, dla większych - wyższe wartości.
- Metryka Czebyszewa nie powinna być wykorzystywana w klasyfikacji tekstów, gdyż osiąga bardzo słabe wyniki.
- Istotny jest podział tekstów na zbiory testowe oraz treningowe. W przypadku zbyt małego zbioru treningowego osiągamy zjawisko niedouczenia, w przypadku zbyt dużego - przeuczenia.
- Cechy odpowiedzialne za liczbę słów oraz liczbę krótkich słów (do 3 znaków) nie sprawdzają się przy klasyfikacji tekstów.
- Wektor cech powinien się składać z przynajmniej kilku cech, żeby osiągnąć większą skuteczność.

Literatura

- [1] Methods for the linguistic summarization of data - applications of fuzzy sets and their extensions, Adam Niewiadomski, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008
- [2] <http://home.agh.edu.pl/~horzyk/lectures/miw/KNN.pdf>
- [3] <https://github.com/hklemp/dotnet-stop-words>
- [4] <http://snowball.tartarus.org/algorithms/english/stemmer.html>