Informatyka, studia niestacjonarne, I st. semestr VI

Komputerowe systemy rozpoznawania **2019**

Prowadzący: dr hab. inż. Adam Niewiadomski Niedziela, 12:00

Data oddania: _____ Ocena: _____

Konrad Jachimstal 211807

Patryk Janicki 211951

Zadanie 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja

1. Cel

Celem zadania jest zbadanie wpływu ekstrakcji cech oraz wykorzystanych miar podobieństwa w procesie klasyfikacji tekstu. Klasyfikacja tekstów ma zostać zrealizowania z wykorzystaniem algorytmu najbliższych sąsiadów (KNN).

2. Wprowadzenie

Do wykonania tego zadania niezbędne będzie skorzystanie z uczenia maszynowego. Klasyfikacja tekstów odbywać się będzie za pomocą algorytmu KNN. Polega ona na przypisaniu tekstu do odpowiedniej kategorii. Odbywa się to na podstawie wartości poszczególnych wyekstrahowanych cech, które posiada każdy z tekstów.
Do ekstrakcji wykorzystywany jest znormalizowany zbiór słów badanego tekstu. Normalizacja ma na celu wyeliminowanie niepożądanych słów oraz sprowadzenie odmian słów o tym samym znaczeniu do jednego (określenie rdzenia słowa).

2.1. Wykorzystane cechy

2.1.1. Występowanie słowa kluczowego

Na podstawie dostarczonego słowa kluczowego określamy czy to słowo występuje w dokumencie (reprezentacja binarna). Przy użyciu tej cechy nie bierzemy pod uwagę liczności występowania tego słowa kluczowego. Cecha przyjmuje wartości $\{0; 1\}$.

$$F = \begin{cases} 1, & kiedy\ w = t_i \\ 0, & w\ przeciwnym\ przypadku \end{cases} \tag{1}$$

gdzie:

w - dane słowo kluczowe;
t_1 - kolejne słowa występujące w tekście;

2.1.2. Liczba wystąpień słowa kluczowego

Na podstawie dostarczonego słowa kluczowego określamy liczbę wystąpień tego słowa w dokumencie. Wykorzystanie tej cechy dla listy słów kluczowych pozwala na stwierdzenie jak licznie występuje każde słowo kluczowe w dokumencie. Chcemy sprawdzić, czy określenie liczności słów kluczowych ma istotny wpływ na klasyfikację dokumentu.

$$F = \sum_{i=0}^{n} S(w) \tag{2}$$

gdzie:

$$S(w) = \begin{cases} 1, & kiedy\ w = t_i \\ 0, & w\ przeciwnym\ przypadku \end{cases} \tag{3}$$

oraz:
t_i - kolejne słowa występujące w tekście;
n - liczba słów kluczowych;

2.1.3. Suma wystąpień słów kluczowych

Wystąpienie w dokumencie wszystkich słów kluczowych z listy ma istotny wpływ na to w jakim stopniu dany tekst przynależy do danej etykiety. Cecha to liczba całkowita przyjmująca wartości z przedziału $< 0; n >$, gdzie n liczba słów w tekście.

$$F = S_1 + ... + S_n \tag{4}$$

gdzie:

S - liczba wystąpień słowa kluczowego;

2.1.4. Gęstość występowania słów kluczowych

Wyliczenie gęstości słów kluczowych pozwala na sprawdzenie czy cały dokument stanowi na dany temat czy jest to jedynie wzmianka. Cecha to liczba przyjmująca wartości z zakresu $< 0; 1 >$.

$$F = \frac{S}{L} \tag{5}$$

gdzie:

S - suma wystąpień słów kluczowych;
L - liczba wszystkich słów w tekście;

2.1.5. Odległość słowa kluczowego od początku tekstu

Odległość słowa kluczowego od początku tekstu wyrażona za pomocą liczby wyrazów poprzedzających dane słowo kluczowe. Cecha przyjmuje wartości liczbowe z zakresu $< 0; n >$.

$$F = \sum_{0}^{n} 1 \tag{6}$$

gdzie:

n - liczba wyrazów poprzedzających dane słowo kluczowe;

2.1.6. Średnia odległość słów kluczowych od początku tekstu

Średnia odległość słów kluczowych wyrażona jako iloraz sumy odległości słów kluczowych od początku tekstu i liczby słów kluczowych w tekście. Cecha to suma wartości liczbowych odzwierciedlających odległość danego słowa kluczowego od początku tekstu, gdzie pojedyncze słowo odpowiada odległości równej jeden. Cecha przyjmuje wartości z przedziału $< 0; n >$, gdzie n jest maksymalną średnią z sumy odległości słów kluczowych od początku tekstu.

$$F = \frac{S_1 + ... + S_n}{L} \tag{7}$$

gdzie:

S - odległość słowa kluczowego od początku tekstu;
L - liczba słów kluczowych w tekście;

2.1.7. Pierwsze słowo kluczowe

Pierwsze słowo kluczowe to cecha, która jest reprezentowana przez pierwsze napotkane słowo w tekście należące do zbioru słów kluczowych. Cecha przyjmuje wartość tekstową odpowiadającą pierwszemu napotkanemu słowu kluczowemu.

$$F = h(i) = \{w, \quad w = t_i \tag{8}$$

gdzie:

w - dane słowo kluczowe;
t_i - kolejne słowa występujące w tekście;

2.1.8. Najczęściej występujące słowo kluczowe

Najczęściej występujące słowo kluczowe wyrażone jako tekst odpowiadający największej liczbie wystąpień danego słowa kluczowego w tekście. Cecha przyjmuje wartość tekstową odpowiadającą najczęściej występującemu słowu kluczowemu.

$$F = w_{max_i(h(w_i),h(w_i))} \tag{9}$$

gdzie:

$$h(w) = \sum_{i=0}^{n} S(w) \tag{10}$$

$S(w)$ - według wzoru (3);
w - dane słowo kluczowe;
t_i - kolejne słowa występujące w tekście;

2.1.9. Liczba wszystkich słów

Określenie liczby wszystkich słów występujących w tekście. Cecha przyjmuje wartość n równą liczbie słów w tekście.

$$F = |S| \tag{11}$$

gdzie:

S - to zbiór słów występujących w tekście;

2.2. Określanie istotności słów

Podczas analizy tekstu można zauważyć, że niektóre słowa są bardziej lub mniej istotne. W związku z tym zależy nam na eliminacji słów nieznaczących i wyodrębnieniu słów istotnych w kontekście dokumentu. Istotność słów możemy określić na podstawie algorytmów opisanych poniżej.

2.2.1. Częstość słów (ang. Term Frequency)

Liczba wystąpień słowa w dokumencie w stosunku do wszystkich słów pozwala nam na określenie częstotliwości występowania określonego słowa. Zakładamy, że słowo, które występuje stosunkowo rzadko w dokumencie jest wysoce istotne. Zależność tą można określić za pomocą wzoru:

$$F = \frac{K}{W} \tag{12}$$

gdzie:

K - liczba wystąpień danego słowa kluczowego w dokumencie;
W - liczba wszystkich słów w dokumencie;

2.2.2. IDF (ang. inverse document frequency)

Możemy określić w ilu dokumentach występuje dane słowo. Jeśli słowo występuje w małej liczbie dokumentów lub tylko w jednym, można stwierdzić,

że to słowo jest ściśle powiązane z treścią tych dokumentów. Zależność tą można określić za pomocą wzoru:

$$F = log\frac{W}{D} \tag{13}$$

gdzie:

W - liczba wszystkich dokumentów;
D - liczba dokumentów w których wystąpiło słowo kluczowe;

2.2.3. TF-IDF

Jest to połączenie częstości słów występujących w dokumencie zestawione z stosunkiem wytępowania tego słowa we wszystkich dokumentach. Połączenie tych cech pozwala na uzależnienie istotności słowa nie tylko od dokumentu w którym występuje, ale również od występowania w całym zbiorze dokumentów. Zależność tą można określić za pomocą wzoru:

$$F = TF \cdot IDF \tag{14}$$

gdzie:

TF - częstość słów w danym dokumencie;
IDF - częstość występowania na tle innych dokuemntów;

2.2.4. Generowanie stop listy

Tworzenie stop listy opieramy na algorytmie IDF. Słowo które ma najmniejszą wartość IDF występuje najczęściej. Czyli jest nieistotne. Dodajemy do stoplisty jeśli dana wartość będzie poniżej określonego progu.

2.2.5. Nauka słów kluczowych

Do nauki słów kluczowych wykorzystujemy ekstraktory TF oraz TF-IDF, za pomocą których określamy istotność danego słowa. Kiedy dla danego słowa wartość ta jest większą od założonej, uznajemy to słowo za istotne tym samym dopisując je do listy słów kluczowych. Słowa kluczowe mogą zostać również określone odgórnie. Wtedy niezależnie od liczności danej etykiety zostanie wybrana tylko określona liczba słów.

2.3. Metryki - miara odległości

Wykorzystane metryki służa do określenia odległości pomiędzy elementami tego samego zbioru w naszym przypadku tym zbiorem będzie zbiór artykułów. Do obliczenia odległości pomiędzy dwoma artykułami na potrzeby algorytmu KNN zostały wykorzystane metryki opisane poniżej.

2.3.1. Metryka euklidesowa

Odległość euklidesowa jest to odległość między dwoma wektorami określona jako pierwiastek kwadratowy sumy różnic między wartościami podniesionymi do kwadratu, wyrażona wzorem:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}((x_i - y_i)^2)} \tag{15}$$

gdzie:

d - miara odległości;
x, y - wartości cech;

2.3.2. Metryka czebyszewa

Odległość czebyszewa jest to różnica pomiędzy znormalizowanymi cechami wartości obiektów, określona wzorem:

$$d(x,y) = max_i|x_i - y_i| \tag{16}$$

gdzie:

d - miara odległości;
x, y - wartości cech obiektów;

2.3.3. Metryka Manhattan

Metryka Manhattan jest metryką podobną do metryki euklidesowej z tą różnicą, że odległość wyliczana jest z bezwzględnych różnic pomiędzy wektorami. Odległość tę wyraża się wzorem:

$$d(x,y) = \sum_{i=1}^{n}|x_i - y_i| \tag{17}$$

gdzie:

d - miara odległości;
x, y - wartości cech obiektów;

2.4. Miary podobieństwa tekstów

Miara podobieństwa tekstów to miara określająca, w jakim stopniu dany tekst A jest podobny do tekstu B.

2.4.1. Metoda n-gramów

Metoda n-gramów określa w jakim stopniu łańcuch znaków x jest podobny do łańcucha znaków y, na podstawie podciągów.

$$sim_n(x,y) = \frac{1}{N-n+1}\sum_{i=1}^{N-n+1} h(i) \tag{18}$$

gdzie:

$h(i)$ - przyjmuje 1 jeżeli dany podciąg występuje w łancuchu znaków y, w przeciwnym wypadku przyjmuje wartość 0;
N - liczba liter w słowie;
n - długość n-grama;
$N - n + 1$ - ilość n-elementowych podciągów w łańcuchu znaków;

2.4.2. Algorytm KMP (Knutha-Morrisa-Pratta)

Algorytm KMP wyszukuje podany wzorzec x w tekście y, jeżeli podany wzorzec zostaje znaleziony zwracana jest jego pozycja w tekście.

2.5. Normalizacja (Zero-Mean)

Normalizacji poddawany jest każdy wektor cech. Działanie to eliminuje sytuację, w której wartość pewnej cechy może zdominować cały wektor cech.

$$V' = \frac{V - \bar{U}}{\sigma} \tag{19}$$

gdzie:

V' - znormalizowana wartość cechy;
V - wartość cechy;
$\bar{U}$ - średnia z wektora cech;
σ - odchylenie standardowe z wektora cech;

2.6. Określanie poprawności klasyfikacji

Poprawność sklasyfikowanych danych określa w jakim stopniu dokumenty zostały poprawnie przyporządkowane etykietom. Wartość poprawności będzie oscylować w przedziale $< 0; 1 >$. Kiedy wartość będzie będzie zbliżać się do jedynki będzie to oznaczać, że większość dokumentów została sklasyfikowana poprawnie.

$$F = \frac{P}{W} \tag{20}$$

gdzie:

P - liczba poprawnie sklasyfikowanych dokumentów;
W - liczba wszystkich klasyfikowanych obiektów;

3. Opis implementacji

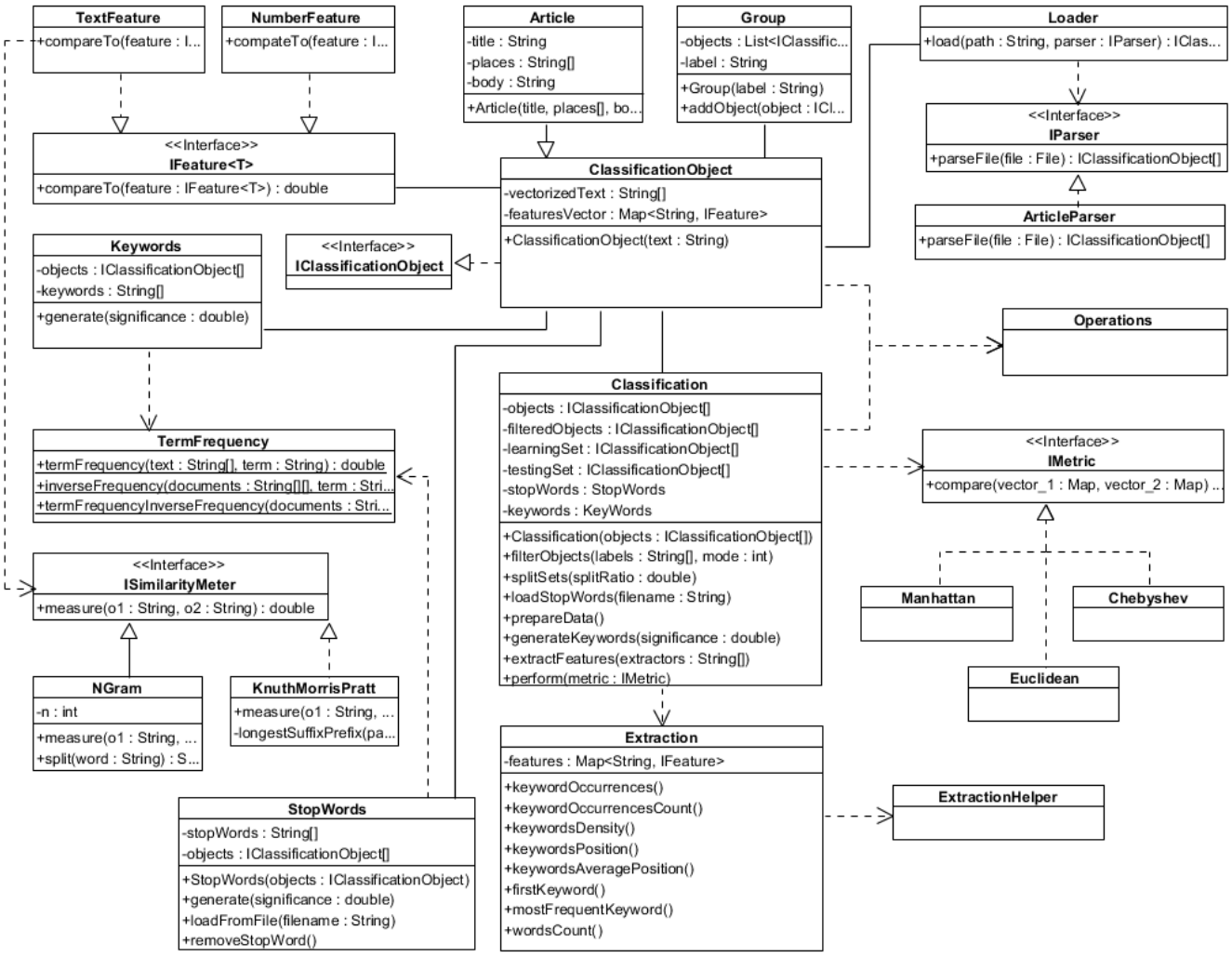
3.1. Struktura projektu

Projekt został zrealizowany w języku Java. W celu uporządkowania projektu utworzona została następująca struktura:
— **data_models** - obsługa przygotowania obiektów do klasyfikacji
— **features** - ekstraktory oraz implementacja cech
— **helpers** - metody pomocnicze
— **metrics** - klasy realizujące miary odległości

— **similarity** - klasy realizujące miary podobieństwa tekstu
— **utils** - narzędzia do manipulacji danymi (stemizacja, słowa kluczowe, stoplista, zapis/odczyt, ...)

3.2. Najważniejsze składowe

Aby zachować niezależność od obiektów, które podlegają klasyfikacji, utworzono interfejs `IClassificationObject`. Interfejs gwarantuje, aby klasyfikowany obiekt posiadał metody udostępniające dane wymagane do klasyfikacji. Dodatkowo stworzono interfejsy `IMetric` oraz `ISimilarity`, których implementacja pozwala na dostarczenie nowych metryk lub miar podobieństwa. Klasa `Extraction` zawiera ekstraktory cech, które tworzą mapę obiektów `IFeature` wewnątrz klasy. Każde użycie ekstraktora powoduje dodanie cech do mapy, która staje się wektorem cech danego obiektu. Klasa `Classification` dokonuje podziału dostarczonych obiektów na dwie grupy zgodnie z ustalonym współczynnikiem podziału oraz odpowiada za realizację wszystkich zadań - wywołuje metody realizujące stemizację, normalizację, ekstrakcję, generowanie słów kluczowych i ostatecznie klasyfikację.



Rysunek 1: Diagram klas

4. Materiały i metody

Klasyfikacja tekstów dotycząca kategorii *PLACES* została przeprowadzna na zbiorze tekstów Reutersa, za pomocą wszystkich zaimplmentowanych ekstraktorów cech (2.1). Badania zostały przeprowadzone dla wszystkich trzech metryk opisanych w punkcie 2.3. Wartości parametru k w algorytmie klasyfikującym K najbliższych sąsiadów zostały dobrane tak, aby wykazać jego wpływ na klasyfikację. Parametr ten przyjmuje wartości ze zbioru {2, 5, 10, 15, 20}. Zbiór tekstów został podzielony odpowiednio - 60% jako dane treningowe oraz 40% jako dane testowe. Jako stop listę wykorzystano gotowy zbiór[1] słów nieznaczących dla języka angielskiego. Słowa kluczowe zostały wygenerowane na podstawie algorytmu częstości słów (TF-IDF). Podczas generowania słów kluczowych została przyjęta wartość 0.75 (TF-IDF), powyżej której słowo traktowane jest jako istotne i tym samym trafia na listę słów kluczowych. Miara odległości pomiędzy wekstrahowanymi wektorami cech zawierającymi słowa została zrealizowana za pomocą metody N-Gramów opisanej w punkcie 2.4.1 z parametrem $N = 3$ (metoda trigramów).

Aby zbadać poprawność klasyfikacji posłużyliśmy się parameterm zgodności, którego wartość wskazuje jaki procent tekstów został sklasyfikowany poprawnie (opisany w punkcie 2.6).

Podczas klasyfikowania kategorii *TOPICS* przyjęto etykiety: ship, tea, oraz silver. Zbiór podzielono równo po 50% dla zbioru uczącego jak i dla zbioru testowego. Stop listę wczytano z pliku jak w poprzednim badaniu, jednak współczynnik generowania słów kluczowych przyjął wartość 0.1. W celu lepszego porównania wyników, badania wykonano dla takich samych parametrów k jak w poprzednim badaniu. Porównanie odbyło się w oparciu o te same metryki liczenia odległości jak i te same miary podobieństwa.

Klasyfikację kategorii *TOPICS* powtórzono zmieniając jedynie metodę generowania słów kluczowych algorytmu TF-IDF na TF z liczbą słów równą 15 dla każdej etykiety.

Własny zbiór tekstów zawierał sześć etykiet: siliconvalley, drwho, twinpeaks, got, friends, simpsons oraz składał się ze 120 obietków. Badań dokonano na wszystkich etykietach. Zbiór podzielono w stosunku 60% : 40%. Jako stop listę wykrzystano gotowy zbiór słów nieznaczączych wymieniony powyżej. Słowa kluczowe zostały wygenerowane za pomocą algorytmu TF-IDF ze współczynnikiem 0.7. Jako miarę podobieństwa słów wybrano metodę Knutha-Morrisa-Pratta. Klasyfikacji dokonano dla każdej z wcześniej przyjętych wartości k, dla każdej z metryk oraz dla wszystkich wcześniej przedstawionych cech.

Badanie dotyczące optymalnego podziału zbioru rozpoczęto od podziału 90% tekstów jako zbiór uczący i 10% jako testowy, kolejno do wartości 10% zbiór uczący oraz 90% zbiór testowy. Skorzystano z etykiet: hongkong, sweeden, philipines. Etykiety dobrano tak, aby ich liczebność była zbliżona. Stop listę wczytano z pliku jak w poprzednich badaniach. Słowa kluczowe generowano ze współczynnikiem TF-IDF wynoszącym $0,2$. Generowania słów kluczowych dokonywano przy każdej zmianie podziału zbioru ze względu na

[1] https://gist.github.com/sebleier/554280

zmianę liczebności zbioru uczącego. Wykorzystano wszystkie dostępne ekstraktory cech. W badaniu parametr k był stały i wynosił 20. Jako metrykę przyjęto metrykę Euklidesową. Liczebność zbioru na którym przeprowadzono badanie wynosiła 206 elementów.

Podczas badania jak miara podobieństwa słów wpływa na poprawność klasyfikacji wykorzystano ustawienia programu takie, jak te wykorzystane podczas badania podziału zbioru. Zmianie uległa jedynie miara podobieństwa słów. Zbiór uczący zawiarał 60% wszystkich tekstów.

Badanie wpływu danych cech na poprawność klasyfikacji oraz czas jej trwania przeprowadzono na etykietach: *canada*, *uk*, *west − germany* z parametrem $k = 20$. Natomiast jako miarę odległości przyjęto metrykę euklidesową. Zbiór podzielono w stosunku 60% : 40%. Słowa kluczowe zostały wygenerowane z parametrem TF-IDF równym 0.4. Stop listę wczytano z pliku.

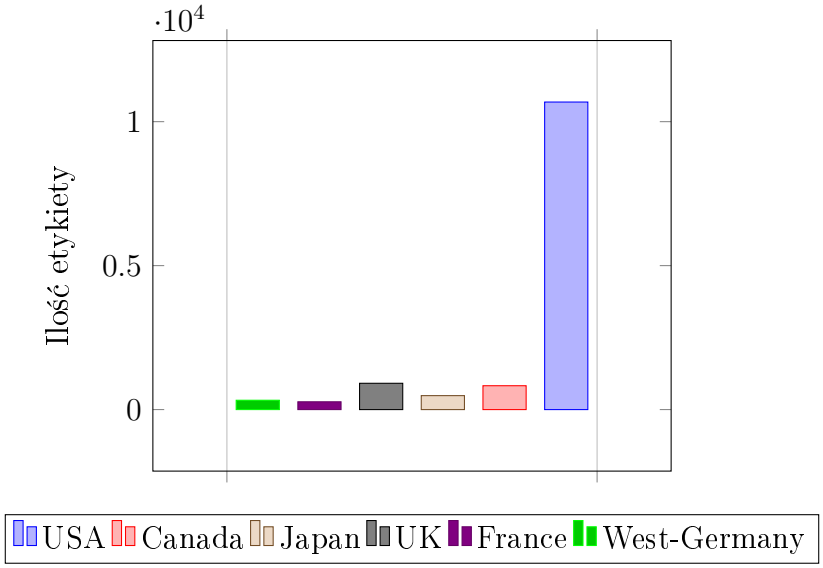
Wpływ miary podobieństwa słów w metodzie N-Gramów zbadano dla wartości N równych kolejno $1, 2, 3, 4, 5, 6, 7$. Do badania wybrano tylko cechy: najczęściej występujące słowo kluczowe oraz pierwsze słowo kluczowe, tak aby inne cechy nie wpływały na wynik badania. Pozostałe ustawienia programu zostały ustawione tak jak podczas badania wpływu danych cech na poprawność klasyfikacji.

5. Wyniki

Wyniki zostały podzielone na sekcje, każda z sekcji przedstawia wyniki klasyfikacji tekstu dla danego badania.

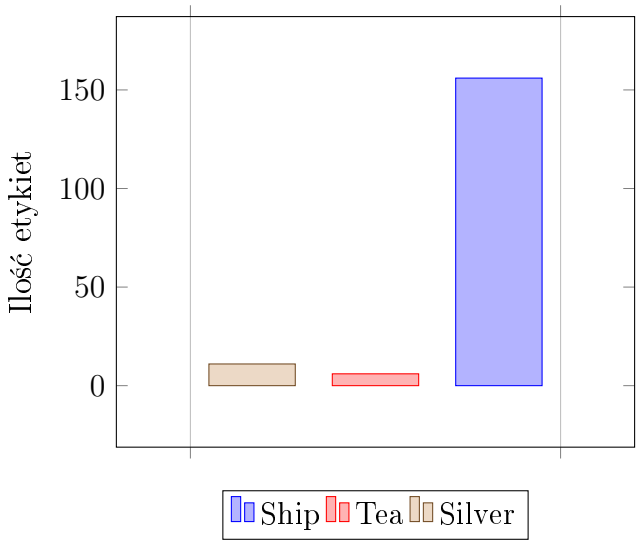
5.1. Liczebności poszczególnych badanych zbiorów

5.1.1. Liczebność zbioru dla kategorii *PLACES*



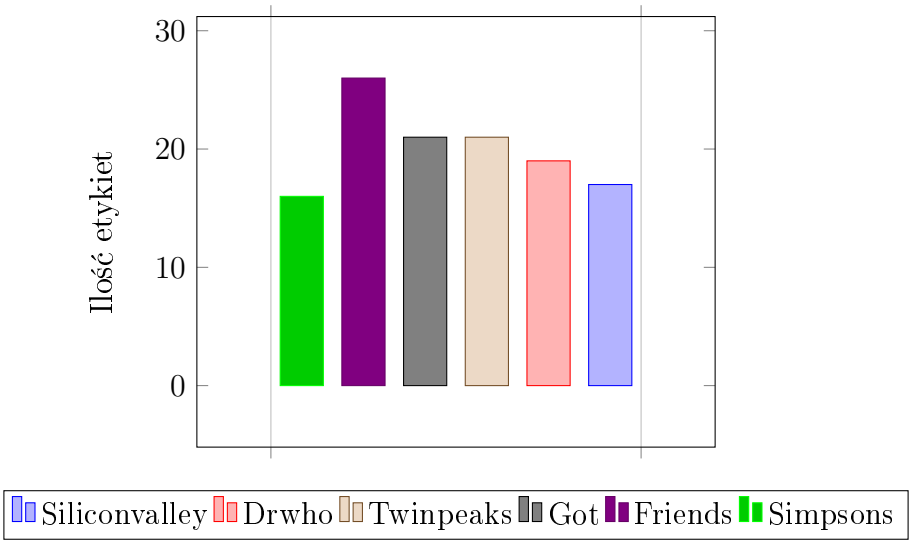
Rysunek 2: Liczebność poszczególnych etykiet dla kategorii *PLACES*.

5.1.2. Liczebność zbioru dla kategorii *TOPICS*



Rysunek 3: Liczebność poszczególnych etykiet dla kategorii *TOPICS*.

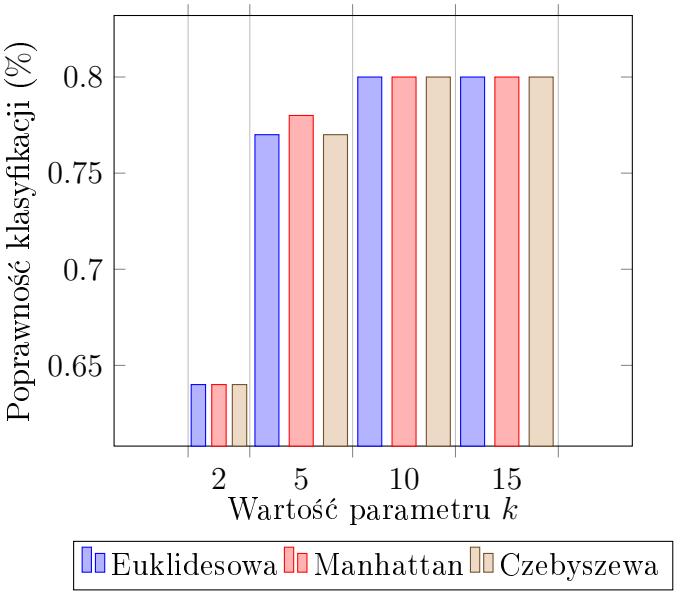
5.1.3. Liczebność własnego zbioru



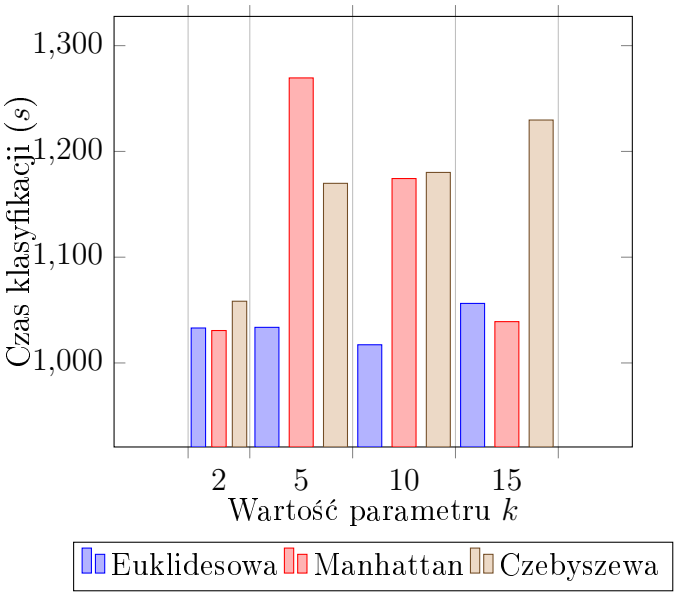
Rysunek 4: Liczebność poszczególnych etykiet dla własnego zbioru tekstów.

5.2. Wyniki klasyfikacji dla kategorii *PLACES*

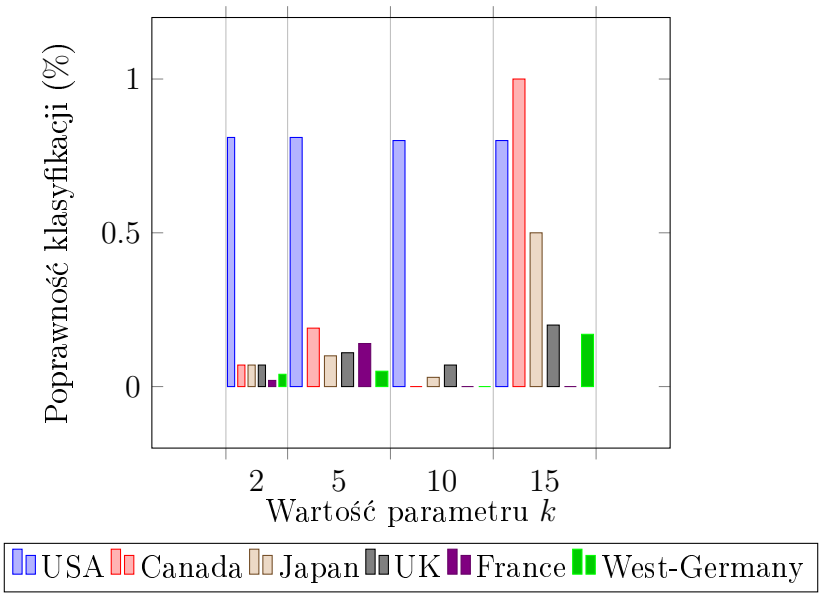
Klasyfikację przeprowadzono z wykorzystaniem każdej z wcześniej wypisanych metod ekctrakcji cech. Wektor klasyfikacji składa się z wartości tekstowych jak i liczbowych. Wyniki klasyfikacji przedstawiono poniżej na wykresach.



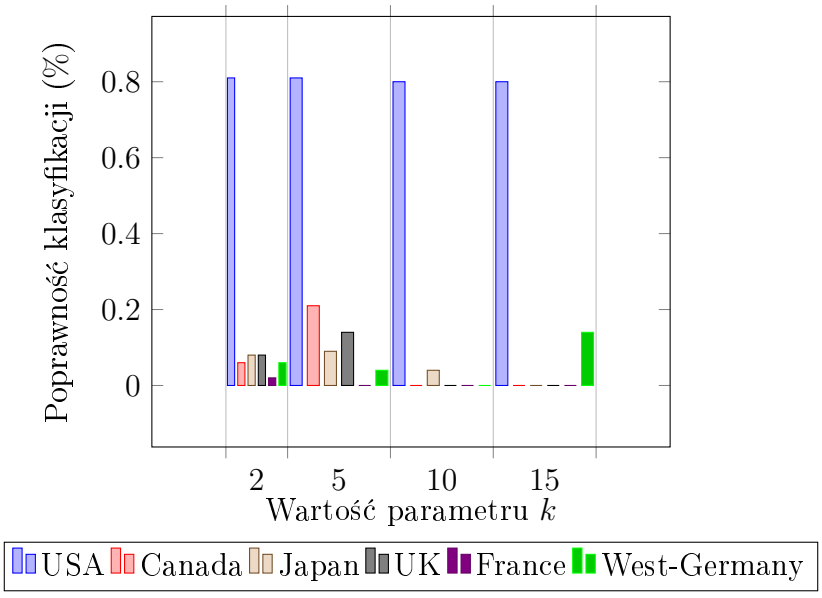
Rysunek 5: Poprawność klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla kategorii $PLACES$.



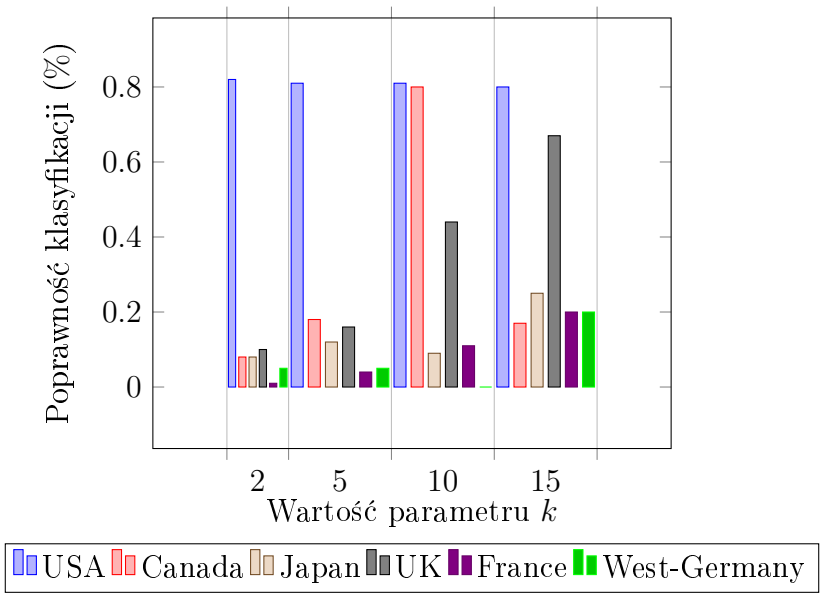
Rysunek 6: Czas klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla kategorii $PLACES$.



Rysunek 7: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla kategorii $PLACES$ z użyciem metryki Euklidesowej.



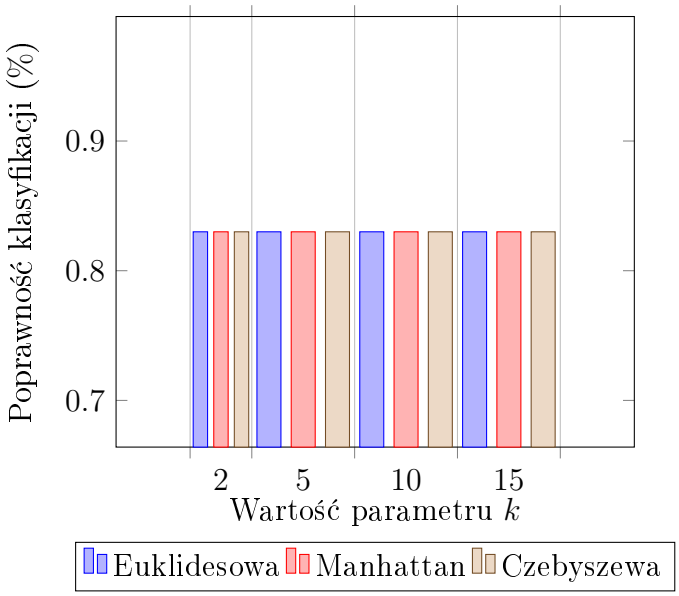
Rysunek 8: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla kategorii $PLACES$ z użyciem metryki Manhattan.



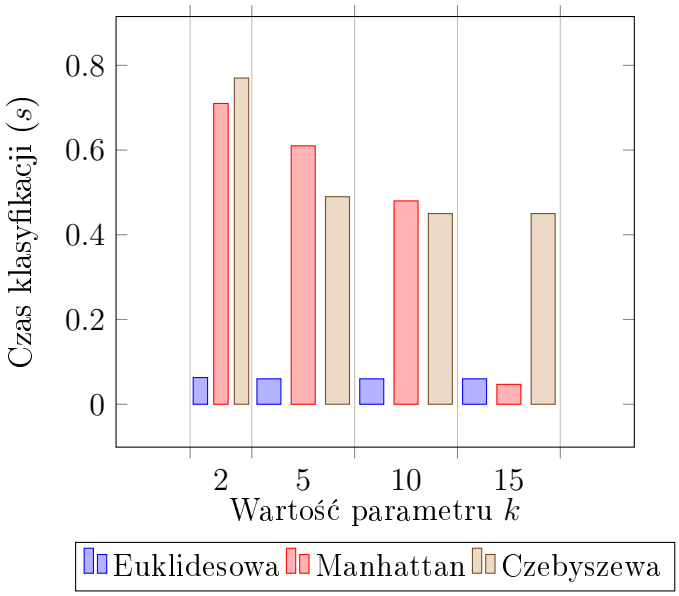
Rysunek 9: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla kategorii *PLACES* z użyciem metryki Czebyszewa.

5.3. Wyniki klasyfikacji dla kategorii *TOPICS*

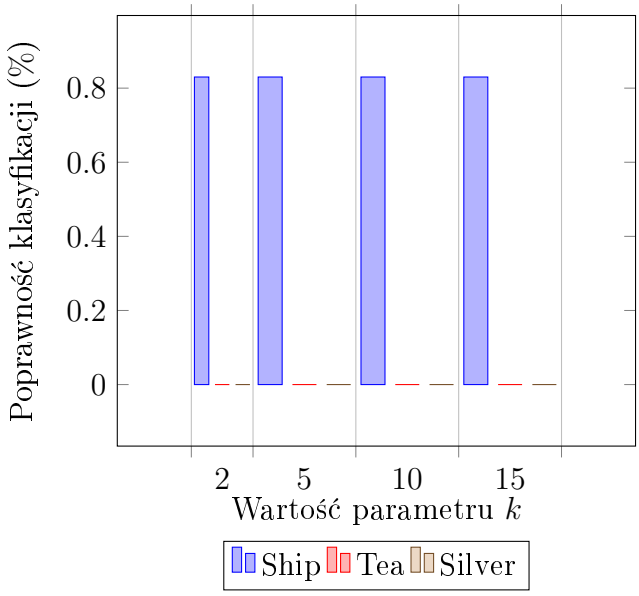
Klasyfikację przeprowadzono dla podanych w punkcie 4 danych. Zbiór danych jest znacznie mniejszy od poprzedniego, zawiera on jedynie 173 elementy.



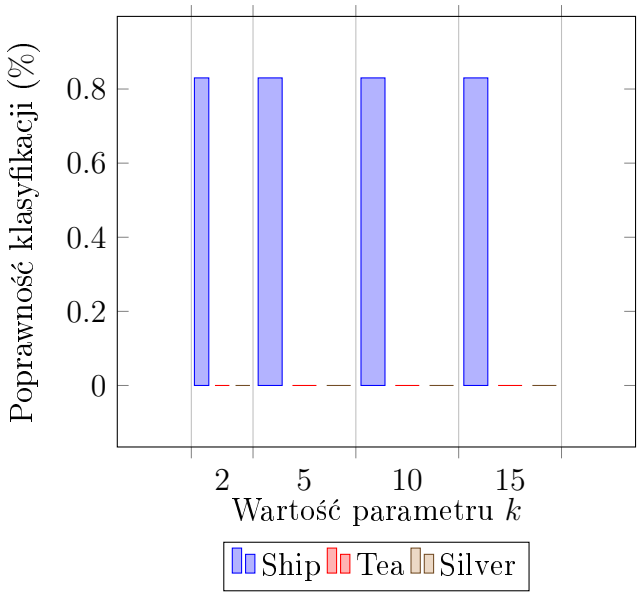
Rysunek 10: Poprawność klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla kategorii *TOPICS*.



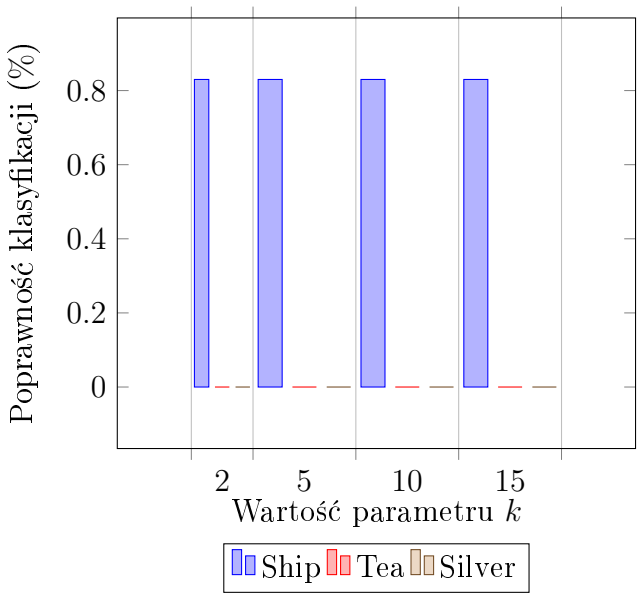
Rysunek 11: Czas klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla kategorii *TOPICS*.



Rysunek 12: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla kategorii $TOPICS$ z użyciem metryki Euklidesowej.

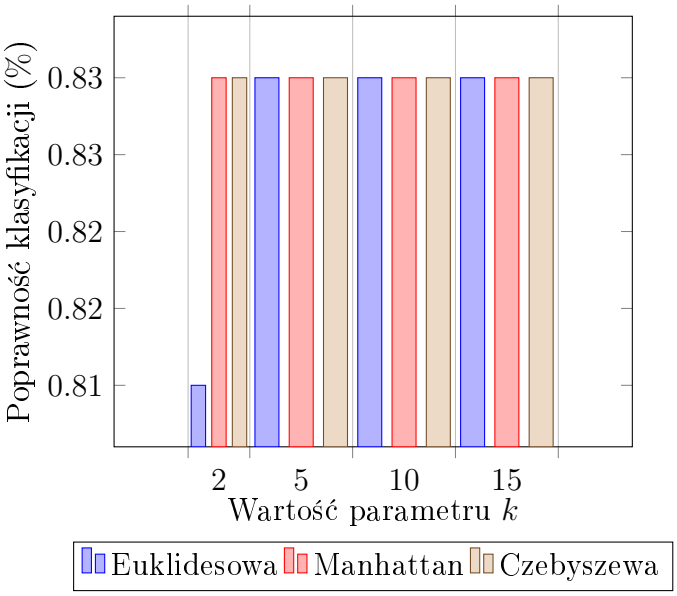


Rysunek 13: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla kategorii $TOPICS$ z użyciem metryki Manhattan.

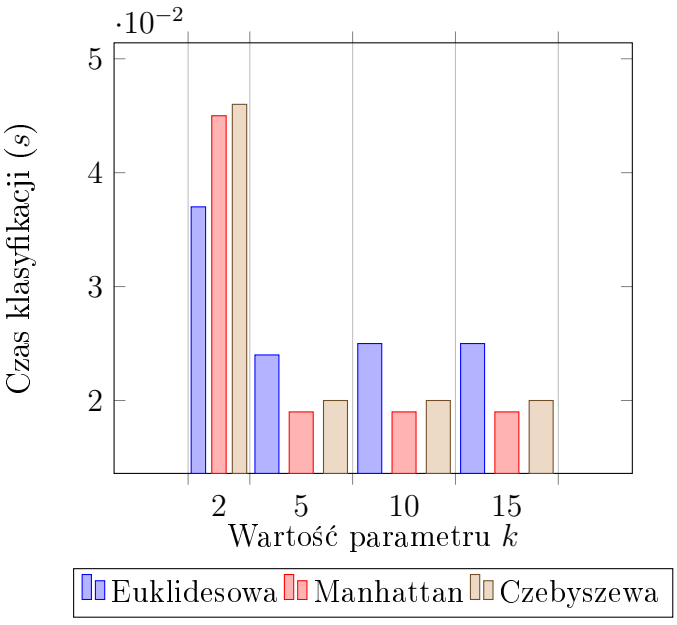


Rysunek 14: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla kategorii $TOPICS$ z użyciem metryki Czebyszewa.

5.4. Wyniki klasyfikacji dla kategorii *TOPICS* z użyciem ekstraktora TF

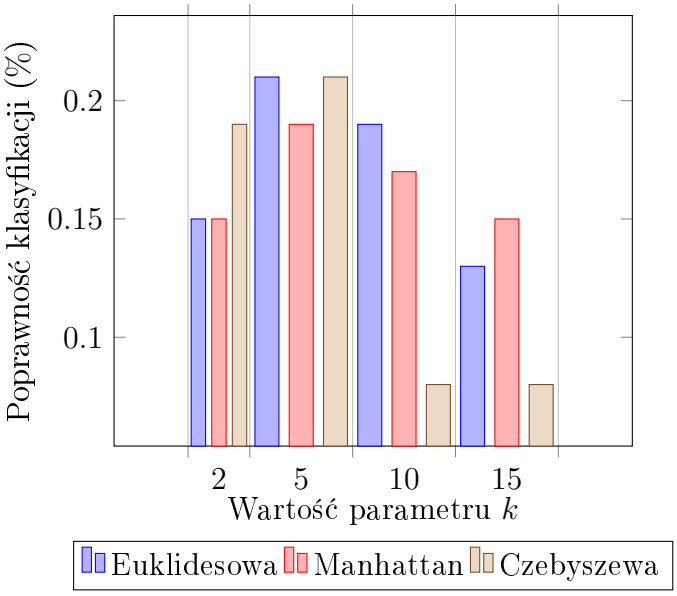


Rysunek 15: Poprawność klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla kategorii *TOPICS* z użyciem ekstraktora TF.

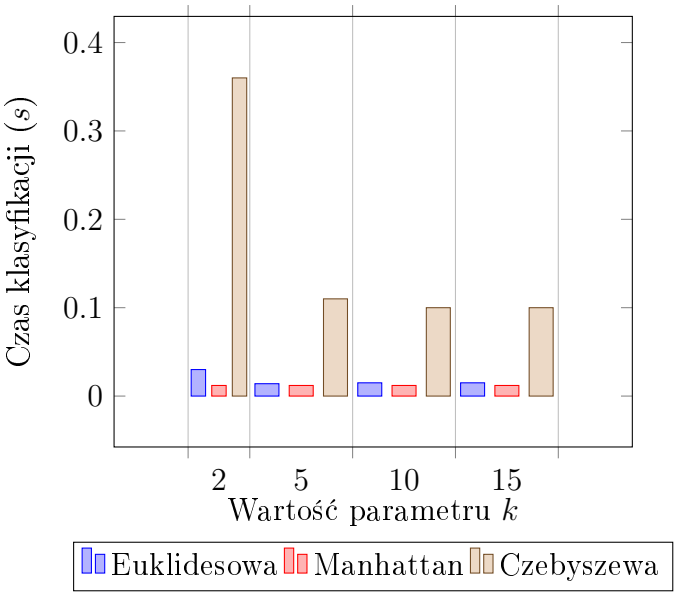


Rysunek 16: Czas wykonywania dla wybranych wartości parametru k i zawartych metryk dla kategorii *TOPICS* z użyciem ekstraktora TF.

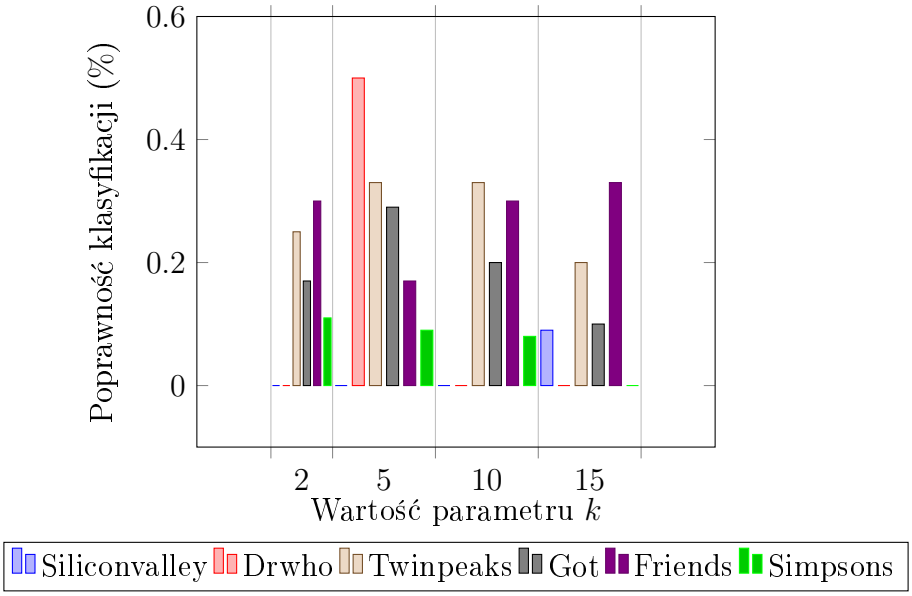
5.5. Wyniki klasyfikacji dla własnego zbioru



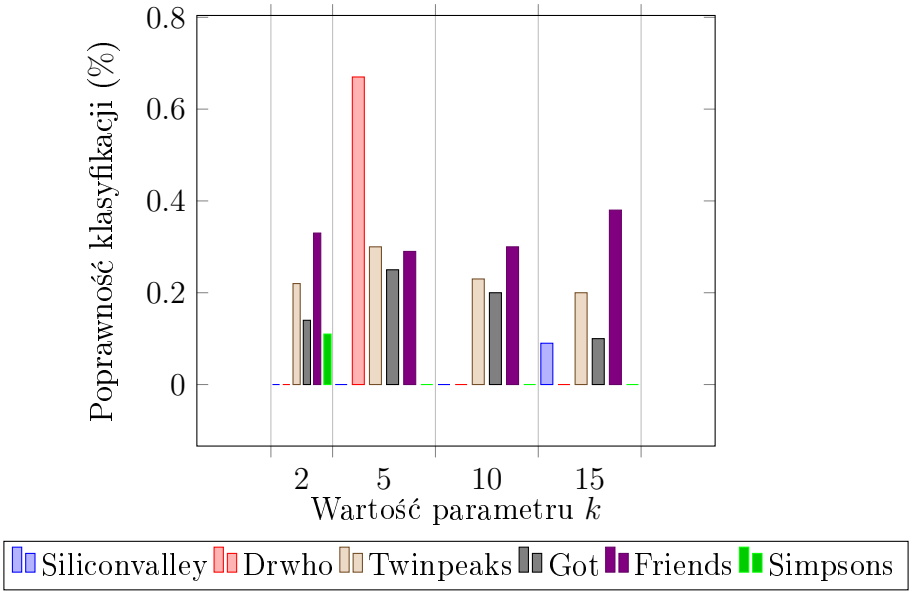
Rysunek 17: Poprawność klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla własnych tekstów.



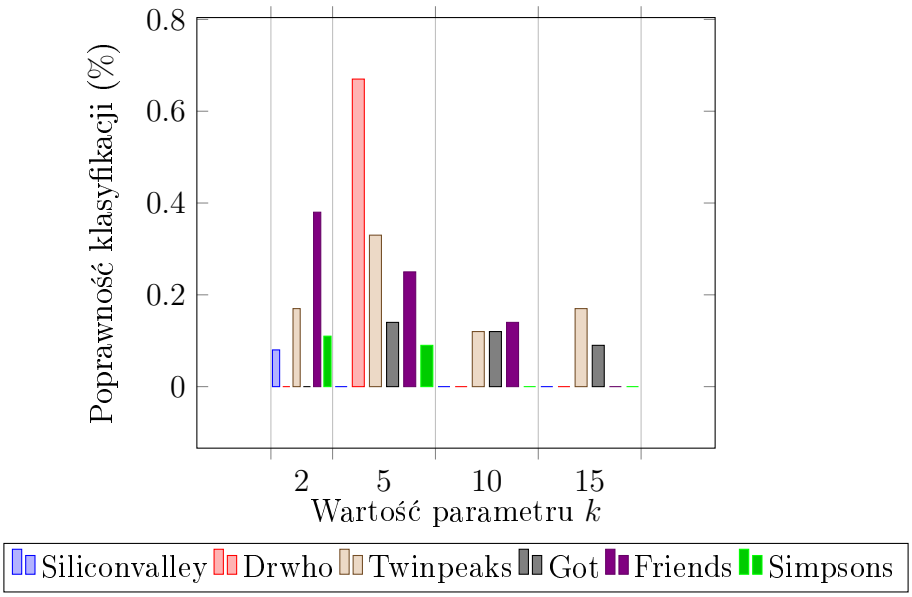
Rysunek 18: Czas wykonywania dla wybranych wartości parametru k i zawartych metryk dla własnych tekstów.



Rysunek 19: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla własnych tekstów z użyciem metryki Euklidesowej.

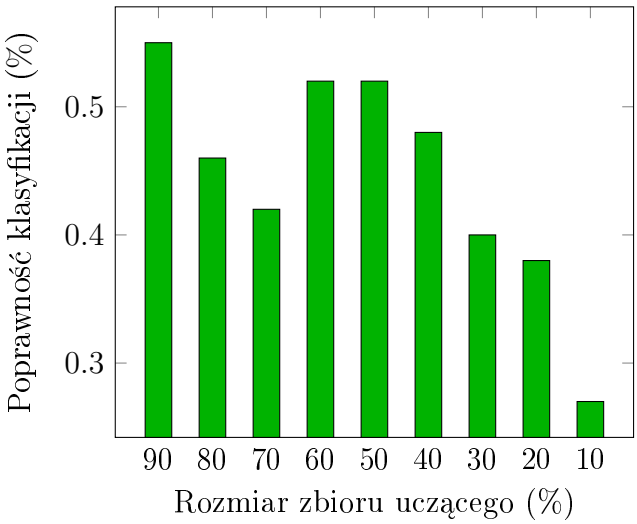


Rysunek 20: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla własnych tekstów z użyciem metryki Manhattan.



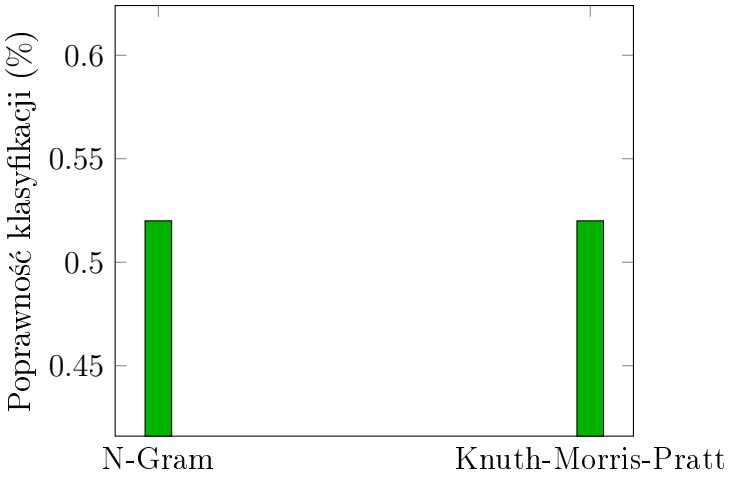
Rysunek 21: Poprawność klasyfikacji etykiety dla wybranych wartości parametru k dla własnych tekstów z użyciem metryki Czebyszewa.

5.6. Wpływ rozmiaru zbioru uczącego na skuteczność klasyfikacji

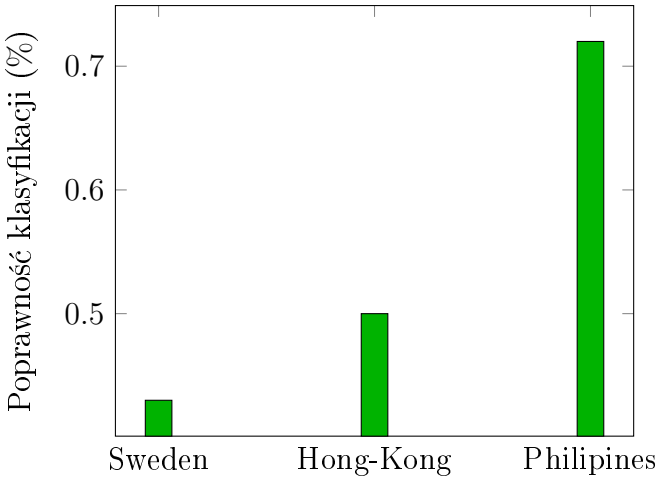


Rysunek 22: Poprawność klasyfikacji dla danego podziału zbioru w kategorii *PLACES*.

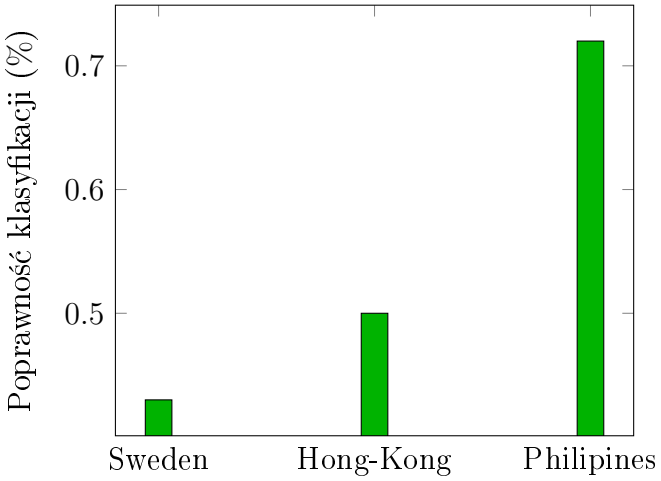
5.7. Wpływ miary podobieństwa słów na poprawnośc klasyfikacji



Rysunek 23: Poprawność klasyfikacji dla miar podobieństwa słów w kategorii *PLACES*.



Rysunek 24: Poprawność klasyfikacji etykiet dla miary N-Gramów w kategorii *PLACES*.

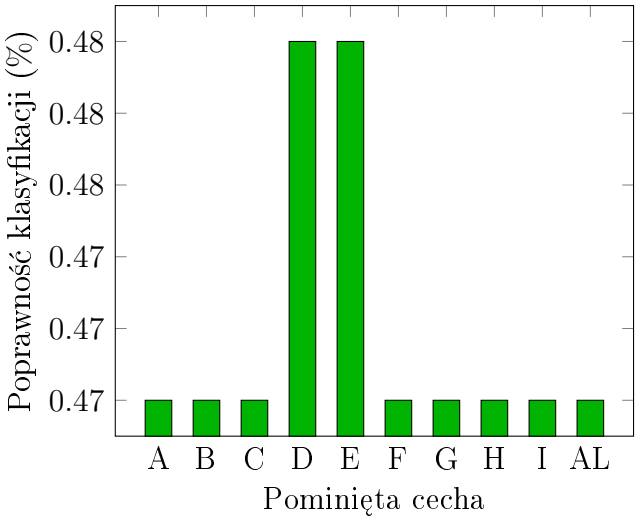


Rysunek 25: Poprawność klasyfikacji etykiet dla miary Knutta-Morrisa-Pratta.

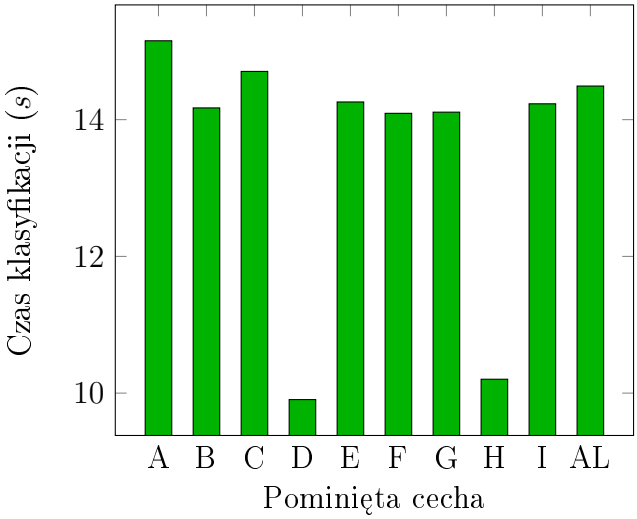
5.8. Wpływ danej cechy na poprawność klasyfikacji oraz jej czas trwania

Dla lepszej czytelności wykresów wprowadziliśmy oznaczenia odpowiadające danym cechom przedstawionym poniżej.

— *A* - Gęstość występowania słów kluczowych
— *B* - Występowanie słowa kluczowego
— *C* - Liczba wystąpień słów kluczowych
— *D* - Odległość słowa kluczowego od początku tekstu
— *E* - Liczba wszystkich słów w dokumencie
— *F* - Pierwsze słowo kluczowe z dokumentu
— *G* - Suma wyspień słów kluczowych w dokumencie
— *H* - Średnia odległość słów kluczowych w dokumencie
— *I* - Najczęściej występujęce słowo kluczowe
— *AL* - Wszystkie cechy

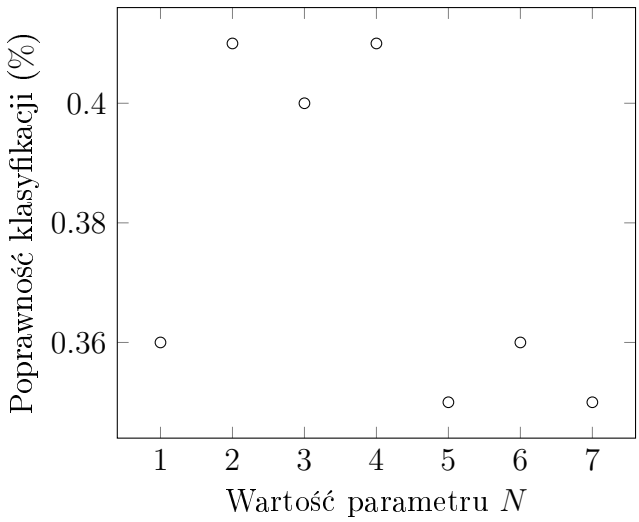


Rysunek 26: Poprawność klasyfikacji z pominięciem danej cechy.

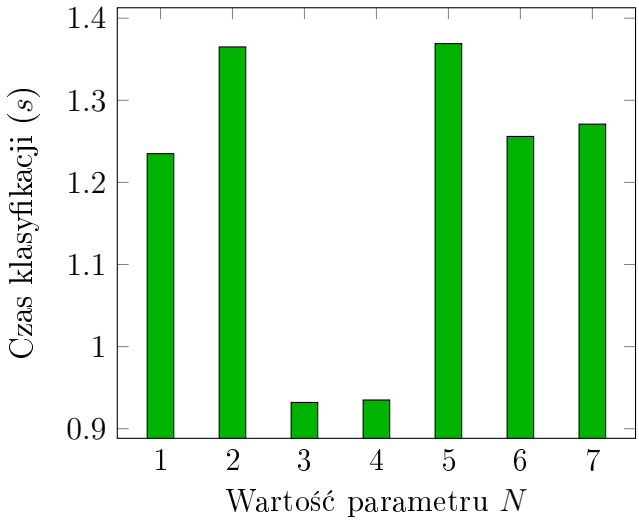


Rysunek 27: Czas klasyfikacji z pominięciem danej cechy.

5.9. Wpływ parametru N w metodzie N-Gramów na poprawność klasyfikacji



Rysunek 28: Poprawność klasyfikacji w zależności od parametru N w metodzie N-Gramów.



Rysunek 29: Czas klasyfikacji w zależności od parametru N w metodzie N-Gramów.

5.10. Podsumowanie wyników

5.10.1. Liczebność poszczególnych zbiorów

Rodzaj etykiety	Nazwa etykiety	Liczebność
Euklidesowa	USA	10682
	Canada	829
	Japan	485
	UK	914
	France	271
	West-Germany	325
Manhattan	Ship	156
	Tea	6
	Silver	11
Czebyszewa	Siliconvalley	17
	Drwho	19
	Twinpeaks	21
	Got	21
	Friends	26
	Simpsons	16

Tabela 1: Podsumowanie liczebności zbiorów.

5.10.2. Zestawienie badań poprawności klasyfikacji

Metryka	k	Poprawność klasyfikacji [%]							Czas [s]
		usa	canada	japan	uk	france	west-ger	ogółem	
Euklidesowa	2	81	7	7	7	2	4	64	1033,047
	5	81	19	10	11	14	5	77	1033,665
	10	80	0	3	7	0	0	80	1017,195
	15	80	100	50	20	0	17	80	1056,296
	20	80	0	0	0	0	0	80	1072,138
Manhattan	2	81	6	8	8	2	6	64	1030,663
	5	81	21	9	14	0	4	78	1269,484
	10	80	0	4	0	0	0	80	1174,352
	15	80	0	0	0	0	14	80	1039,059
	20	80	0	0	0	0	11	80	1089,03
Czebyszewa	2	82	8	8	10	1	5	64	1058,355
	5	81	18	12	16	4	5	77	1169,824
	10	81	8	9	44	11	0	80	1180,133
	15	80	17	25	67	20	20	80	1229,68
	20	80	50	25	0	0	20	80	978,71

Tabela 2: Podsumowanie wyników dla etykiety *PLACES* przy podziale 60% : 40%.

Metryka	**k**	**Poprawność klasyfikacji [%]**				**Czas [s]**
		ship	tea	silver	ogółem	
Euklidesowa	2	83	0	0	83	0,063
	5	83	0	0	83	0,06
	10	83	0	0	83	0,06
	15	83	0	0	83	0,06
	20	83	0	0	83	0,061
Manhattan	2	83	0	50	83	0,71
	5	83	0	0	83	0,61
	10	83	0	0	83	0,48
	15	83	0	0	83	0,047
	20	83	0	0	83	0,044
Czebyszewa	2	83	0	0	83	0,077
	5	83	0	0	83	0,049
	10	83	0	0	83	0,045
	15	83	0	0	83	0,045
	20	83	0	0	83	0,046

Tabela 3: Podsumowanie wyników dla etykiety *TOPICS* przy podziale 50% : 50%.

Metryka	**k**	**Poprawność klasyfikacji [%]**							**Czas [s]**
		sill.	drwho	twin.	got	friends	simp.	ogółem	
Euklidesowa	2	0	0	25	17	30	11	15	0,03
	5	0	50	33	29	17	9	21	0,014
	10	0	0	33	20	30	8	19	0,015
	15	9	0	20	10	33	0	13	0,015
	20	11	0	18	9	33	0	13	0,019
Manhattan	2	0	0	22	14	33	11	15	0,012
	5	0	67	30	25	29	0	19	0,012
	10	0	0	23	20	30	0	17	0,012
	15	9	0	20	10	38	0	15	0,012
	20	11	0	17	8	25	0	10	0,012
Czebyszewa	2	8	0	17	0	38	11	19	0,36
	5	0	67	33	14	25	9	21	0,11
	10	0	0	12	13	14	0	8	0,1
	15	0	0	17	9	0	0	8	0,1
	20	0	0	17	18	0	0	8	0,1

Tabela 4: Podsumowanie wyników dla własnych obiektów przy podziale 60% : 40%

Pominięta cecha	**Poprawność klasyfikacji [%]**				**Czas [s]**
	canada	uk	west-g.	ogółem	
Gęstość słów kluczowych	63	44	32	47	14,492
Występowanie słowa kluczowego	63	44	32	47	15,154
Liczba wystąpień słowa kluczowego	63	44	32	47	14,172
Odległość słowa kluczowego	64	44	32	48	14,706
Liczba wszystkich słów	66	44	39	48	9,905
Pierwsze słowo kluczowe	63	44	32	47	14,259
Suma wystąpień słów kluczowych	63	44	32	47	14,093
Średnia odległość słów kluczowych	64	44	31	47	14,11
Najczęściej występujące słowo klucz.	63	44	32	47	10,203

Tabela 5: Podsumowanie wpływu cechy na poprawność i czas klasyfikacji w kategorii *PLACES*.

5.10.3. Zestawienie miar podobieństwa

Miara podobieństwa	Nazwa etykiety	Poprawność klasyfikacji [%]	Poprawność ogółem [%]	Czas [s]
N-Gramów	Sweden	43	52	0.056
	Hong-Kong	50		
	Philipines	72		
Knutta-Morrisa-Pratta	Sweden	43	52	0.038
	Hong-Kong	50		
	Philipines	72		

Tabela 6: Podsumowanie wpływu miary podobieństwa na poprawność klasyfikacji oraz jej czas trwania.

6. Dyskusja

Na podstawie przeprowadzonych badań klasyfikacji tekstów można stwierdzić, że klasyfikacja teskstów jest złożonym zagadnieniem. W związku z tym dzielimy obszary badań na kilka kategorii.

6.1. Wpływ parametru k

Przeprowadzone badania pozwalają stwierdzić, że wartośc parametru k ma znaczy wpływ na poprawność klasyfikacji, lecz do pewnego momentu. Na podstawie rysunku 5 można stwierdzić, że zbyt niska wartość parametru k negatywnie wpływa na poprawnośc klasyfikacji. Podczas klasyfikacji dla kategorii $PLACES$ i $k = 2$ skuteczność dla każdej metryki wyniosła 64%. Natomiast, jak wynika z rysunku 10, dla kategorii $TOPICS$ poprawność klasyfikacji wyniosła 83% i nie była użależniona od wartości k. Wyniki przedstawione na rysunku 17 wskazują na niski próg poprawności dla własnych tekstów. Dla parametru $k = 5$ wyniki były najlepsze.

6.2. Wpływ metryki

Pomiędzy metrykami wykorzystanymi w klasyfikacji kategorii $PLACES$ oraz $TOPICS$ nie wykazano różnic ze względu na poprawność klasyfikacji, co przedstawia rysunek 5 oraz 10. Badając własny zbiór tekstów (rysunek 17) zaobserwowano, że wpływ metryki jest znacznie większy niż w poprzednich przypadkach. Metryka czebyszewa klasyfikuje obiekty w sposób mniej dokładny niż pozostałe. Znaczące róznice pojawiają się z czasie klasyfikacji. Metryka Euklidesowa zarówno dla kategorii $PLACES$ jak i $TOPICS$ wykazuje złożoność obliczeniową liniową co można zaobserwować na rysunkach: 6, 11. Należy pamiętać że wyniki te są najmniej miarodajne ponieważ mogą się znacząco różnić w zależności od zasobów sprzętowych, które są dostępne w danym momencie.

6.3. Liczebność zbiorów

Badania wykazały, że na poprawność klasyfikacji znaczący wpływ ma liczebność zbiorów. W przypadku kiedy dana etykieta znacząco przeważa liczebnością pozostałe, wyniki klasyfikacji dla pozostałych etykiet będą niskie. Spowodowane jest to znikomą liczbą pozostałych etykiet w zbiorze uczącym. Liczebność zbiorów pokazują rysunki 2, 3 oraz 4.

6.4. Wpływ podziału zbiorów

Z rysunku 22 można odczytać, że dla podziału zbioru w stosunku: 90% : 10% klasyfikacja uzyskała najlepsze wyniki. Klasyfikacja dla podziału w stosunku 30% (zbiór uczący) do 70% (zbiór testowy) oraz niższych wartości dla zbioru uczącego skuteczność klasyfikacji uległa pogorszeniu. Podobnych obserwacji można dokonać w przypadku, gdy zbiór uczący stanowi 70% klasyfikowanych obiektów.

6.5. Wpływ miary podobieństwa

Rysunek 23 przedstawia takie same wartości poprawności klasyfikacji dla obydwu miar podobieństwa słów. Nie wykazano różnic na poziomie poprawności klasyfikacji dla poszczególnych etykiet (rysunki 24, 25). Zróżnicowanie czasu klasyfikacji w przypadku obydwu miar podobieństwa było na tyle małe, że nie byliśmy w stanie określić która miara jest optymalna. Wynika to z odmiennego czasu wykonywania klasyfikacji przy każdym uruchomieniu.

6.6. Wpływ danych cech

Badanie ma na celu sprawdzenie uniwersalnych metod ekstrakcji cech i ich wpływ na skuteczność klasyfikacji (rysunki 26, 27). Przeprowadzone badania wskazują, że pominięcie cechy w postaci liczby wysąpień słowa kluczowego lub odległości słowa kluczowego od początku tekstu nieznacznie wpływa na zwiększenie poprawności klasyfikacji. Ponadto wyłączając cechę odległości słowa kluczowego od początku tekstu oprócz lepszej klasyfikacji zyskujemy zmniejszony jej czas trwania. Biorąc pod uwagę cechę średniej odległości słów kluczowych od początku tekstu zyskujemy zmniejszony czas klasyfikacji jednocześnie nie zwiększając jej poprawności. W przypadku pozostałych cech nie można jednoznacznie swierdzić przewagi jednej cechy nad drugą, ponieważ podobnie jak w trakcie badania wpływu metryki występuje tu różnica w czasie wykonania zależna od zasobów sprzętowych dostępnych w danym momencie.

6.7. Wpływ parametru N

Z rysunku 28 wynika, że dla parametrów N z zakresu $< 2; 4 >$ wyniki klasyfikacji były najlepsze. Wartość parametru powyżej lub poniżej podanego zakresu negatywnie wpływała na poprawność klasyfikacji. Nie zaobserwowano znaczących różnic w czasie działania klasyfikacji. Różnice w czasie klasyfikacji na poziomie 400ms nie są wystarczające aby swierdzić przewagę danej konfiguracji.

6.8. Wpływ ekstraktora TF

Podczas badania dotyczącego zmiany ekstraktora słów kluczowych (rysunek 15) z algorytmu TF-IDF na algorytm TF, nie wykazano znaczących różnic w poprawności klasyfikacji w stosunku do uprzednio wykorzystywanego ekstraktora. Natomiast jak wynika z rysunku 16 czas klasyfikacji uległ nieznacznej poprawie.

7. Wnioski

— Dla parametru k z zakresu $< 2; 5 >$ skuteczność klasyfikacji jest najwyższa.
— Metryka euklidesowa ze względu na jej złożoność liniową jest najszybsza, nie zmniejszając tym samym poprawności klasyfikacji.
— Liczebność etykiet w zbiorze uczącym powinna być zbliżona.
— Podział zbioru (90 do 10) jest najbardziej optymalny.
— Badania nie wykazały wpływu miary podobieństwa słów na poprawność klasyfikacji.
— W badaniach dotyczących poszczególnych cech wyłączenie cechy: "Odległości słowa kluczowego od początku tesktu" pozytywnie wpłyneła na poprawność klasyfikacji oraz na czas jej trwania.
— Dla parametru N z zakresu $< 2; 4 >$ wyniki klasyfikacji były najlepsze.
— Nie wykazano znaczących różnic pomiędzy generowaniem słów kluczowych za pomocą algorytmów TF-IDF oraz TF.

Literatura

[1] Adam Niewiadomski; *Materiały, przykłady i ćwiczenia do przedmiotu Komputerowe Systemy Rozpoznawania*; 21 września 2009;
[2] Isabelle Guyon, Steve Gunn, Masoud Nikravesh, Lofti A. Zadeh; *Feature Extraction: Foundations and Applications*; Springer; 16 listopada 2008;
[3] David D. Lewis; *Feature Selection and Feature Extract ion for Text Categorization*; University of Chicago; 26 września 1992;
[4] B.S.Charulatha, Paul Rodrigues, T.Chitralekha, Arun Rajaraman; *A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms*; 2013
[5] Stop lista; `https://gist.github.com/sebleier/554280`