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, która odbywa się za pomocą algorytmu KNN. Polega ona na przypisaniu tekstu do odpowiedniej kategorii. Odbywa się to na podstawie watrości poszczególnych wyekstrachowanych cech, które posiada każdy z tekstów.
Do ekstrakcji wykorzystywany jest znormalizowany zbiór słów badanego tekstu, oznaczony przez T. 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 klyczowych 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 sumy liczby wyrazów poprzedzających dane słowo kluczowe. Cecha przyjmuje wartości liczbowe z zakresu $< 0; n >$, gdzie n - liczba wyrazów poprzedzających słowo kluczowe.

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

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 $< 0; n >$ 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 przyjmuje wartość pierwszego napotkanego słowa kluczowego w tekście. 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 następujących 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 algorytm TF-IDF, za pomocą którego 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.

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 Manhattana

Metryka Manhattana 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 tesktów to miara mówiąca o tym 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. Uogólniona miara n-gramów

Uogólniona miara n-gramów sprawdza podobieństwo słów w oparciu o podciągi o określonej długości. Wyrażona jest wzorem:

$$u_N(x,y) = \frac{2}{N^2+N} \sum_{i=1}^{N(x)} \sum_{j=1}^{N(x-i+1)} h(i,j) \tag{19}$$

gdzie:

$h(i,j)$ przyjmuje wartość 1 jeżeli dany podciąg ze słowa x znajduje się w słowie y;
$N(x), N(y)$ - ilość liter w słowach x, y, $N = maxN(x), N(y)$;
$\frac{N^2+N}{2}$ - ilość możliwych podciągów 1-elementowych do N-elementowych w słowie o długości N;

2.4.3. 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. 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 klasyfikacji

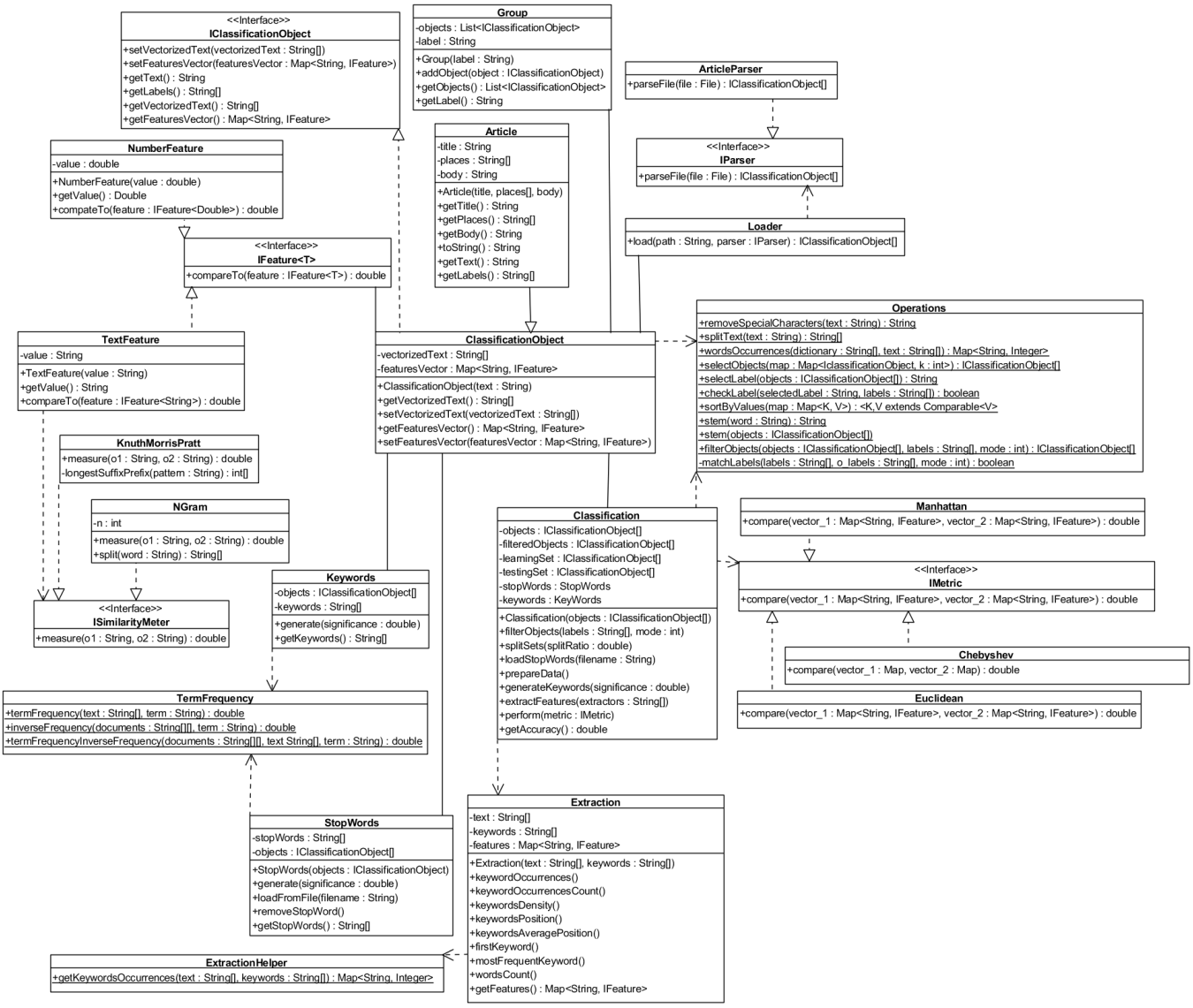
3. Opis implementacji

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. Stop lista została wczytana z pliku *stopwords.txt* natomiast słowa kluczowe zostały wygenerowane ze współczynnikiem $0,6$. Miara odległości pomiędzy wekstrachowanymi wektorami cech zawierającymi słowa została zrealizowana za pomocą metody N-Gramów opisanej w punkcie 2.4.1.

Podczas klasyfikowania kategorii *TOPICS* przyjęto etykiety: ship, tea, cocoa, silver oraz housing. 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 $0,03$. Badanie wykonano dla takich samych parametrów k jak poprzednio dla lepszego porównania wyników. Porównanie odbyło się w oparciu o te same metryki liczenia odległości pomiędzy wektorami dokumentów jak i tę samą metodę obliczania odległości między cechami o wartości tekstowej.

Aby zbadać poprawność klasyfikacji posłużyliśmy się parameterm zgodności, którego wartość wskazuje jaki procent tekstów został sklasyfikowany poprawnie 2.5.

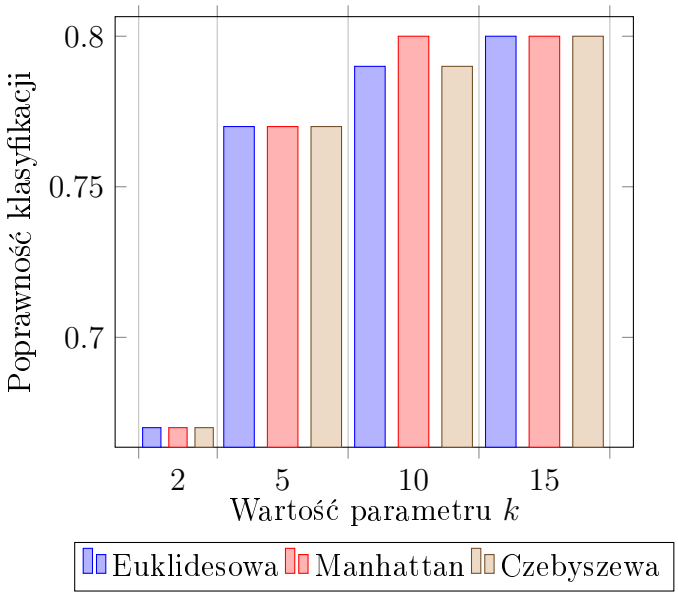
W tym miejscu należy opisać, jak przeprowadzone zostały wszystkie badania, których wyniki i dyskusja zamieszczane są w dalszych sekcjach. Opis ten powinien być na tyle dokładny, aby osoba czytająca go potrafiła wszystkie przeprowadzone badania samodzielnie powtórzyć w celu zweryfikowania ich poprawności (a zatem m.in. należy zamieścić tu opis architektury sieci, wartości współczynników użytych w kolejnych eksperymentach, sposób inicjalizacji wag, metodę uczenia itp. oraz informacje o danych, na których prowadzone były badania). Przy opisie należy odwoływać się i stosować do opisanych w sekcji drugiej wzorów i oznaczeń, a także w jasny sposób opisać cel konkretnego testu. Najlepiej byłoby wyraźnie wyszczególnić (ponumerować) poszczególne eksperymenty tak, aby łatwo było się do nich odwoływać dalej.

5. Wyniki

Wyniki zostały podzielone na sekcje, każda z sekcji przedstawia wyniki klasyfikacji tekstu dla danej kategorii.

5.1. Wyniki klasyfikacji dla kategorii *PLACES*

Klasyfikację przeprowadzono z wykorzystaniem każdej z wcześniej wypisanych metod ekctrakcji cech. Wektor klasyfikacji składa się z wartości

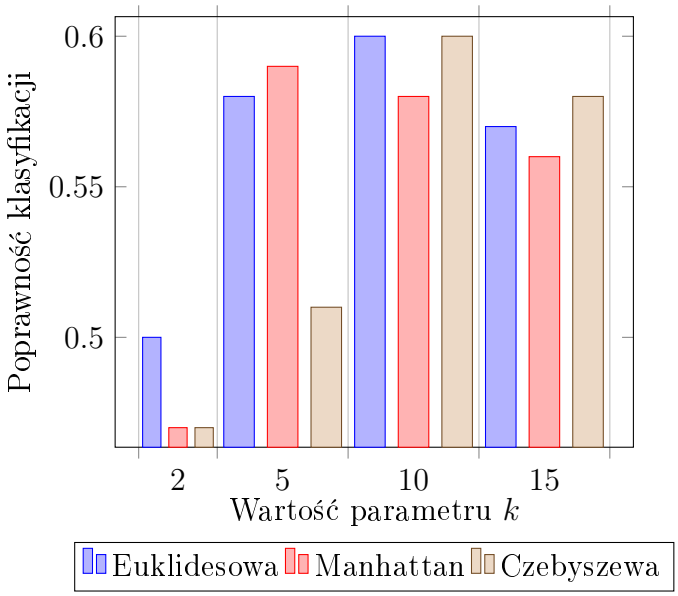


Rysunek 2. Poprawność klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla kategorii $PLACES$.

tekstowych jak i liczbowych. Wyniki klasyfikacji przedstawiono poniżej na wykresie 2.

5.2. Wyniki klasyfikacji dla kategorii $TOPICS$

W tej sekcji należy zaprezentować, dla każdego przeprowadzonego eksperymentu, kompletny zestaw wyników w postaci tabel, wykresów itp. Powinny być one tak ponazywane, aby było wiadomo, do czego się odnoszą. Wszystkie tabele i wykresy należy oczywiście opisać (opisać co jest na osiach, w kolumnach itd.) stosując się do przyjętych wcześniej oznaczeń. Nie należy tu komentować i interpretować wyników, gdyż miejsce na to jest w kolejnej sekcji. Tu również dobrze jest wprowadzić oznaczenia (tabel, wykresów) aby móc się do nich odwoływać poniżej.



Rysunek 3. Poprawność klasyfikacji dla wybranych wartości parametru k i zawartych metryk dla kategorii *TOPICS*

6. Dyskusja

Klasyfikacja tekstów jest znacznie bardziej złożona niż mogło się to w pierwszej chwili wydawać. Aby dokonać poprawnej klasyfikacji należy uprzednio przeanalizować badane teksty między innymi w jakim języku zostały one napisane. Ze względu na język dokumentu należy odpowiednio dokonać stemizacji i lematyzacji. W zależności od typu tekstu należy dobrać odpowiednie ekstraktory cech. Zadajemy sobie sprawę, iż badane przez nas dokumenty mają charakter artykułów z gazet. Dla danego typu badanego tekstu cecha odległości słów kluczowych od początku tekstu będzie kluczowa w poprawnej klasyfikacji. Natomiast na przykład dla opowiadania istotność tej cechy na tle innych znacząco zmaleje. Użyty w badaniu wektor cech składający się zarówno z wartości metajęzykowych (czyli takich dla, których znaczenie mają jedynie położenie danego słowa w tekście czy liczba liter składowych) oraz wartości semantycznych (takich, które zwracają szczególną uwagę na budowę słowa), pozwala na dokładniejsze zbadanie tekstu bez względu na to jak został on zbudowany. Stwierdzenie to można smiało postawić zwracając uwagę na wykres 2 na, którym tylko dla metryki *Manhattan* przy wartości parametu $k = 10$ poprawność klasyfikacji była nieznacznie wyższa. W pozostałych przypadkach poprawność klasyfikacji nie zależała od zastosowanej metryki liczenia odległości.

Sekcja ta powinna zawierać dokładną interpretację uzyskanych wyników eksperymentów wraz ze szczegółowymi wnioskami z nich płynącymi. Najcenniejsze są, rzecz jasna, wnioski o charakterze uniwersalnym, które mogą być istotne przy innych, podobnych zadaniach. Należy również omówić i wyjaśnić wszystkie napotakane problemy (jeśli takie były). Każdy wniosek powinien mieć poparcie we wcześniej przeprowadzonych eksperymentach (odwołania

do konkretnych wyników). Jest to jedna z najważniejszych sekcji tego sprawozdania, gdyż prezentuje poziom zrozumienia badanego problemu.

7. Wnioski

W tej, przedostatniej, sekcji należy zamieścić podsumowanie najważniejszych wniosków z sekcji poprzedniej. Najlepiej jest je po prostu wypunktować. Znów, tak jak poprzednio, najistotniejsze są wnioski o charakterze uniwersalnym.

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] Knuth–Morris–Pratt algorithm; `https://en.wikipedia.org/wiki/Knuth-Morris-Pratt_algorithm`

[5] 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