

Data oddania: \_\_\_\_\_

Ocena: \_\_\_\_\_

Rafał Komorowski 236563  
Kamil Presler 236629

## Projekt 1. Klasyfikacja dokumentów tekstowych

### 1. Cel projektu

Celem zadania pierwszego było stworzenie aplikacji do klasyfikacji zbioru dokumentów tekstowych [13]. Zbiór pochodzi z 1987 roku i zawiera dokumenty z agencji prasowej Reuters. Określany jest jako *Reuters-21578* z uwagi na liczbę dokumentów w zbiorze. Dokumenty są oznaczone różnymi etykietami. W zadaniu metodą *k*-NN [1] będziemy klasyfikować dokumenty do odpowiednich państw, w związku z tym pod uwagę bierzemy etykietę "places". Spośród wszystkich dokumentów wybrano jedynie te, dla których wartość ta jest pojedynczą wartością i określa jedną z sześciu predefiniowanych klas: **west-germany, canada, usa, france, uk, japan**. Z uwagi na filtrację zbiór został ograniczony do 13766 dokumentów. Na przefiltrowanym zbiorze przeprowadzono ekstrakcję określonych cech. Następnie metodą *k*-NN i przy pomocy odpowiednich metryk oraz miar podobieństwa tekstów, określono przynależność do predefiniowanych klas.

### 2. Klasyfikacja nadzorowana metodą *k*-NN

Aby sklasyfikować dokumenty tekstowe wykorzystana została jedna z metod klasyfikacji nadzorowanej – metoda *k* Najbliższych Sąsiadów, określana również jako *k*-NN [1]. Naszym zadaniem jest sklasyfikowanie artykułów do odpowiednich klas. Rozważany zbiór dokumentów (a więc tych, które mają przypisane jedno państwo w etykiecie **places** i jest to jedno z następujących sześciu państw: **canada, france, japan, uk, usa, west-germany**)

dzielimy na dwa podzbiory: zbiór uczący i zbiór testowy. W doświadczeniu wykorzystano różne proporcje podziału zbioru:

- 80% – zbiór uczący, 20% – zbiór testowy
- 70% – zbiór uczący, 30% – zbiór testowy
- 60% – zbiór uczący, 40% – zbiór testowy
- 50% – zbiór uczący, 50% – zbiór testowy
- 40% – zbiór uczący, 60% – zbiór testowy
- 30% – zbiór uczący, 70% – zbiór testowy

Aby skorzystać z metody  $k$ -NN, musimy również określić parametr  $k$ , stanowiący liczbę najbliższych sąsiadów braną pod uwagę w trakcie działania algorytmu. W zadaniu wzięto pod uwagę różne parametry  $k$  (1, 2, 3, 5, 10, 20, 35, 50, 75, 100). Następnie każdy artykuł przedstawiamy za pomocą wektorów cech. Poszczególne cechy zostały opisane w rozdziałach od 2.1.1 do 2.1.12, a sam wektor w rozdziale 2.1. Cechy mogą być opisane liczbami, słowami, a także wartościami binarnymi. Cechy liczbowe zostały poddane normalizacji (2.2) do przedziału [0;1]. Po wykonaniu normalizacji cech liczbowych, przy pomocy wybranych metryk (rozdział 3.3) i miar podobieństwa tekstu (rozdział 3.1), liczymy odległość między wektorem opisującym konkretny dokument z części testowej i wektorami opisującymi dokumenty z części uczącej. Następnie ze zbioru uczącego wybieramy  $k$  wektorów, do których klasyfikowany wektor ma najmniejszą odległość. Obiektyowi przydzielimy taki kraj, który ma najczęściej przedstawicieli pośród  $k$  najbliższych wektorów.

### Przykład I :

$$k = 1$$

Dla wybranego artykułu ze zbioru testowego znajdujemy 1 artykuł ze zbioru uczącego, wobec którego odległość między wektorami cech będzie najmniejsza. Następnie sprawdzamy etykietę artykułu, ze zbioru uczącego. Oznaczmy tą etykietę jako  $e_1$ . Niech:

$$e_1 = usa$$

Zatem klasyfikowany artykuł zostanie przyporządkowany do klasy *usa*.

### Przykład II :

$$k = 3$$

Dla wybranego artykułu ze zbioru testowego znajdujemy 3 artykuły ze zbioru uczącego, wobec którego odległość między wektorami cech będzie najmniejsza. Następnie sprawdzamy etykietę artykułów, ze zbioru uczącego. Oznaczmy te etykiety jako  $e_1, e_2, e_3$ . Niech:

$$e_1 = usa$$

$$e_2 = uk$$

$$e_3 = uk$$

Ponieważ klasą *uk* występuje pośród  $k$  najbliższych sąsiadów 2 razy, a klasa *usa* tylko raz, to zgodnie z regułą większości klasyfikowany wybrany do klasyfikacji artykuł zostanie przyporządkowany do klasy *uk*.

### Przykład III :

$$k = 5$$

Dla wybranego artykułu ze zbioru testowego znajdujemy 5 artykułów ze zbioru uczącego, wobec którego odległość między wektorami cech będzie najmniejsza. Następnie sprawdzamy etykietę artykułów, ze zbioru uczącego. Oznaczmy te etykiety jako  $e_1, e_2, e_3, e_4, e_5$ . Niech:

$$e_1 = usa$$

$$e_2 = uk$$

$$e_3 = uk$$

$$e_4 = france$$

$$e_5 = france$$

Zarówno klasy *uk*, jak i *france* występują dwukrotnie pośród  $k$  najbliższych sąsiadów. Wiemy jednak, że klasa *uk* występuje jako 2 i 3 najbliższy sąsiad, z kolei klasa *france* jako 4 i 5. Zatem klasyfikowany artykuł zostanie przypisany do klasy *uk*, ponieważ średnia odległość od najbliższych sąsiadów tej klasy, będzie mniejsza niż średnia odległość od najbliższych sąsiadów należących do klasy *france*.

## 2.1. Ekstrakcja cech, wektory cech

Ekstrakcja cech to proces w którym dla każdego obiektu tworzony jest reprezentujący go wektor cech. Wektor cech składa się z wartości binarnych, liczb lub wartości tekstowych, które identyfikują dany obiekt:

$$w = (C_1, C_2, \dots, C_n) \quad (2.1)$$

gdzie

$n$  - liczba cech, będąca liczbą naturalną; wymiar wektora cech

W zadaniu rozpatrujemy 12 cech, więc wektor cech będzie miał postać :

$$w = (C_1, C_2, \dots, C_{12}) \quad (2.2)$$

Przykładowy wektor cech dla artykułu rozważany w zadaniu:

$$w = (1, 0, U.S., U.S., Texas, "", "", acquisition, Austin, U.S., 0.56, 0.03)$$

Poszczególne cechy zostały opisane poniżej, w rozdziałach od [2.1.1](#) do [2.1.12](#).

### 2.1.1. Występowanie nazw walut w treści artykułu.

Jest to cecha binarna i jej wartość zależy od wystąpienia nazwy waluty, używanej w danych państwach w czasie powstania zbioru (w roku 1987). Nazwy walut zawiera słownik [16]. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie sąbrane pod uwagę w zadaniu. Wartość cechy może przyjmować tylko wartość 0 lub 1.

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.3)$$

$$W = \{\$, deutsche mark, \dots, yen\} \quad (2.4)$$

$$C_1 = \text{currency}(a) \quad (2.5)$$

$$\text{currency}(a) = \begin{cases} 1, & \text{gdy } a \text{ zawiera element z } W \\ 0, & \text{gdy } a \text{ nie zawiera elementu z } W \end{cases} \quad (2.6)$$

$$C_1 \in \{0, 1\} \quad (2.7)$$

gdzie

$W$  - zbiór wszystkich nazw walut występujących w słowniku walut [16];

$A$  - zbiór treści artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$\text{currency}(a)$  - funkcja, która sprawdza, czy w treści artykułu  $a$  występuje nazwa waluty ze zbioru  $W$

**Przykład:**

**Otrzymana wartość cechy dla treści artykułu:** "Average yen cd rates fall in latest week":

$$C_1 = 1$$

### 2.1.2. Występowanie nazwy kraju w tytule dokumentu.

Jest to cecha binarna i jej wartość zależy od wystąpienia nazwy kraju w tytule dokumentu. Nazwy państw zawiera słownik [17]. Dziedzinę stanowi zbiór  $B$ , zbiór tytułów wszystkich artykułów, jakie sąbrane pod uwagę w zadaniu. Wartość cechy może przyjmować tylko wartość 0 lub 1.

$$B = \{b_1, b_2, \dots, b_m\} \quad (2.8)$$

$$P = \{\text{america}, \text{canada}, \dots, \text{west-germany}\} \quad (2.9)$$

$$C_2 = \text{countryTitle}(b) \quad (2.10)$$

$$\text{countryTitle}(b) = \begin{cases} 1, & \text{gdy } b \text{ zawiera element z } P \\ 0, & \text{gdy } b \text{ nie zawiera elementu z } P \end{cases} \quad (2.11)$$

$$C_2 \in \{0, 1\} \quad (2.12)$$

gdzie

$P$  - zbiór wszystkich nazw krajów występujących w słowniku państw [17];

$B$  - zbiór tytułów artykułów;

$b$  - tytuł artykułu;

$m$  - liczba wszystkich artykułów;

$\text{countryTitle}(b)$  - funkcja, która sprawdza, czy w tytule artykułu  $b$  występuje nazwa kraju ze zbioru  $P$

**Przykład:**

**Otrzymana wartość cechy dla tytułu:** "Ronald Reagan visited Canada."

$$C_2 = 1$$

### 2.1.3. Pierwsze słowo z ogólnego słownika, które wystąpi w tytule dokumentu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $B$ , zbiór tytułów wszystkich artykułów, jakie są brane pod uwagę w zadaniu. Do zbioru wartości należą słowa z ogólnego słownika [18].

$$B = \{b_1, b_2, \dots, b_m\} \quad (2.13)$$

$$S = \{\$, A. Lange & Sohne, \dots, yen\} \quad (2.14)$$

$$C_3 = firstInTitle(b) \quad (2.15)$$

gdzie

$S$  - niepusty zbiór słów z ogólnego słownika pojęć [18];

$B$  - zbiór tytułów artykułów;

$b$  - tytuł artykułu;

$m$  - liczba wszystkich artykułów;

$firstInTitle(b)$  - funkcja, zwracająca pierwszą nazwę z tytułu artykułu  $b$ , która należy do zbioru  $S$ . W przypadku, gdy nie będzie takich nazw, zwróci pusty łańcuch tekstowy

**Przykład:**

Otrzymana wartość cechy dla tytułu: "Ronald Reagan visited Canada.":

$$C_3 = Reagan$$

### 2.1.4. Najliczniej występująca nazwa państwa w treści dokumentu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie są brane pod uwagę w zadaniu. Do zbioru wartości należą słowa ze słownika nazw państw [17].

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.16)$$

$$P = \{america, canada, \dots, west - germany\} \quad (2.17)$$

$$C_4 = countryName(a) \quad (2.18)$$

gdzie

$P$  - niepusty zbiór słów ze słownika nazw państw [17];

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$countryName(a)$  - funkcja, zwracająca nazwę ze zbioru  $P$ , która najliczniej występuje w treści artykułu  $a$ . Jeśli liczba wystąpień kilku nazw będzie taka sama, wówczas zwraca nazwę o największej liczbie wystąpień, która wystąpi jako pierwsza. W przypadku, kiedy żadna nazwa państwa ze zbioru  $P$  nie wystąpi w treści artykułu  $a$ , funkcja zwróci pusty łańcuch tekstowy

**Przykład:**

Otrzymana wartość cechy dla tekstu artykułu: *Japan had agreed to try to increase purchases of U.S.-made parts by Japanese car makers and to begin long term contracts for parts purchases, a Commerce department official said.*

$$C_4 = Japan$$

### 2.1.5. Najliczniej występująca nazwa geograficzna ze słownika w treści dokumentu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie są brane pod uwagę w zadaniu. Do zbioru wartości należą słowa ze słownika nazw geograficznych [19].

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.19)$$

$$G = \{Aachen, Abbotsford, \dots, Yukon River\} \quad (2.20)$$

$$C_5 = landName(a) \quad (2.21)$$

gdzie

$G$  - niepusty zbiór słów ze słownika nazw geograficznych [19];

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$landname(a)$  - funkcja, zwracająca nazwę ze zbioru  $G$ , która najliczniej występuje w treści artykułu  $a$ . Jeśli liczba wystąpień kilku nazw będzie taka sama, wówczas zwraca nazwę o największej liczbie wystąpień, która wystąpi jako pierwsza. W przypadku, kiedy żadna nazwa geograficzna ze zbioru  $G$  nie wystąpi w treści artykułu  $a$ , funkcja zwróci pusty łańcuch tekstowy

**Przykład:**

**Otrzymana wartość cechy dla tekstu artykułu:** *"Another on Tuesday registering a preliminary 6.5 on the scale struck the Sea of Japan coast northwest of Tokyo."*:

$$C_5 = Sea of Japan$$

### 2.1.6. Najliczniej występująca sławna postać ze słownika w treści dokumentu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie są brane pod uwagę w zadaniu. Do zbioru wartości należą nazwiska ze słownika sławnych ludzi [20].

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.22)$$

$$O = \{Adenauer, Akihito, \dots, Weizsacker\} \quad (2.23)$$

$$C_6 = famousPerson(a) \quad (2.24)$$

gdzie

$O$  - niepusty zbiór słów ze słownika nazwisk sławnych postaci [20];

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$famousPerson(a)$  - funkcja, zwracająca nazwę ze zbioru  $O$ , która najliczniej występuje w treści artykułu  $a$ . Jeśli liczba wystąpień kilku nazw będzie taka sama, wówczas zwraca nazwę o największej liczbie wystąpień, która wystąpi jako pierwsza. W przypadku kiedy żadne nazwisko ze zbioru  $O$  nie wystąpi w treści artykułu  $a$ , funkcja zwróci pusty łańcuch tekstowy

**Przykład:**

**Otrzymana wartość cechy dla tekstu dokumentu:** *An experienced successor, therefore, would seem a necessity. One widely mentioned possibility is Secretary of State George Shultz, whose experience as Treasury Secretary under President Nixon and background as a trained economist would make him ideal.”:*

$$C_6 = Shultz$$

#### 2.1.7. Najliczniej występująca nazwa firmy ze w treści dokumentu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie sąbrane pod uwagę w zadaniu. Do zbioru wartości należą nazwy ze słownika nazw firm [21].

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.25)$$

$$F = \{ARAG SE, ARMCO, \dots, Ziehl - Abegg SE\} \quad (2.26)$$

$$C_7 = companyName(a) \quad (2.27)$$

gdzie,

$F$  - niepusty zbiór słów ze słownika nazw firm [21];

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$companyName(a)$  - funkcja, zwracająca nazwę ze zbioru  $F$ , która najliczniej występuje w treści artykułu  $a$ . Jeśli, liczba wystąpień kilku nazw będzie taka sama, wówczas zwraca nazwę o największej liczbie wystąpień, która wystąpi jako pierwsza. W przypadku kiedy żadna nazwa firmy ze zbioru  $F$  nie wystąpi w treści artykułu  $a$ , funkcja zwróci pusty łańcuch tekstowy

**Przykład:**

**Otrzymana wartość cechy dla tekstu dokumentu:** *”Nissan Motor Corp U.S.A. said domestic car sales in the February 21 to 28 period rose to 3,501 from 2,578 cars at the same time last year.”*

$$C_7 = Nissan$$

#### 2.1.8. Najdłuższy wyraz w treści artykułu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie sąbrane pod uwagę w zadaniu. Zbiór  $A$  jest również zbiorem wartości cechy.

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.28)$$

$$C_8 = longest(a) \quad (2.29)$$

gdzie

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$longest(a)$  - funkcja zwracająca najdłuższy wyraz pojawiający się w tekście artykułu  $a$ . Jeśli jest kilka takich wyrazów, to funkcja zwróci ten, który pojawi się jako pierwszy;

**Przykład:**

**Dla treści dokumentu:**

*"MTS said it does not expect problems in obtaining New Jersey and Nevada regulatory approval for the acquisition, since ownership in a Caesars stake has already been cleared. In June 1986, Sosnoff requested a seat on the Caesars World board, a request that has not yet been granted."*

**Otrzymana wartość cechy dla przykładowego tekstu artykułu wynosi**

$$C_8 = \text{acquisition}$$

#### 2.1.9. Najdłuższy wyraz ze słownika ogólnego występujący w treści artykułu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie są brane pod uwagę w zadaniu. Zbiorem wartości cechy jest zbiór  $S$  zawierający pojęcia ze słownika ogólnego [18].

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.30)$$

$$S = \{\$, A. Lange & Sohne, \dots, yen\} \quad (2.31)$$

$$C_9 = \text{longestDict}(a) \quad (2.32)$$

gdzie

$S$  - niepusty zbiór słów z ogólnego słownika pojęć [18];

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$\text{longestDict}(a)$  - funkcja zwracająca najdłuższy wyraz pojawiający się w tekście artykułu  $a$  występujący w słowniku ogólnym [18]. Jeśli jest kilka takich wyrazów, to funkcja zwróci ten, który pojawi się jako pierwszy;

**Przykład:**

**Dla tekstu dokumentu:** *"Honda Motor Co Ltd of Japan's American Honda Motor Co Inc unit said its February sales rose to 56,704 from 48,443 a year ago. The sales figures include sales of 7,056 cars from its new Acura division, which was not in place a year ago. Year to date sales totaled 102,751 at the end of February, up from 98,724. This included sales of 12,723 from the Acura division. In the company's Honda division, sales of the Accord model led the monthly and year-to-date totals, followed by Civic sales, then Prelude sales. In the Acura division, Intera sales outpaced Legend sales."*

**Otrzymana wartość cechy dla przykładowego tekstu artykułu wynosi:**

$$C_9 = \text{American}$$

#### 2.1.10. Ostatni wyraz ze słownika ogólnego występujący w treści artykułu.

Jest to cecha tekstowa. Dziedzinę stanowi zbiór  $A$ , zbiór tytułów wszystkich artykułów, jakie są brane pod uwagę w zadaniu. Do zbioru wartości należą słowa z ogólnego słownika [18].

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.33)$$

$$S = \{\$, A. Lange \& Sohne, ..., yen\} \quad (2.34)$$

$$C_{10} = lastInDict(a) \quad (2.35)$$

gdzie

$S$  - niepusty zbiór słów z ogólnego słownika pojęć [18];

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$lastInDict(a)$  - funkcja zwracająca ostatni wyraz ze słownika ogólnego [18] pojawiający się w tekście artykułu  $a$ . Jeśli w tekście artykułu  $a$  nie wystąpi żaden wyraz ze słownika, funkcja zwraca pusty łańcuch tekstu; **Przykład:**

**Dla tekstu artykułu:**

*"Honda Motor Co Ltd of Japan's American Honda Motor Co Inc unit said its February sales rose to 56,704 from 48,443 a year ago. The sales figures include sales of 7,056 cars from its new Acura division, which was not in place a year ago. Year to date sales totaled 102,751 at the end of February, up from 98,724. This included sales of 12,723 from the Acura division. In the company's Honda division, sales of the Accord model led the monthly and year-to-date totals, followed by Civic sales, then Prelude sales. In the Acura division, Intera sales outpaced Legend sales."*

$$C_{10} = Prelude$$

### 2.1.11. Liczba słów wielką literą w treści dokumentu.

Jest to cecha liczbowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie sąbrane pod uwagę w zadaniu. Oznacza liczbę słów zaczynających się wielką literą, jaka wystąpiła w treści dokumentu. Liczba ta jest liczbą naturalną, zatem należy dokonać jej normalizacji (2.2). Po tym procesie wzór cechy ostatecznie przyjmuje postać (2.37). Dzięki temu finalnie zbiór wartości cechy zawiera się w przedziale  $[0;1]$ .

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.36)$$

$$C_{11} = \frac{big(a)}{len(a)} \quad (2.37)$$

$$C_{11} \in [0; 1] \quad (2.38)$$

gdzie

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$big(a)$  - funkcja, zwracająca liczbę wyrazów zaczynających się wielką literą w treści artykułu  $a$ ;

$len(a)$  - funkcja, zwracająca liczbę wyrazów w treści artykułu  $a$ ;

**Przykład:**

**Dla tekstu artykułu:** *"Morrison Knudsen Corp said its National Projects Inc subsidiary was awarded a contract worth about 27 mln dlrs for construction of improvements at the U.S. Navy's Fallon Naval Air Station in Nevada."*

**Liczba słów zaczynających się wielką literą w przykładowym tekście artykułu:**

$$big(a) = 13$$

**Liczba słów w tekście przykładowego artykułu:**

$$len(a) = 32$$

**Otrzymana wartość cechy:**

$$C_{11} = \frac{13}{32} \approx 0.41$$

### 2.1.12. Liczba słów z ogólnego słownika występująca w treści dokumentu

Jest to cecha liczbowa. Dziedzinę stanowi zbiór  $A$ , zbiór treści wszystkich artykułów, jakie sąbrane pod uwagę w zadaniu. Oznacza liczbę słów z ogólnego słownika [18], jaka wystąpiła w treści dokumentu. Liczba ta jest liczbą naturalną, zatem należy dokonać jej normalizacji (2.2). Po tym procesie wzór cechy ostatecznie przyjmuje postać (2.41). Dzięki temu finalnie zbiór wartości cechy zawiera się w przedziale  $[0;1]$ .

$$A = \{a_1, a_2, \dots, a_m\} \quad (2.39)$$

$$S = \{\$, A. Lange & Sohne, \dots, yen\} \quad (2.40)$$

$$C_{12} = \frac{countDictOcc(a)}{len(a)} \quad (2.41)$$

$$C_{12} \in [0; 1] \quad (2.42)$$

gdzie

$A$  - zbiór artykułów;

$a$  - treść artykułu;

$m$  - liczba wszystkich artykułów;

$S$  - niepusty zbiór słów ze ogólnego słownika [18];

$countDictOcc(a)$  - funkcja, zwracająca liczbę wystąpień wyrazów ze słownika [18] w treści artykułu  $a$ ;

$len(a)$  - funkcja, zwracająca liczbę wyrazów w treści artykułu  $a$ ;

**Przykład:**

**Dla tekstu dokumentu:** "Canada's industrial product price index rose 0.2 pct in January after falling 0.2 pct in each of the two previous months, Statistics Canada said. The rise was led by price gains for papers, pharmaceuticals and petroleum and coal products. Price declines were recorded for meat products, lumber and motor vehicles. On a year over year basis, the federal agency said the index fell 0.9 pct in January, the largest yearly decline on record."

**Liczba słów ze słownika ogólnego [18] występujące w przykładowym tekście artykułu:**

$$countDictOcc(a) = 2$$

**Liczba słów w tekście artykułu:**

$$len(a) = 73$$

**Otrzymana wartość cechy:**

$$C_{12} = \frac{2}{73} \approx 0.03$$

## 2.2. Normalizacja

Niektóre cechy liczbowe wymagają normalizacji. Normalizacja to proces przekształcenia wartości w taki sposób, aby spełniała określone kryteria. W naszym przypadku, wartość cechy musi zawierać się w przedziale  $[0;1]$ . W celu dokonania normalizacji dzielimy otrzymaną liczbę słów spełniających daną cechę przez liczbę wszystkich słów w przetwarzanym tekście.

**Przykład:**

Cecha przedstawiająca liczbę słów występujących w tekście na literę  $m$ :

Liczba słów na literę  $m$  w tekście artykułu: 12;

Liczba słów w tekście artykułu: 80;

$$C = \frac{12}{80} = 0.15$$

## 2.3. Miary jakości klasyfikacji

Aby ocenić jakość klasyfikacji stosuje się następujące miary jakości klasyfikacji:

- Accuracy,
- Precision,
- Recall,
- F1.

Do korzystania z powyższych miar należy użyć tablicy pomyłek [14].

W zadaniu zdefiniowane jest 6 klas **west-germany, canada, usa, france, uk, japan**. Należy zmodyfikować tablice pomyłek tak aby można było ustalić liczbę przypadków przewidzianych poprawnie i błędnie dla zdefiniowanej w zadaniu liczbie klas.

### 2.3.1. Accuracy (ACC)

*Accuracy* - jest to miara dokładności która określa stosunek liczby poprawnie sklasyfikowanych artykułów danej klasy do liczby wszystkich artykułów w zbiorze testowym. *Accuracy* określa udział poprawnych predykcji do całkowitej liczby predykcji. *Accuracy* jest liczona tylko jeden raz i określa jakość całego procesu klasyfikacji.

$$\text{ACC} = \frac{TP_1 + TP_2 + \dots + TP_n}{N} \quad (2.43)$$

gdzie

$TP_n$  - liczba poprawnych klasyfikacji artykułów dla  $n$ -tej klasy,

$N$  - liczby wszystkich artykułów w zbiorze testowym,

**Przykład:**

Liczba wszystkich artykułów dla każdej klasy wynosi 20. Założymy, że liczby poprawnie sklasyfikowanych dokumentów dla następujących klas wynoszą:

$$TP_{usa} = 10, TP_{uk} = 9, TP_{japan} = 15, TP_{west-germany} = 7,$$

$$TP_{france} = 3, TP_{canada} = 11,$$

$$ACC = \frac{10 + 9 + 15 + 7 + 3 + 11}{120} = \frac{55}{120} = 0.46 \quad (2.44)$$

### 2.3.2. Precision (PPV)

*Precision* - miara jakości klasyfikacji która określa, jak wiele z wykrytych przez klasyfikator artykułów faktycznie należy do konkretnej klasy w stosunku do wszystkich wykrytych artykułów tej klasy. *Precision* jest obliczana indywidualnie dla każdej klasy zdefiniowanej w zadaniu. Miara *Precision* wyrażona jest wzorem:

$$PPV_n = \frac{TP_n}{TP_n + FP_n} \quad (2.45)$$

$TP_n$  - liczba poprawnych klasyfikacji artykułów dla  $n$ -tej klasy,

$FP_n$  - liczba błędnie sklasyfikowanych artykułów dla  $n$ -tej klasy,

#### Przykład:

Dla klasy *japan* liczebność zbioru wynosi 20 artykułów, poprawnie sklasyfikowanych artykułów dla klasy *japan* zostało 10 i 8 zostało sklasyfikowanych niepoprawnie. Wartość miary *Precision* dla klasy *japan* wynosi:

$$PPV_{japan} = \frac{10}{10 + 8} = \frac{10}{18} \approx 0.55$$

Aby obliczyć miarę *Precision* dla klasyfikacji wszystkich klas stosujemy średnią ważoną, w której wagą jest udział  $n$ -tej klasy w testowym zbiorze artykułów.

$$\overline{PPV} = \frac{PPV_1 \cdot v_1 + PPV_2 \cdot v_2 + \dots + PPV_n \cdot v_n}{v_1 + v_2 + \dots + v_n}$$

gdzie

$v_n$  - waga, liczba artykułów  $n$ -tej klasy w testowym zbiorze artykułów.

#### Przykład:

Podane są następujące wyniki klasyfikacji dla poszczególnych klas:

- Dla klasy *canada* liczącej 30 artykułów, poprawnie sklasyfikowanych artykułów zostało 10 i 16 zostało sklasyfikowanych niepoprawnie.
- Dla klasy *france* liczącej 10 artykułów, poprawnie sklasyfikowanych artykułów zostało 6 i 2 zostało sklasyfikowanych niepoprawnie.
- Dla klasy *japan* liczącej 20 artykułów, poprawnie sklasyfikowanych artykułów zostało 10 i 8 zostało sklasyfikowanych niepoprawnie.
- Dla klasy *uk* liczącej 40 artykułów, poprawnie sklasyfikowanych artykułów zostało 22 i 33 zostało sklasyfikowanych niepoprawnie.
- Dla klasy *usa* liczącej 120 artykułów, poprawnie sklasyfikowanych artykułów zostało 100 i 15 zostało sklasyfikowanych niepoprawnie.
- Dla klasy *west-germany* liczącej 5 artykułów, poprawnie sklasyfikowanych artykułów zostało 2 i 1 zostało sklasyfikowanych niepoprawnie.

$$PPV_{canada} = \frac{10}{10 + 16} = \frac{5}{13} \approx 0.38$$

$$PPV_{france} = \frac{6}{6 + 2} = \frac{3}{4} = 0.75$$

$$PPV_{japan} = \frac{10}{10+8} = \frac{5}{9} \approx 0.55$$

$$PPV_{uk} = \frac{22}{22+33} = \frac{2}{5} = 0.40$$

$$PPV_{usa} = \frac{100}{100+15} = \frac{20}{23} \approx 0.87$$

$$PPV_{west-germany} = \frac{2}{2+1} = \frac{2}{3} \approx 0.66$$

Wartość miary precyzji dla całego procesu klasyfikacji wynosi:

$$\overline{PPV} = \frac{\frac{5}{13} \cdot 26 + \frac{3}{4} \cdot 8 + \frac{5}{9} \cdot 18 + \frac{2}{5} \cdot 30 + \frac{20}{23} \cdot 115 + \frac{2}{3} \cdot 3}{26 + 8 + 18 + 30 + 115 + 3} = 0.70$$

Wyższa wartość *Precision* oznacza, że klasyfikator lepiej radzi sobie z poprawnym rozpoznaniem artykułów należących do właściwej klasy.

### 2.3.3. Recall (TPR)

*Recall* (czułość) jest to stosunek liczby poprawnie sklasyfikowanych artykułów tej samej klasy do liczby wszystkich artykułów tej klasy. Miara jakości *Recall* określa, jak wiele klas prawidłowo zostało wykrytych i poprawnie sklasyfikowanych. Im większa wartość czułości tym liczba błędnych predykcji jest mniejsza. Dla każdej klasy zdefiniowanej w zadaniu czułość jest obliczana indywidualnie i wyrażana jest wzorem:

$$TPR_n = \frac{TP_n}{TP_n + FN_n} \quad (2.46)$$

$TP_n$  - liczba poprawnych klasyfikacji artykułów dla  $n$ -tej klasy,

$FN_n$  - liczba błędnie sklasyfikowanych artykułów dla  $n$ -tej klasy jako artykuły innej  $n$ -tej klasy,

**Przykład:**

Dla klasy *usa* liczebność zbioru artykułów wynosi 120, poprawnie sklasyfikowanych artykułów dla klasy *usa* jest 100 a 20 artykułów zostało sklasyfikowane jako artykuły innych klas.

Zakładamy, że  $TP_{usa} = 100$  a  $FN_{usa} = 20$

Wtedy *Recall* dla klasy *usa* wynosi:

$$TPR_{usa} = \frac{TP_{usa}}{TP_{usa} + FN_{usa}} = \frac{100}{100 + 20} = \frac{100}{120} \approx 0.83$$

Aby obliczyć miarę *Recall* dla klasyfikacji wszystkich klas zastosujemy średnią ważoną, w której wagą jest  $v_n$ , liczebność  $n$ -tej klasy w testowym zbiorze artykułów.

$$\overline{TPR} = \frac{TPR_1 \cdot v_1 + TPR_2 \cdot v_2 + \dots + TPR_n \cdot v_n}{v_1 + v_2 + \dots + v_n}$$

gdzie

$v_n$  - waga, liczba artykułów  $n$ -tej klasy w testowym zbiorze artykułów.

**Przykład:**

Podane są następujące wyniki klasyfikacji dla poszczególnych klas:

- Dla klasy *canada* liczącej 30 artykułów, poprawnie sklasyfikowanych artykułów zostało 10 i 20 zostało sklasyfikowanych jako artykuły innej klasy.
- Dla klasy *france* liczącej 10 artykułów, poprawnie sklasyfikowanych artykułów zostało 6 i 4 zostało sklasyfikowanych jako artykuły innej klasy.
- Dla klasy *japan* liczącej 20 artykułów, poprawnie sklasyfikowanych artykułów zostało 10 i 10 zostało sklasyfikowanych jako artykuły innej klasy.
- Dla klasy *uk* liczącej 40 artykułów, poprawnie sklasyfikowanych artykułów zostało 22 i 18 zostało sklasyfikowanych jako artykuły innej klasy.
- Dla klasy *usa* liczącej 120 artykułów, poprawnie sklasyfikowanych artykułów zostało 100 i 20 zostało sklasyfikowanych jako artykuły innej klasy.
- Dla klasy *west-germany* liczącej 5 artykułów, poprawnie sklasyfikowanych artykułów zostało 2 i 3 zostało sklasyfikowanych jako artykuły innej klasy.

$$TPR_{canada} = \frac{10}{10 + 20} = \frac{1}{3} \approx 0.33$$

$$TPR_{france} = \frac{6}{6 + 4} = \frac{3}{5} = 0.60$$

$$TPR_{japan} = \frac{10}{10 + 10} = \frac{1}{2} = 0.50$$

$$TPR_{uk} = \frac{22}{22 + 18} = \frac{11}{20} = 0.55$$

$$TPR_{usa} = \frac{100}{100 + 20} = \frac{5}{6} \approx 0.83$$

$$TPR_{west-germany} = \frac{2}{2 + 3} = \frac{2}{5} = 0.40$$

Wartość miary czułości dla całego procesu klasyfikacji wynosi:

$$\overline{TPR} = \frac{\frac{1}{3} \cdot 30 + \frac{3}{5} \cdot 10 + \frac{1}{2} \cdot 20 + \frac{11}{20} \cdot 40 + \frac{5}{6} \cdot 120 + \frac{2}{5} \cdot 5}{30 + 10 + 20 + 40 + 120 + 5} = 0.66$$

#### 2.3.4. F1

*F1* - jest to miara klasyfikacji która przedstawiona jest jako średnia harmoniczna precyzji i czułości. Dla każdej klasy zdefiniowanej w zadaniu miara *F1* jest obliczana indywidualnie i wyrażana jest wzorem:

$$F1_n = 2 \cdot \frac{PPV_n \cdot TPR_n}{PPV_n + TPR_n} \quad (2.47)$$

gdzie

*PPV<sub>n</sub>* - precyza (2.45) klasyfikacji artykułów dla *n*-tej klasy,

*TPR<sub>n</sub>* - czułość (2.46) klasyfikacji artykułów dla *n*-tej klasy,

*v<sub>n</sub>* - waga, liczba artykułów *n*-tej klasy w testowym zbiorze artykułów.

**Przykład:**

Zakładając, że wartość precyzji dla klasy *canada* *PPV<sub>canada</sub>* = 0.38 i wartość czułości dla klasy *canada* *TPR<sub>canada</sub>* = 0.33 można wyznaczyć miarę *F1* dla klasy *canada*:

$$F1_{canada} = 2 \cdot \frac{PPV_{canada} \cdot TPR_{canada}}{PPV_{canada} + TPR_{canada}} = 2 \cdot \frac{0.38 \cdot 0.33}{0.38 + 0.33} \approx 0.35$$

Aby obliczyć miarę  $F1$  dla klasyfikacji wszystkich klas stosujemy średnią ważoną, w której wagą jest  $v_n$ , liczebność dla n-tej klasy w zbiorze testowym.

$$\overline{F1} = \frac{F1_1 \cdot v_1 + F1_2 \cdot v_2 + \dots + F1_n \cdot v_n}{v_1 + v_2 + \dots + v_n}$$

### Przykład:

- Klasa *canada* licząca 30 artykułów,
- Klasy *france* licząca 10 artykułów,
- Klasy *japan* licząca 20 artykułów,
- Klasy *uk* licząca 40 artykułów,
- Klasy *usa* licząca 120 artykułów,
- Klasy *west-germany* licząca 5 artykułów,

Wykorzystane zostaną policzone wcześniej miary jakości PPV i TPR dla poszczególnych klas ([2.3.2](#), [2.3.3](#))

$$\begin{aligned} PPV_{canada} &= \frac{5}{13}, \quad TPR_{canada} = \frac{1}{3} \\ PPV_{france} &= \frac{3}{4}, \quad TPR_{france} = \frac{3}{5} \\ PPV_{japan} &= \frac{5}{9}, \quad TPR_{japan} = \frac{1}{2} \\ PPV_{uk} &= \frac{2}{5}, \quad TPR_{uk} = \frac{11}{20} \\ PPV_{usa} &= \frac{20}{23}, \quad TPR_{usa} = \frac{5}{6} \\ PPV_{west-germany} &= \frac{2}{3}, \quad TPR_{west-germany} = \frac{2}{5} \end{aligned}$$

Wartość miary F1 dla poszczególnych klas:

$$\begin{aligned} F1_{canada} &= 2 \cdot \frac{\frac{5}{13} \cdot \frac{1}{3}}{\frac{5}{13} + \frac{1}{3}} = \frac{5}{14} \approx 0.36 \\ F1_{france} &= 2 \cdot \frac{\frac{3}{4} \cdot \frac{3}{5}}{\frac{3}{4} + \frac{3}{5}} = \frac{2}{3} \approx 0.66 \\ F1_{japan} &= 2 \cdot \frac{\frac{5}{9} \cdot \frac{1}{2}}{\frac{5}{9} + \frac{1}{2}} = \frac{10}{19} \approx 0.53 \\ F1_{uk} &= 2 \cdot \frac{\frac{2}{5} \cdot \frac{11}{20}}{\frac{2}{5} + \frac{11}{20}} = \frac{44}{95} \approx 0.46 \\ F1_{usa} &= 2 \cdot \frac{\frac{20}{23} \cdot \frac{5}{6}}{\frac{20}{23} + \frac{5}{6}} = \frac{40}{47} \approx 0.85 \\ F1_{west-germany} &= 2 \cdot \frac{\frac{2}{3} \cdot \frac{2}{5}}{\frac{2}{3} + \frac{2}{5}} = \frac{1}{2} = 0.50 \end{aligned}$$

Wartość miary czułości dla całego procesu klasyfikacji wynosi:

$$\overline{F1} = \frac{\frac{5}{14} \cdot 30 + \frac{2}{3} \cdot 10 + \frac{10}{19} \cdot 20 + \frac{44}{95} \cdot 40 + \frac{40}{47} \cdot 120 + \frac{1}{2} \cdot 5}{30 + 10 + 20 + 40 + 120 + 5} = 0.67$$

### 3. Klasyfikacja z użyciem metryk i miar podobieństwa tekstów

Odległość między wektorami została policzona z użyciem różnych metryk. Ponieważ wektory zawierają również cechy tekstowe, w przypadku tych cech policzone zostało podobieństwo tekstów. Następnie podobieństwo tekstów zostało przekształcone na odległość zgodnie ze wzorem (3.4) W zadaniu wykorzystano 4 cechy liczbowe (2.1.1, 2.1.2, 2.1.11, 2.1.12) i 8 cech tekstowych (od 2.1.3 do 2.1.10). Wykorzystane miary podobieństwa przedstawiają wzory 3.1 i 3.2. Wykorzystane metryki w zadaniu zostały przedstawione w rozdziałach od 3.3.1 do 3.3.3.

#### 3.1. Wybrane miary podobieństwa

##### 3.1.1. Metoda n-gramów

Metoda n-gramów [23] w naszym zadaniu zostanie użyta dla  $n = 3$  i  $n = 4$ . Dla  $n = 3$  metoda jest zwana metodą *trigramów*, a jej wzór przybiera postać:

$$sim_3(s_1, s_2) = \frac{1}{N-2} \sum_{i=1}^{N-2} h(i) \quad (3.1)$$

Z kolei dla  $n = 4$ , metoda określana jest jako metoda *tetragramów*, a jej wzór przybiera postać:

$$sim_4(s_1, s_2) = \frac{1}{N-3} \sum_{i=1}^{N-3} h(i) \quad (3.2)$$

gdzie

$N$  - długość dłuższego z porównywanych słów;

$$h(i) = \begin{cases} 1, & \text{gdy } n\text{-literowy wyraz od } i\text{ - tej pozycji } s_1 \text{ występuje w } s_2 \\ 0, & \text{gdy nie występuje} \end{cases} \quad (3.3)$$

##### Przykład:

Porównanie słów  $s_1 = \text{Missouri}$  i  $s_2 = \text{Mississippi}$ ,  $N(s_1) = 8, N(s_2) = 11, N = \max\{N(s_1), N(s_2)\} = 11$

Dla wzoru (3.1):

$$sim_3(s_1, s_2) = \frac{1}{9} \sum_{i=1}^9 h(i) = \frac{3}{9} \approx 0.33$$

Ponieważ 3 trigramy w Missouri: Mis, iss i iss występują w Mississippi.  
Dla wzoru (3.2):

$$sim_4(s_1, s_2) = \frac{1}{8} \sum_{i=1}^8 h(i) = \frac{1}{8} = 0.125$$

Ponieważ tylko 1 tetagram w Missouri: Miss występuje w Mississippi.

### 3.2. Podobieństwo a metryka

Obliczone podobieństwo należy przekształcić na odległość. Aby określić odległość między tekstami, skorzystano ze wzoru:

$$d(t_1, t_2) = 1 - sim(t_1, t_2) \quad (3.4)$$

gdzie

$t_1$  - cecha tekstowa należąca do pierwszego wektora;

$t_2$  - cecha tekstowa należąca do drugiego wektora, równoległa do  $t_1$ ;

$sim(t_1, t_2)$  - wartość funkcji miary podobieństwa między tekstami;  $t_1$  i  $t_2$

**Przykład:**

$$t_1 = "Missouri", t_2 = "Mississippi"$$

Zgodnie ze wzorem (3.1) wartość funkcji podobieństwa wynosi:

$$sim_3(t_1, t_2) = 0.33$$

Odległość tych cech wynosi:

$$d(t_1, t_2) = 1 - 0.33 = 0.67$$

### 3.3. Wybrane metryki

#### 3.3.1. Metryka euklidesowa

Metryka euklidesowa [22] w zadaniu określona jest wzorem:

$$\rho_E(u, t) = \sqrt{\sum_{i=1}^{12} d_E(t_i, u_i)^2} \quad (3.5)$$

$$t = \{t_1, t_2, \dots, t_{12}\} \quad u = \{u_1, u_2, \dots, u_{12}\} \quad (3.6)$$

gdzie

$t, u$  - wektory cech dla dwóch różnych artykułów ze zbioru rozważanych artykułów o postaci przedstawionej w sekcji (2.1);

$t_i, u_i$  - wartości i-tych cech z wektorów  $t, u$ ;

$d_E$  - funkcja porównująca wartości cech (3.7);

$$d_E(t_i, u_i) = \begin{cases} t_i - u_i, & \text{dla } i \in \{1, 2, 11, 12\} \\ 1 - sim(t_i, u_i), & \text{dla } i \in \{3, 4, 5, 6, 7, 8, 9, 10\} \end{cases} \quad (3.7)$$

gdzie

$sim()$  - funkcja podobieństwa dla wartości cech tekstowych. W zależności od wyboru użytkownika określona wzorem (3.1) lub (3.2).

**Przykład:**

$$t = (0, 1, \text{canada}, \text{Canada}, \text{Canadian}, \text{Jobs}, \text{Chrysler}, \text{inflation-indexed}, \text{Canadian}, \text{canada}, 0.16, 0.04)$$

$$u = (1, 1, \text{canada}, \text{Canada}, \text{Canadian}, "", "", \text{development}, \text{Canadian}, \text{Canadian}, 0.16, 0.09)$$

Wybrana miara podobieństwa tekstów: metoda *trigramów* (3.1).

$$\begin{aligned}
d_E(t_1, u_1) &= 0 - 1 = 1 \\
d_E(t_2, u_2) &= 1 - 1 = 0 \\
d_E(t_3, u_3) &= 1 - sim(canada, canada) = 1 - 1 = 0 \\
d_E(t_4, u_4) &= 1 - sim(Canada, Canada) = 1 - 1 = 0 \\
d_E(t_5, u_5) &= 1 - sim(Canadian, Canadian) = 1 - 1 = 0 \\
d_E(t_6, u_6) &= 1 - sim(Jobs, "") = 1 - 0 = 1 \\
d_E(t_7, u_7) &= 1 - sim(Chrysler, "") = 1 - 0 = 1 \\
d_E(t_8, u_8) &= 1 - sim(inflation-indexed, development) = 1 - 0 = 1 \\
d_E(t_9, u_9) &= 1 - sim(Canadian, Canadian) = 1 - 1 = 0 \\
d_E(t_{10}, u_{10}) &= 1 - sim(canada, Canadian) = 1 - \frac{1}{2} = 0.5 \\
d_E(t_{11}, u_{11}) &= 0.16 - 0.16 = 0 \\
d_E(t_{12}, u_{12}) &= 0.04 - 0.09 = -0.05 \\
\rho_E(t, u) &= \sqrt{1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0.5^2 + 0^2 + (-0.05)^2} \\
\rho_E(t, u) &\approx 2.06
\end{aligned}$$

### 3.3.2. Metryka miejska

Metryka *miejska* [24] znana jest także jako metryka Manhattan lub metryka taksówkowa. W zadaniu określona jest wzorem:

$$\rho_M(t, u) = \sqrt{\sum_{i=1}^{12} |d_M(t_i, u_i)|} \quad (3.8)$$

$$t = \{t_1, t_2, \dots, t_{12}\} \quad u = \{u_1, u_2, \dots, u_{12}\} \quad (3.9)$$

gdzie

$t, u$  - wektory cech dla dwóch różnych artykułów ze zbioru rozważanych artykułów o postaci przedstawionej w sekcji (2.1);

$t_i, u_i$  - wartości i-tych cech z wektorów  $t, u$ ;

$d_M$  - funkcja porównująca wartości cech (3.10);

$$d_M(t_i, u_i) = \begin{cases} t_i - u_i, & \text{dla } i \in \{1, 2, 11, 12\} \\ 1 - sim(t_i, u_i), & \text{dla } i \in \{3, 4, 5, 6, 7, 8, 9, 10\} \end{cases} \quad (3.10)$$

gdzie

$sim()$  - funkcja podobieństwa dla wartości cech tekstowych. W zależności od wyboru użytkownika określona wzorem (3.1) lub (3.2).

### Przykład:

$$t = (0, 1, \text{canada}, \text{Canada}, \text{Canadian}, \text{Jobs}, \text{Chrysler}, \text{inflation-indexed}, \text{Canadian}, \text{canada}, 0.16, 0.04)$$

$$u = (1, 1, \text{canada}, \text{Canada}, \text{Canadian}, , \text{Development}, \text{Canadian}, \text{Canadian}, 0.16, 0.09)$$

Wybrana miara podobieństwa tekstów: metoda *trigramów* (3.1).

$$d_M(t_1, u_1) = 0 - 1 = -1$$

$$d_M(t_2, u_2) = 1 - 1 = 0$$

$$d_M(t_3, u_3) = 1 - sim(\text{canada}, \text{canada}) = 1 - 1 = 0$$

$$d_M(t_4, u_4) = 1 - sim(\text{Canada}, \text{Canada}) = 1 - 1 = 0$$

$$d_M(t_5, u_5) = 1 - sim(\text{Canadian}, \text{Canadian}) = 1 - 1 = 0$$

$$d_M(t_6, u_6) = 1 - sim(\text{Jobs}, "") = 1 - 0 = 1$$

$$d_M(t_7, u_7) = 1 - sim(\text{Chrysler}, "") = 1 - 0 = 1$$

$$d_M(t_8, u_8) = 1 - sim(\text{inflation-indexed}, \text{development}) = 1 - 0 = 1$$

$$d_M(t_9, u_9) = 1 - sim(\text{Canadian}, \text{Canadian}) = 1 - 1 = 0$$

$$d_M(t_{10}, u_{10}) = 1 - sim(\text{canada}, \text{Canadian}) = 1 - \frac{1}{2} = 0.5$$

$$d_M(t_{11}, u_{11}) = 0.16 - 0.16 = 0$$

$$d_M(t_{12}, u_{12}) = 0.04 - 0.09 = -0.05$$

$$\rho_M(t, u) = |-1| + |0| + |0| + |0| + |0| + |1| + |1| + |1| + |0| + |0.5| + |0| + |-0.05|$$

$$\rho_M(t, u) = 4.55$$

### 3.3.3. Metryka Czebyszewa

Metryka Czebyszewa [24] w zadaniu określona jest wzorem:

$$\rho_{Ch}(t, u) = \max_i |d_{Ch}(t_i, u_i)| \quad (3.11)$$

$$t = \{t_1, t_2, \dots, t_{12}\} \quad u = \{u_1, u_2, \dots, u_{12}\} \quad (3.12)$$

gdzie

$t, u$  - wektory cech dla dwóch różnych artykułów ze zbioru rozważanych artykułów o postaci przedstawionej w sekcji (2.1);

$t_i, u_i$  - wartości i-tych cech z wektorów  $t, u$ ;

$d_{Ch}$  - funkcja porównująca wartości cech (3.13);

$$d_{Ch}(t_i, u_i) = \begin{cases} t_i - u_i, & \text{dla } i \in \{1, 2, 11, 12\} \\ 1 - sim(t_i, u_i), & \text{dla } i \in \{3, 4, 5, 6, 7, 8, 9, 10\} \end{cases} \quad (3.13)$$

gdzie

$sim()$  - funkcja podobieństwa dla wartości cech tekstowych. W zależności od wyboru użytkownika określona wzorem (3.1) lub (3.2).

### Przykład:

$$t = (0, 1, \text{canada}, \text{Canada}, \text{Canadian}, \text{Jobs}, \text{Chrysler}, \text{inflation-indexed}, \\ \text{Canadian}, \text{canada}, 0.16, 0.04)$$

$$u = (1, 1, \text{canada}, \text{Canada}, \text{Canadian}, , , \text{Development}, \text{Canadian}, \\ \text{Canadian}, 0.16, 0.09)$$

Wybrana miara podobieństwa tekstów: metoda *trigramów* (3.1).

$$d_{Ch}(t_1, u_1) = 0 - 1 = -1$$

$$d_{Ch}(t_2, u_2) = 1 - 1 = 0$$

$$d_{Ch}(t_3, u_3) = 1 - sim(\text{canada}, \text{canada}) = 1 - 1 = 0$$

$$d_{Ch}(t_4, u_4) = 1 - sim(\text{Canada}, \text{Canada}) = 1 - 1 = 0$$

$$d_{Ch}(t_5, u_5) = 1 - sim(\text{Canadian}, \text{Canadian}) = 1 - 1 = 0$$

$$d_{Ch}(t_6, u_6) = 1 - sim(\text{Jobs}, "") = 1 - 0 = 1$$

$$d_{Ch}(t_7, u_7) = 1 - sim(\text{Chrysler}, "") = 1 - 0 = 1$$

$$d_{Ch}(t_8, u_8) = 1 - sim(\text{inflation} - \text{indexed}, \text{development}) = 1 - 0 = 1$$

$$d_{Ch}(t_9, u_9) = 1 - sim(\text{Canadian}, \text{Canadian}) = 1 - 1 = 0$$

$$d_{Ch}(t_{10}, u_{10}) = 1 - sim(\text{canada}, \text{Canadian}) = 1 - \frac{1}{2} = 0.5$$

$$d_{Ch}(t_{11}, u_{11}) = 0.16 - 0.16 = 0$$

$$d_{Ch}(t_{12}, u_{12}) = 0.04 - 0.09 = -0.05$$

$$\rho_{Ch}(t, u) = \max_i |d_{Ch}(t_i, u_i)|$$

$$\rho_{Ch}(t, u) = 1$$

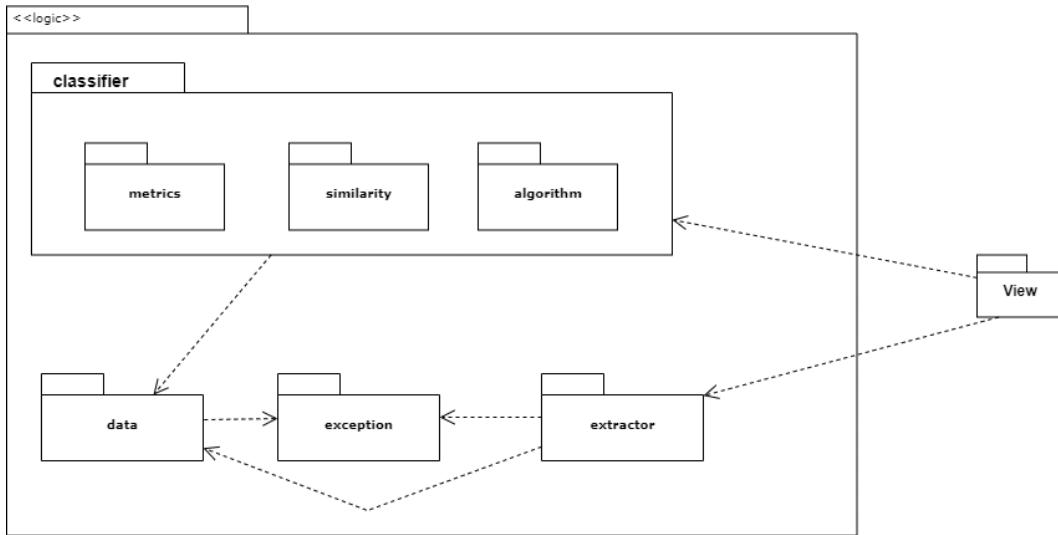
## 4. Budowa aplikacji

Aplikacja składa się z dwóch modułów. W ten sposób oddzielono od siebie graficzny interfejs użytkownika oraz część odpowiedzialną za logikę algorytmu.

### 4.1. Diagramy UML

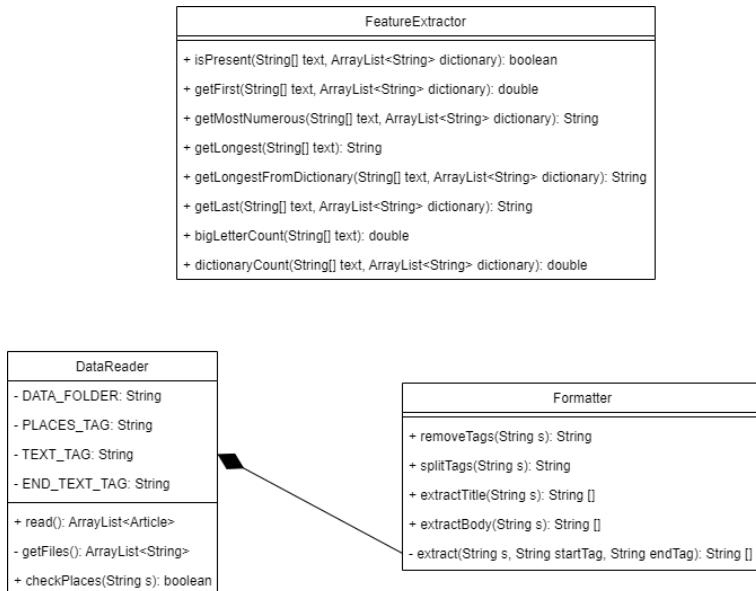
#### 4.1.1. Diagram pakietów

Pierwszy moduł, *view* zawiera graficzny interfejs użytkownika. Drugi moduł, *logic* odpowiada za logikę aplikacji. Moduł ten jest podzielony na mniejsze pakiety. Pierwszym z nich jest *extractor*. Zawiera on część logiki odpowiedzialną za wczytanie plików z danymi, a także wyekstrahowanie wektorów cech. Wczytane i przetworzone dane przechowywane są w klasach z pakietu *data*. Dane te są wykorzystywane przez pakiet *classifier*. Pakiet ten został podzielony na 3 mniejsze podpakiety. Pakiet *metrics* zawiera implementacje wybranych metryk, wykorzystywanych w trakcie algorytmu. Pakiet *similarity* zawiera implementacje wybranych miar podobieństwa tekstów, a pakiet *algorithm* implementację metody klasyfikacji *k* - NN.



Rysunek 1. Diagram pakietów

#### 4.1.2. Diagram klas - pakiet ekstraktora



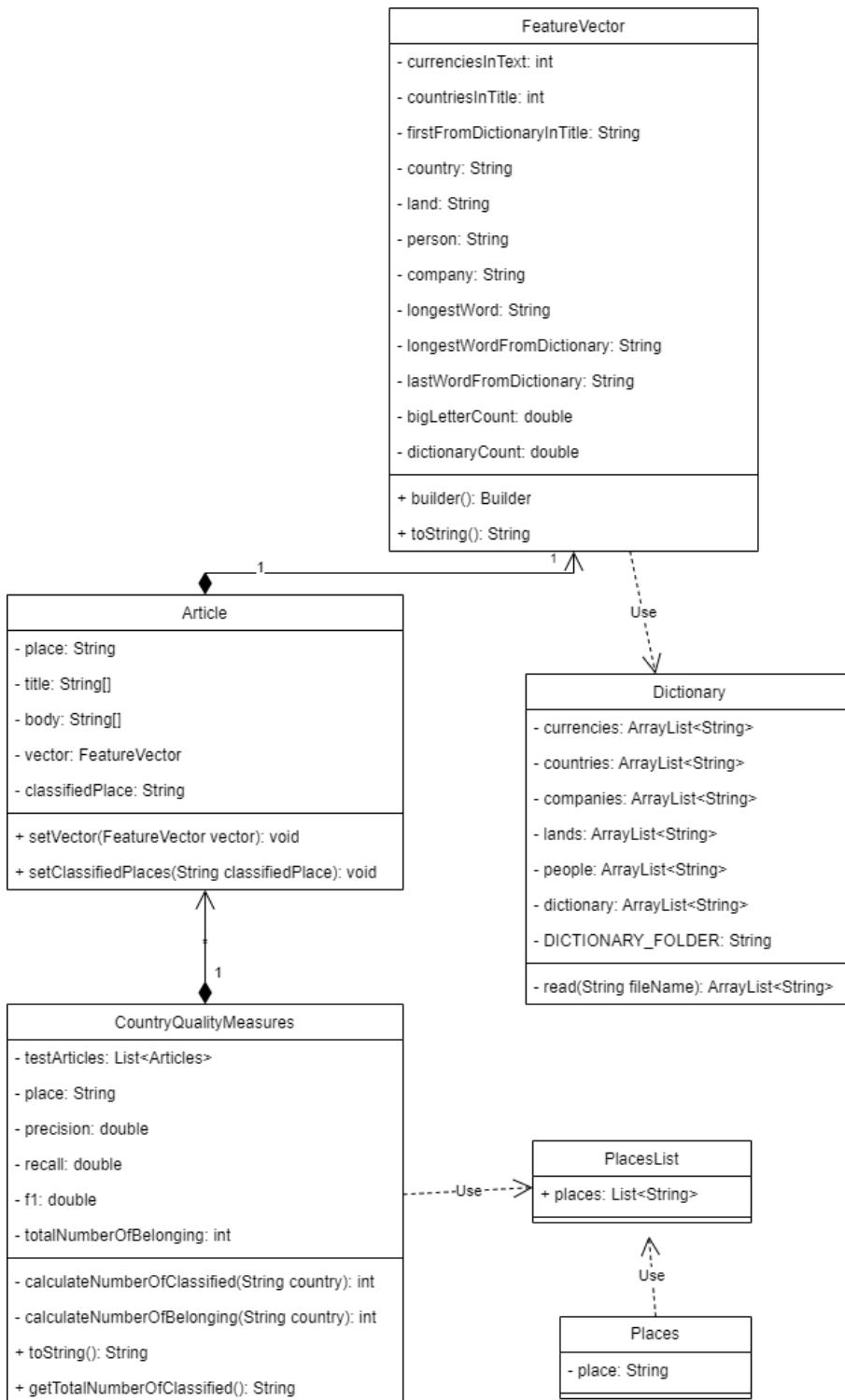
Rysunek 2. Diagram klas - pakiet *extractor*

Pakiet *extractor* odpowiada za wczytanie danych z bazy danych oraz eksplikację wektorów cech. Na diagramie (rys.2) przedstawione zostały klasy należące do tego pakietu:

- *DataReader* - instancja tej klasy pozwala odczytać dokumenty tekstowe z bazy danych, a następnie tworzy oddzielne artykuły (4.1.3).

- *Formatter* - instancja tej klasy przetwarza wczytany tekst i umożliwia utworzenie artykułu (4.1.3).
- *FeatureExtractor* - instancja tej klasy zawiera metody wykorzystywane przez budowniczego klasy (4.1.3) do ekstrakcji wektorów cech z artykułów.

#### 4.1.3. Diagram klas - pakiet danych

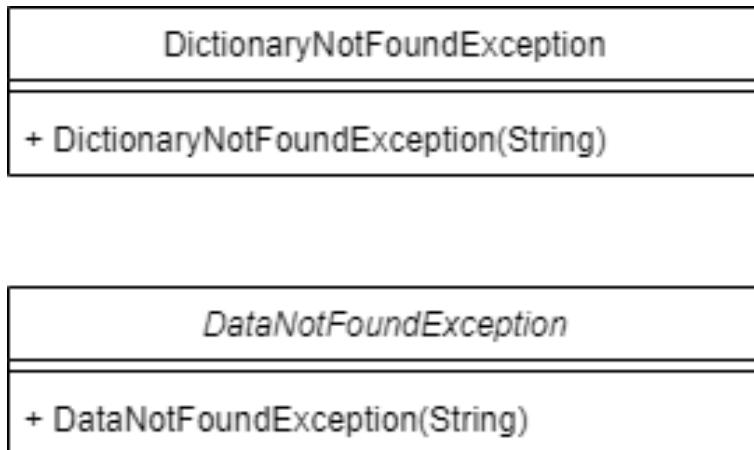


Rysunek 3. Diagram klas - pakiet *data*

Pakiet *data* odpowiada za przechowywanie danych wykorzystywanych w trakcie działania algorytmu. Na diagramie (rys.3) przedstawione zostały klasy należące do tego pakietu :

- *Places* - enum, zawierający nazwy państw w zadaniu.
- *PlacesList* - klasa pomocnicza, tworząca listę wszystkich państw w zadaniu.
- *CountryQualityMeasures* - klasa przechowująca obliczone wyniki miar jakości klasyfikacji.
- *Article* - obiekty tej klasy przechowują informacje o wczytanym artykule z bazy danych.
- *FeatureVector* - obiekty tej klasy określają wektory cech wczytanych artykułów.
- *Dictionary* - instancja tej klasy zawiera wczytane słowniki [15] wykorzystywane podczas ekstrakcji wektorów cech.

#### 4.1.4. Diagram klas - pakiet wyjątków

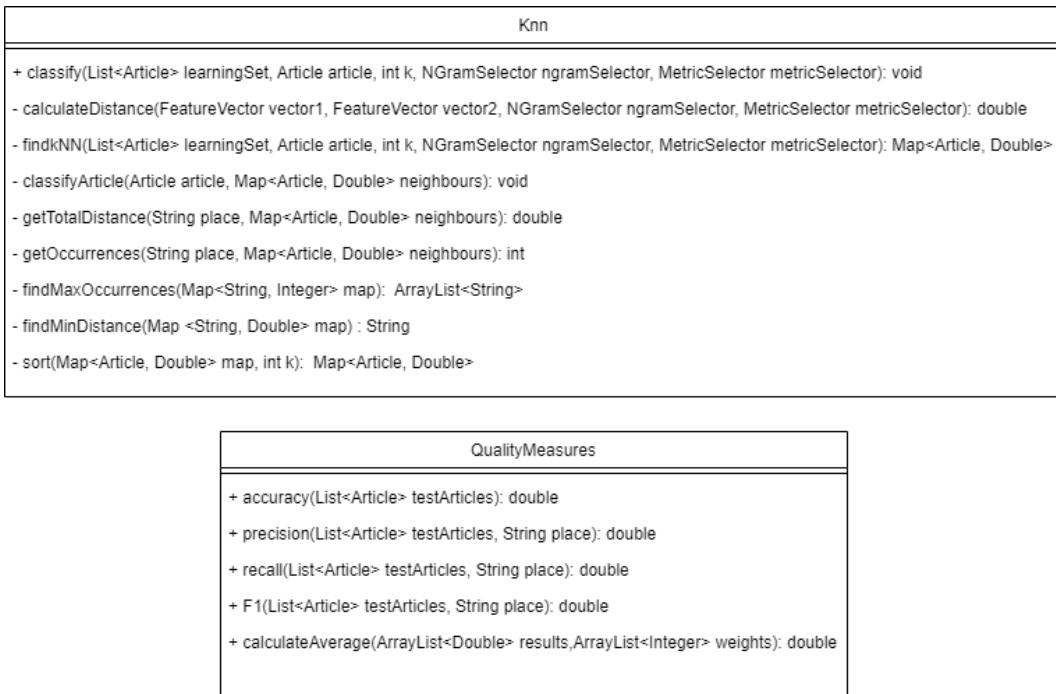


Rysunek 4. Diagram klas - pakiet *exception*

Pakiet wyjątków, zawiera opisane wyjątki, które mogą przerwać działanie programu w niekontrolowany sposób. Na diagramie (rys.4) przedstawione zostały klasy należące do tego pakietu:

- *DictionaryNotFoundException* - wyjątek, pojawiający się, gdy program nie znajdzie jednego z wymaganych słowników.
- *DataNotFoundException* - wyjątek, pojawiający się, gdy program nie znajdzie plików z danymi do wczytania.

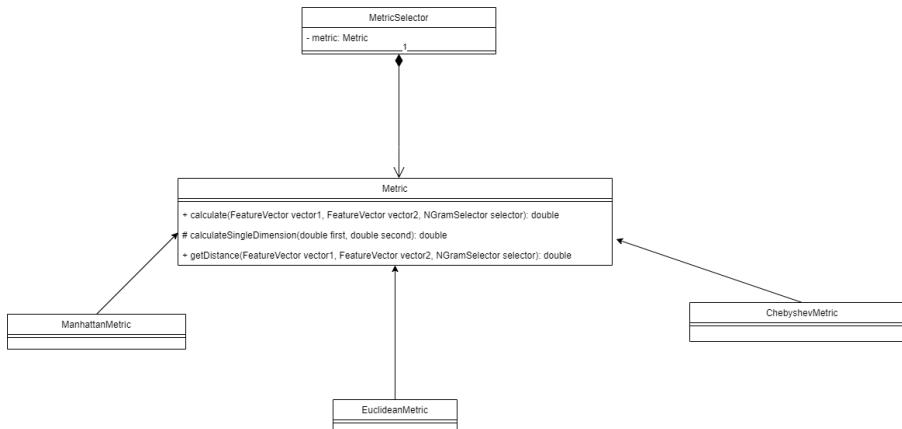
#### 4.1.5. Diagramy klas - pakiet klasyfikatora



Rysunek 5. Diagram klas - pakiet *algorithm*

Pakiet *algorithm* jest jednym z podpakietów pakietu *classifier*. Odpowiada za algorytm klasyfikacji, a także mierzy jego rezultat. Na diagramie (rys.5) przedstawione zostały klasy należące do tego pakietu :

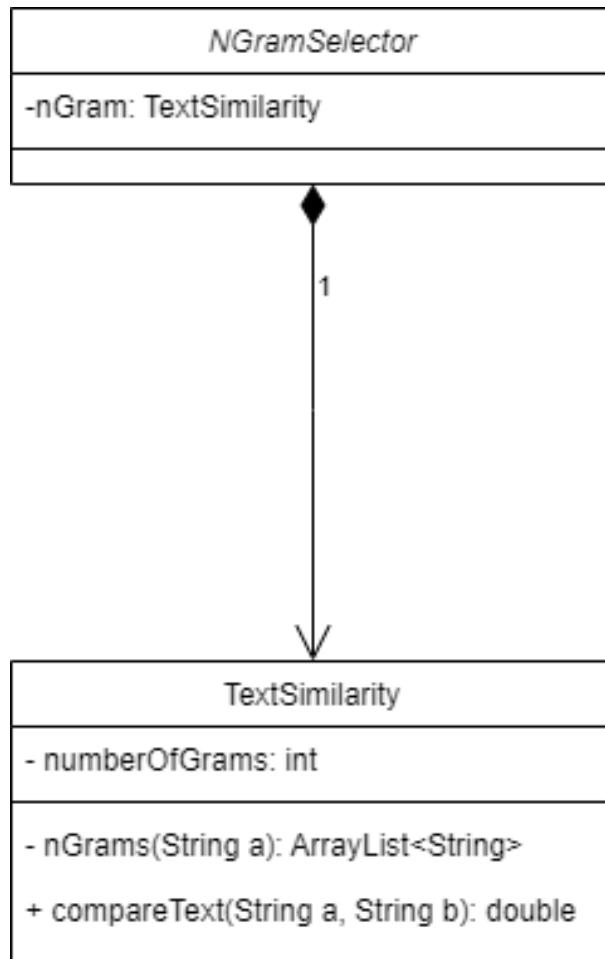
- *QualityMeasures* - instancja tej klasy zawiera implementację miar jakości klasyfikacji (2.3).
- *Knn* - instancja tej klasy zawiera wszystkie niezbędne metody do wykonania klasyfikacji z użyciem algorytmu *k* najbliższych sąsiadów (2).



Rysunek 6. Diagram klas - pakiet *metrics*

Pakiet *metric* jest jednym z podpakietów pakietu *classifier*. Zawiera wykorzystywane w zadaniu metryki (3.3). Na diagramie (rys.6) przedstawione zostały klasy należące do tego pakietu :

- *MetricSelector* - enum, pozwalający na wybór jednej z metryk.
- *Metric* - abstrakcyjna klasa, służąca jako rodzic wykorzystywanych metryk. Ostatecznie w klasyfikacji wykorzystywana jest taka metryka, jaką zostanie stworzone przez opisaną wyżej klasę *MetricSelector*.
- *EuclideanMetric* - klasa zawierająca implementację metryki euklidesowej (3.3.1).
- *ManhattanMetric* - klasa zawierająca implementację metryki miejskiej (3.3.2).
- *ChebyshevMetric* - klasa zawierająca implementację metryki Czebyszewa (3.3.3).

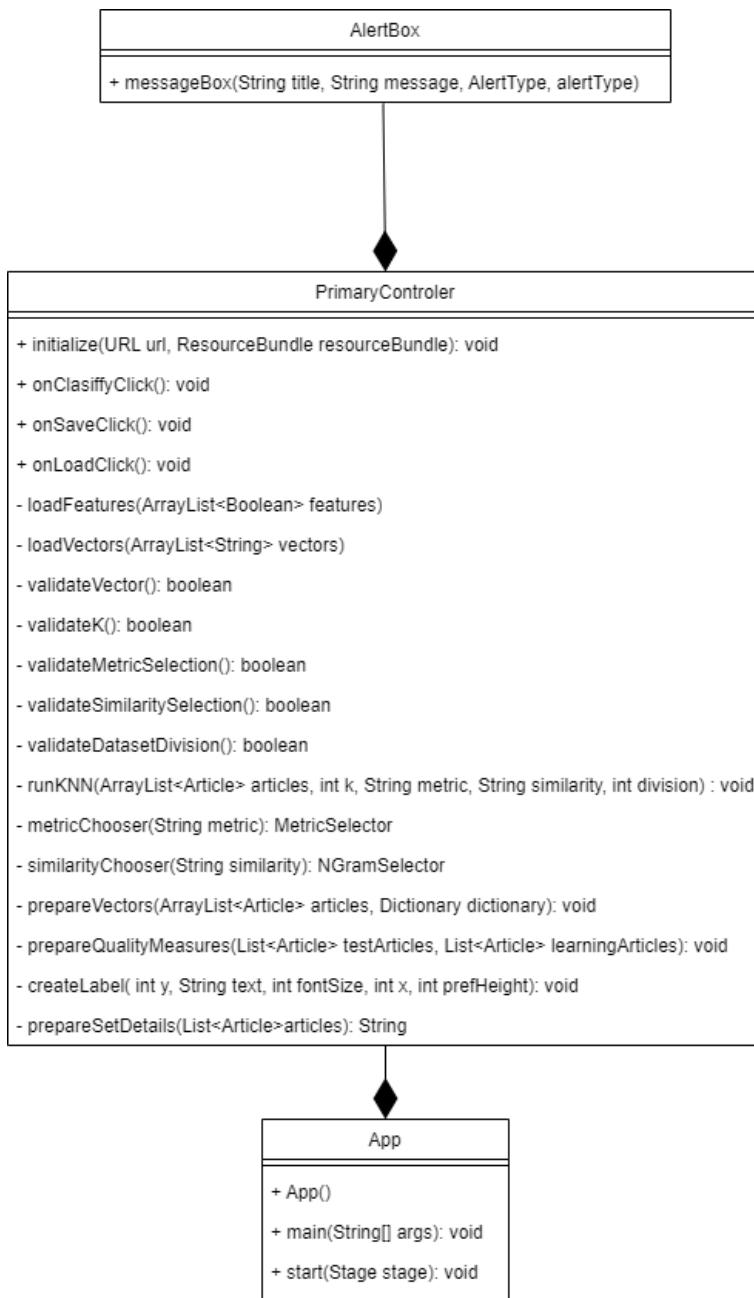


Rysunek 7. Diagram klas - pakiet *similarity*

Pakiet *similarity* jest jednym z podpakietów pakietu *classifier*. Zawiera wykorzystywane w zadaniu miary podobieństwa tekstów (3.1), a także sposób ich przeliczenia na metryki (3.2). Na diagramie (rys.7) przedstawione zostały klasy należące do tego pakietu :

- *NGramSelector* - klasa pozwalająca wybrać, którą odmianę metody n-gramów (3.1.1) użyjemy w zadaniu.
- *TextSimilarity* - klasa inicjalizowana przez opisany wyżej *NGramSelector*. Oblicza miarę podobieństwa tekstów i przekształca ją na odległość z użyciem funkcji 3.4.

#### 4.1.6. Diagram klas - pakiet interfejsu użytkownika



Rysunek 8. Diagram klas - pakiet *view*

Pakiet *view* jest jedynym pakietem modułu *View*. Zawiera kontroler do obsługi graficznego interfejsu użytkownika. Na diagramie (rys.8) przedsta-

wione zostały klasy należące do tego pakietu :

- *App* - klasa ładująca aplikację.
- *PrimaryController* - klasa odpowiedzialna za interakcję z użytkownikiem. Interfejs użytkownika szczegółowo opisano w rozdziale (4.2.2).

## 4.2. Prezentacja wyników, interfejs użytkownika

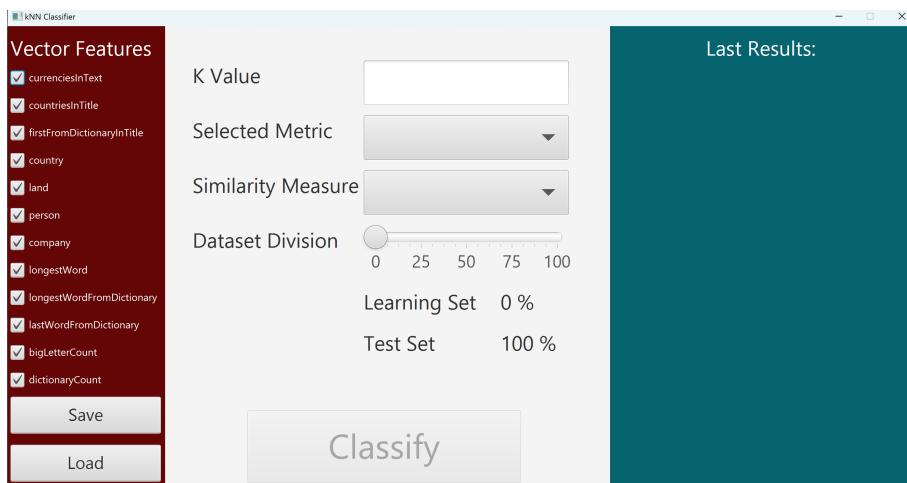
### 4.2.1. Wymagania aplikacji

Aby uruchomić aplikacje potrzebne są:

- JRE (JavaRuntimeEnvironment) - środowisko uruchomieniowe języka Java w wersji co najmniej 17 o architekturze *amd64* [25].
- Apache Maven - narzędzie automatyzujące budowę oprogramowania w wersji co najmniej 3.8.6 [26].

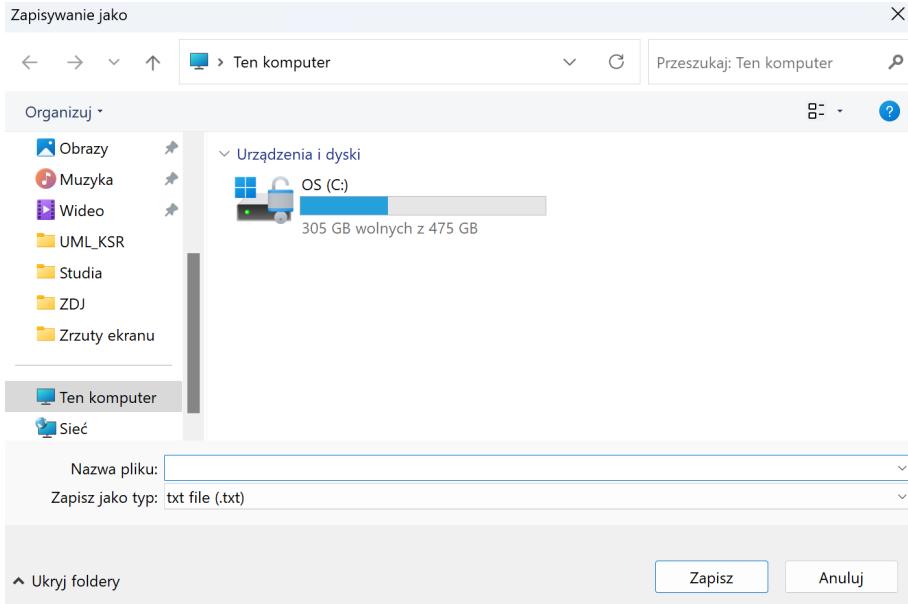
### 4.2.2. Interfejs użytkownika

Aplikacja posiada graficzny interfejs użytkownika. Okno aplikacji przedstawiono poniżej (rys.9).



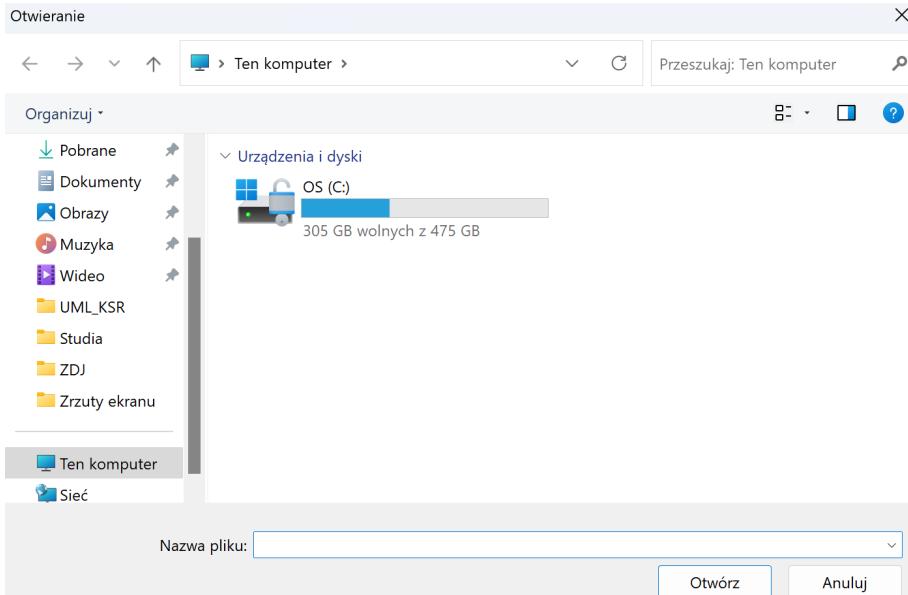
Rysunek 9. Interfejs użytkownika - widok startowy

Panel został podzielony na 3 części. Po lewej stronie znajduje się lista cech wektorów. Użytkownik może zaznaczyć dowolne cechy, które algorytm ma uwzględnić. Pod listą cech znajdują się dwa przyciski. Przycisk *Save* służy do zapisania bieżącej konfiguracji i wygenerowania wektorów z odpowiednimi cechami. Po naciśnięciu tego przycisku pojawi się okno z wyborem miejsca zapisu pliku (rys.10).



Rysunek 10. Interfejs użytkownika - okno zapisu do pliku

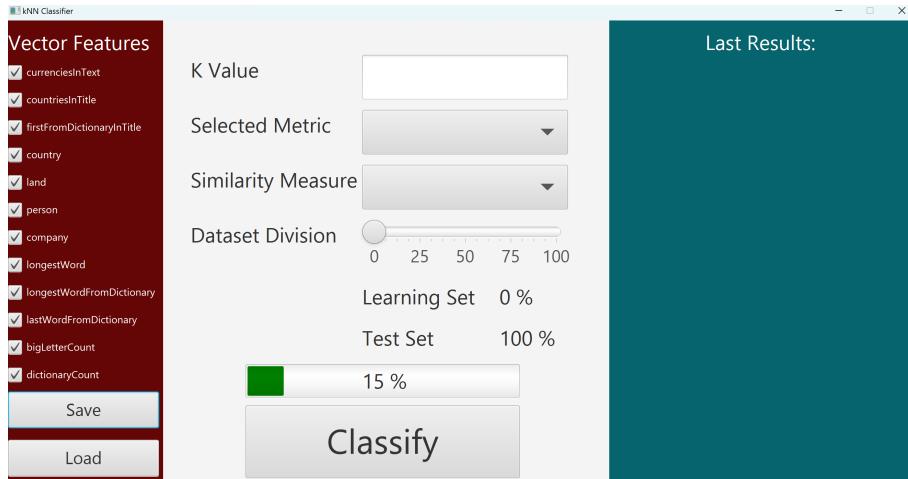
Z kolei przycisk *Load* pozwoli na wczytanie konfiguracji z istniejącego już pliku. Po naciśnięciu tego przycisku pojawi się okno z wyborem pliku do wczytania (rys.11).



Rysunek 11. Interfejs użytkownika - okno wczytania pliku wektorów

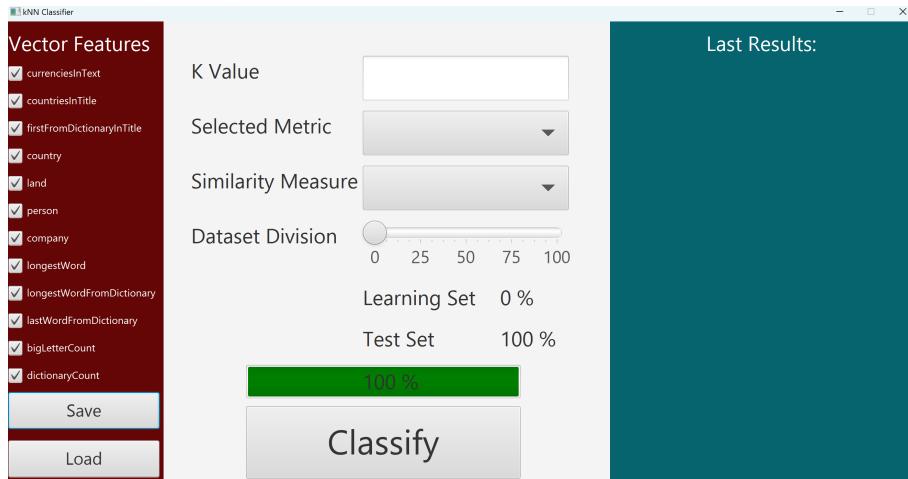
Ułatwi to kontynuowanie użytkownikowi dalszych eksperymentów, w przypadku, gdy aplikacja zostanie zamknięta.

Aby móc rozpoczęć klasyfikację, po pierwszym uruchomieniu aplikacji, należy wcześniej nacisnąć przycisk *Save* lub *Load*, aby przygotować wektory cech. Wówczas pojawi się pasek postępu. (rys. 12).



Rysunek 12. Interfejs użytkownika - w trakcie przygotowania wektorów

Gdy program zakończy przygotowanie wektorów, odblokuje to przycisk *Classify* (rys.13).



Rysunek 13. Interfejs użytkownika - widok po wczytaniu wektorów

Środkowa część panelu użytkownika pozwala na dowolne parametryzowanie algorytmu  $k$ -NN. Pierwszym parametrem jest liczba najbliższych sąsiadów, jaką algorytm będzie uwzględniał. Użytkownik podaje wartość tego parametru w polu tekstowym obok etykiety *K Value*. Jeśli użytkownik nie poda wartości lub poda wartość niecałkowitą i naciśnie przycisk *Classify*, wyświetli się okno informujące o błędny parametrze (rys.14). Informacja o błędzie wyświetli się również, jeśli podana wartość  $k$  będzie większa niż liczba elementów w zbiorze uczącym.

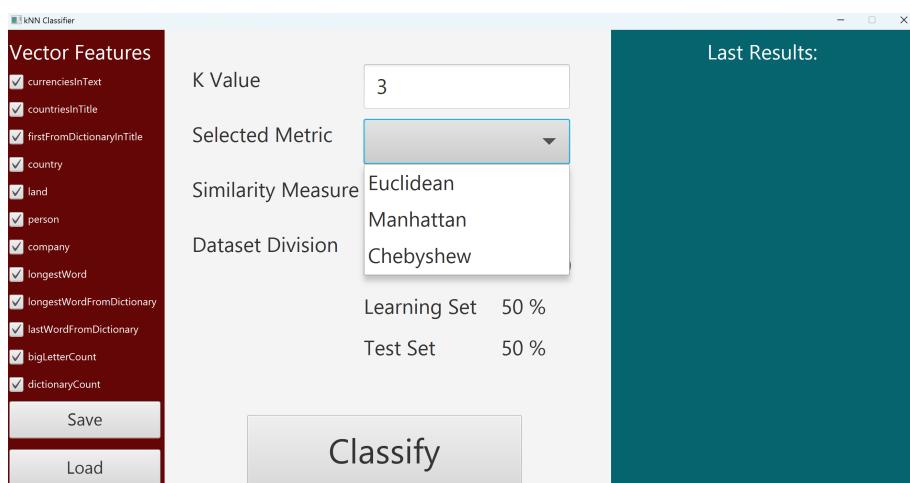


Rysunek 14. Interfejs użytkownika - informacja o błędnych parametrach k

Kolejnym parametrem algorytmu jest wybrana przez użytkownika metryka. Przy etykiecie *Selected Metric* znajduje się z pole z listą rozwijaną, z której użytkownik może wybrać jedną z 3 predefiniowanych metryk:

- *Euclidean* - metryka euklidesowa (3.3.1).
- *Manhattan* - metryka miejska (3.3.2).
- *Chebyshev* - metryka Czebyszewa (3.3.3).

Rys.15 przedstawia wybór metryki.



Rysunek 15. Interfejs użytkownika - metryki do wyboru

Jeśli użytkownik nie wybierze metryki i naciśnie przycisk *Classify*, wyświetli się okno informujące o konieczności wyboru metryki (rys.16). Informacja ta wyświetli się, o ile użytkownik poda prawidłowy parametr  $k$ . W przeciwnym razie, pojawi się informacja o złym parametrze  $k$  (rys.14).

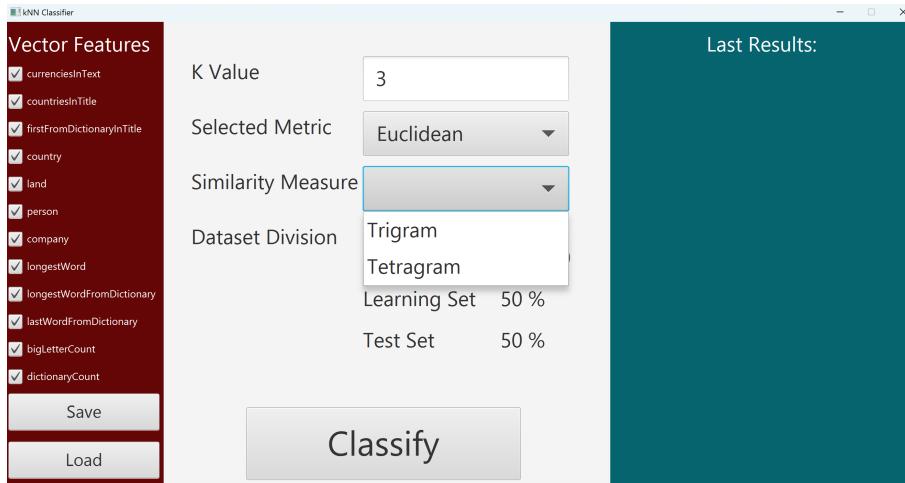


Rysunek 16. Interfejs użytkownika - informacja o braku wyboru metryki

Następnym parametrem algorytmu jest wybrana miara podobieństwa tekstów. Przy etykiecie *Similarity Measure* znajduje się z pole z listą rozwijaną, z której użytkownik może wybrać jedną z 2 predefiniowanych miar podobieństwa tekstów :

- *Trigram* - metoda *trigramów* (3.1).
- *Tetragram* - metoda *tetragramów* (3.2).

Rys. 17 przedstawia wybór miary podobieństwa.



Rysunek 17. Interfejs użytkownika - miary podobieństwa do wyboru

Jeśli użytkownik nie wybierze miary podobieństwa i naciśnie przycisk *Classify*, wyświetli się okno informujące o konieczności wyboru miary podobieństwa (rys.18). Informacja ta wyświetli się, o ile użytkownik poda prawidłowe poprzednie parametry. W przeciwnym razie, pojawi się informacja o tym parametrze, który został wybrany błędnie jako pierwszy.



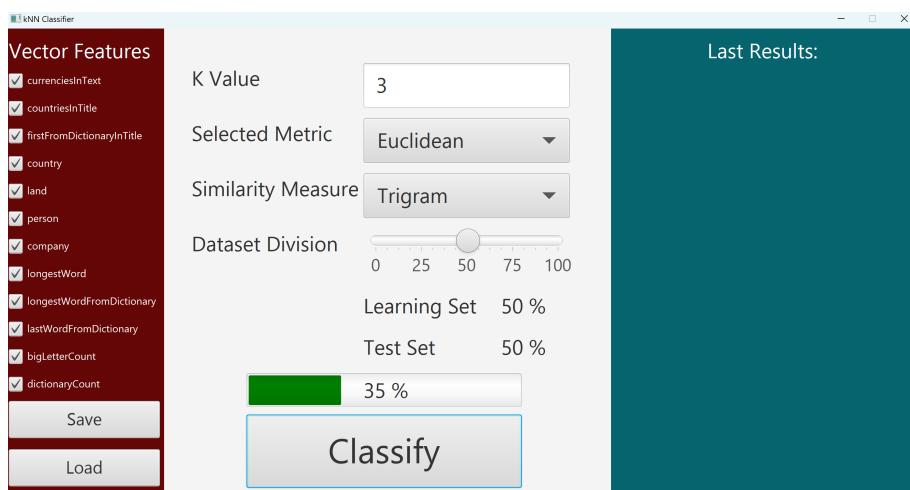
Rysunek 18. Interfejs użytkownika - informacja o braku wyboru podobieństwa

Ostatnim parametrem jest podział zbioru dokumentów. W tym celu użytkownik ustawia suwak znajdujący się na prawo od etykiety *Dataset Division*. W przypadku, gdy użytkownik zbiór testowy lub zbiór uczący będą puste, zostanie wyświetlona informacja o błędny podziale zbioru (rys. 19), o ile poprzednie parametry zostały podane poprawnie.



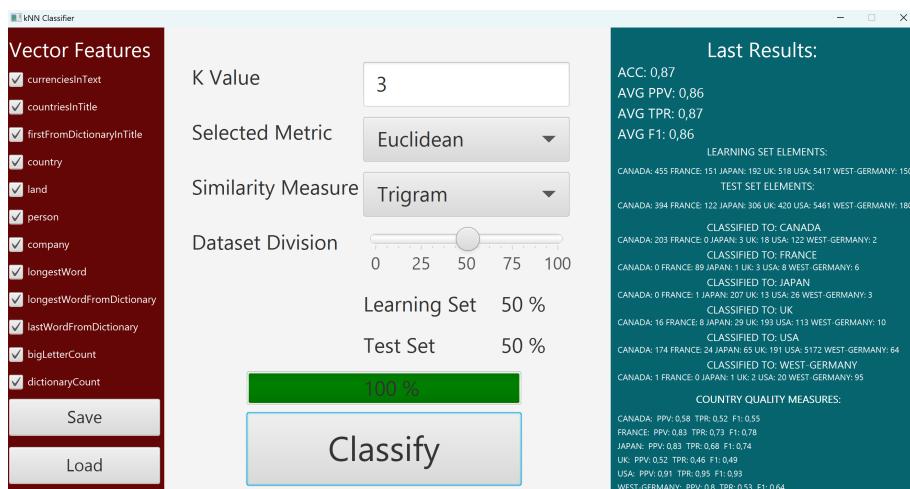
Rysunek 19. Interfejs użytkownika - informacja o błędny podziale zbioru

Po wyborze wszystkich parametrów, należy nacisnąć przycisk *Classify*, w celu wykonania klasyfikacji dla wcześniej dostarczonych parametrów. Wówczas pojawi się pasek progresu (rys.20).



Rysunek 20. Interfejs użytkownika - w trakcie klasyfikacji

Trzecia część panelu użytkownika służy do wypisywania wyników klasyfikacji. Wyniki pojawią się dopiero po zakończeniu klasyfikacji (rys.21).



Rysunek 21. Interfejs użytkownika - wyświetlanie wyników

## 5. Wyniki klasyfikacji dla różnych parametrów wejściowych

Wszystkie eksperymenty przeprowadzane były na ograniczonym zbiorze składającym się z 600 pierwszych tekstów ze zbioru tekstów Reuters [13].

Tabela 1. Liczebność danej klasy w ograniczonym zbiorze wykorzystanym w eksperymentach

Kraj	Liczebność artykułów w zbiorze
Canada	34
France	12
Japan	23
UK	50
USA	473
West Germany	8

Aby przeprowadzić analizę w zadaniu rozpatrujemy 12 cech, więc wektor cech będzie miał postać 2.1. Zostało wyszczególnione pięć podzbiorów cech, które zostały wykorzystane do zbadania, które cechy potencjalnie mają najmniejszy wpływ na wyniki klasyfikacji.

$$I = \{C_3, C_4, \dots, C_{10}\} \quad (5.1)$$

$$J = \{C_1, C_2, \dots, C_{12}\} \quad (5.2)$$

$$K = \{C_1, C_2, C_{11}, C_{12}\} \quad (5.3)$$

$$L = \{C_1, C_4, C_5, C_{11}\} \quad (5.4)$$

$$M = \{C_2, C_3, C_6, C_7, C_8\} \quad (5.5)$$

Podzbiory  $I, J, K, L, M$  składają się z cech wybranych z pośród 12 omawianych w sprawozdaniu cech.

Podzbiór  $I$  (5.1) składa się z cech od trzeciej (2.1.3) do dziesiątej (2.1.10), są to cechy tekstowe,

Podzbiór  $J$  (5.2) składa się ze wszystkich dostępnych cech.

Podzbiór  $K$  (5.3) składa się z cech: pierwsza (2.1.1), druga (2.1.2), jedenasta (2.1.11), dwunasta (2.1.12), są to cechy liczbowe.

Podzbiór  $L$  (5.4) składa się z dwóch cech liczbowych: pierwsza (2.1.1) i jedenasta (2.1.11), oraz z dwóch cech tekstowych: czwarta (2.1.4) i piąta (2.1.5).

Podzbiór  $M$  zawiera jedną cechę liczbową (2.1.2) oraz cztery cechy tekstowe (2.1.3, 2.1.6, 2.1.7, 2.1.8)

### 5.1. Wyniki dla różnych parametrów $k$

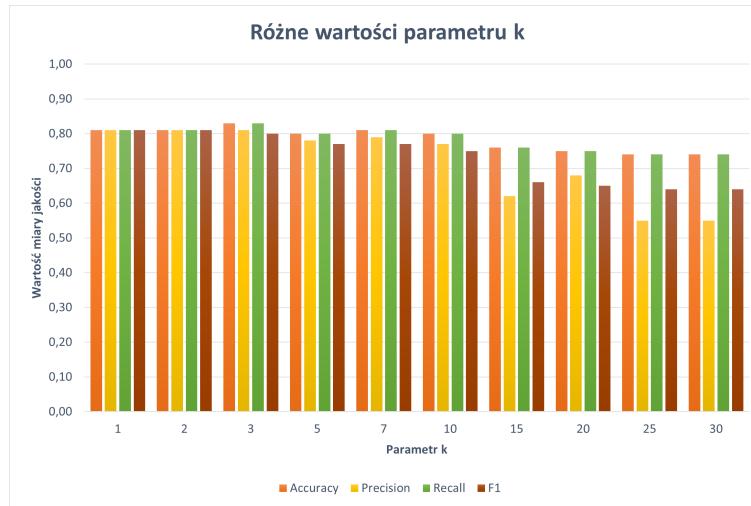
. Podczas przeprowadzania eksperymentów użyliśmy 10 różnych wartości parametru  $k$ : 1, 2, 3, 5, 7, 10, 15, 20, 25, 30.

Tabela 2. Różne wartości parametru  $k$  – eksperyment 1–5, ograniczony zbiór artykułów

Numer eksperymentu	1	2	3	4	5
Parametr $k$	1	2	3	5	7
Metryka	<i>euklidesowa</i>				
Miara podobieństwa	<i>trigramów</i>				
Podział zbioru (uczący/testowy)	70%/30%				
Zbiór cech	$J$				
Accuracy	0.81	0.81	0.83	0.80	0.81
Precision	0.81	0.81	0.81	0.78	0.79
Recall	0.81	0.81	0.83	0.80	0.81
F1	0.81	0.81	0.80	0.77	0.77
Precision Canada	0.33	0.33	0.33	0.14	0.20
Precision France	1.00	1.00	1.00	1.00	1.00
Precision Japan	0.75	0.75	0.75	0.75	0.67
Precision UK	0.61	0.61	0.89	0.86	1.00
Precision USA	0.86	0.86	0.84	0.82	0.81
Precision West Germany	1.00	1.00	0.00	NaN	NaN
Recall Canada	0.38	0.38	0.25	0.12	0.12
Recall France	1.00	1.00	1.00	1.00	0.83
Recall Japan	0.60	0.60	0.60	0.60	0.40
Recall UK	0.50	0.50	0.36	0.27	0.32
Recall USA	0.90	0.90	0.97	0.96	0.97
Recall West Germany	0.40	0.40	0.00	0.00	0.00
F1 Canada	0.35	0.35	0.29	0.13	0.15
F1 France	1.00	1.00	1.00	1.00	0.91
F1 Japan	0.67	0.67	0.67	0.67	0.50
F1 UK	0.55	0.55	0.52	0.41	0.48
F1 USA	0.88	0.88	0.90	0.88	0.88
F1 West Germany	0.57	0.57	NaN	NaN	NaN

Tabela 3. Różne wartości parametru  $k$  – eksperyment 6–10, ograniczony zbiór artykułów

Numer eksperymentu	6	7	8	9	10
Parametr $k$	10	15	20	25	30
Metryka	<i>euklidesowa</i>				
Miara podobieństwa	<i>trigramów</i>				
Podział zbioru (uczący/testowy)	70%/30%				
Zbiór cech	<i>J</i>				
Accuracy	0.80	0.76	0.75	0.74	0.74
Precision	0.77	0.62	0.68	0.55	0.55
Recall	0.80	0.76	0.75	0.74	0.74
F1	0.75	0.66	0.65	0.64	0.64
Precision Canada	0.00	NaN	NaN	NaN	NaN
Precision France	1.00	1.00	NaN	NaN	NaN
Precision Japan	0.67	1.00	NaN	NaN	NaN
Precision UK	1.00	NaN	1.0	NaN	NaN
Precision USA	0.80	0.75	0.75	0.74	0.74
Precision West Germany	NaN	NaN	NaN	NaN	NaN
Recall Canada	0.00	0.00	0.00	0.00	0.00
Recall France	0.83	0.17	0.00	0.00	0.00
Recall Japan	0.40	0.20	0.00	0.00	0.00
Recall UK	0.27	0.00	0.05	0.00	0.00
Recall USA	0.98	1.00	1.00	1.00	1.00
Recall West Germany	0.00	0.00	0.00	0.00	0.00
F1 Canada	NaN	NaN	NaN	NaN	NaN
F1 France	0.91	0.29	NaN	NaN	NaN
F1 Japan	0.50	0.33	NaN	NaN	NaN
F1 UK	0.43	NaN	0.09	NaN	NaN
F1 USA	0.88	0.86	0.86	0.85	0.85
F1 West Germany	NaN	NaN	NaN	NaN	NaN



Rysunek 22. Wykres przedstawiający wyniki eksperymentów 1-10 dla różnych wartości parametru  $k$ , ograniczony zbiór artykułów

Tabela 4. Różne wartości parametru  $k$ , eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	3	0	0	0	6	0
France	0	6	0	0	0	0
Japan	1	0	3	0	0	0
UK	0	0	0	11	7	0
USA	4	0	2	11	121	3
West Germany	0	0	0	0	0	2

Tabela 5. Różne wartości parametru  $k$ , eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	3	0	0	0	6	0
France	0	6	0	0	0	0
Japan	1	0	3	0	0	0
UK	0	0	0	11	7	0
USA	4	0	2	11	121	3
West Germany	0	0	0	0	0	2

Tabela 6. Różne wartości parametru  $k$ , eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	1	3	0
France	0	6	0	0	0	0
Japan	0	0	3	1	0	0
UK	0	0	0	8	1	0
USA	6	0	1	12	130	5
West Germany	0	0	1	0	0	0

Tabela 7. Różne wartości parametru  $k$ , eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	1	0	0	1	5	0
France	0	6	0	0	0	0
Japan	0	0	3	1	0	0
UK	0	0	0	6	1	0
USA	7	0	2	14	128	5
West Germany	0	0	0	0	0	0

Tabela 8. Różne wartości parametru  $k$ , eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	3	0
France	0	5	0	0	0	0
Japan	0	0	2	1	0	0
UK	0	0	0	6	0	0
USA	8	1	3	15	131	5
West Germany	0	0	0	0	0	0

Tabela 9. Różne wartości parametru  $k$ , eksperyment 6, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	1	0	0	0	4	0
France	0	5	0	0	0	0
Japan	0	0	2	1	0	0
UK	0	0	0	7	0	0
USA	7	1	3	14	130	5
West Germany	0	0	0	0	0	0

Tabela 10. Różne wartości parametru  $k$ , eksperyment 7, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	0	0
France	0	1	0	0	0	0
Japan	0	0	1	0	0	0
UK	0	0	0	0	0	0
USA	8	5	4	22	134	5
West Germany	0	0	0	0	0	0

Tabela 11. Różne wartości parametru  $k$ , eksperyment 8, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	0	0
France	0	0	0	0	0	0
Japan	0	0	0	0	0	0
UK	0	0	0	1	0	0
USA	8	6	5	21	134	5
West Germany	0	0	0	0	0	0

Tabela 12. Różne wartości parametru  $k$ , eksperyment 9, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	0	0
France	0	0	0	0	0	0
Japan	0	0	0	0	0	0
UK	0	0	0	0	0	0
USA	8	6	5	22	134	5
West Germany	0	0	0	0	0	0

Tabela 13. Różne wartości parametru  $k$ , eksperyment 10, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	0	0
France	0	0	0	0	0	0
Japan	0	0	0	0	0	0
UK	0	0	0	0	0	0
USA	8	6	5	22	134	5
West Germany	0	0	0	0	0	0

Wartość parametru  $k$  wpływa na jakość klasyfikacji. Przy stałych pozostałych parametrach i zmianie jedynie parametru  $k$  otrzymano różne wyniki wszystkich miar jakości

klasyfikacji. Dla zbioru 600 pierwszych artykułów najlepsze wyniki otrzyma-  
no dla  $k = 3$  (tabela 2 i 3). Kiedy parametr  $k$  podniesiony zostało powyżej  
 $k = 10$  widać znaczny spadek jakości klasyfikacji. Wpływ na takie zacho-  
wanie ma rozważany zbiór (tabela 15). W zbiorze zdecydowanie najwięcej  
jest artykułów z kraju USA, co sprawia, że dla odpowiednio dużej wartości  
 $k$ , wszystkie artykuły zostaną sklasyfikowane do tej klasy, z uwagi na jej  
dominującą pozycję pośród  $k$  najbliższych sąsiadów (tabela 12 i 13).

## 5.2. Wybór podziału na zbiór uczący i testowy

Podczas przeprowadzania eksperymentów używaliśmy podziału zbioru  
w stosunku:

- 80% – zbiór uczący, 20% – zbiór testowy
- 70% – zbiór uczący, 30% – zbiór testowy
- 60% – zbiór uczący, 40% – zbiór testowy
- 50% – zbiór uczący, 50% – zbiór testowy
- 40% – zbiór uczący, 60% – zbiór testowy
- 30% – zbiór uczący, 70% – zbiór testowy

Tabela 14. Liczebność artykułów w zbiorze. Podział 80%/20%, ograniczony zbiór  
artykułów

<b>Podział</b>	<b>80%/20%</b>	
	<b>Kraj</b>	<b>Liczebność artykułów w zbiorze uczącym</b>
<b>Canada</b>	28	6
<b>France</b>	6	6
<b>Japan</b>	18	5
<b>UK</b>	28	22
<b>USA</b>	397	76
<b>West Germany</b>	3	5

Tabela 15. Liczebność artykułów w zbiorze. Podział 70%/30%, ograniczony zbiór  
artykułów

<b>Podział</b>	<b>70%/30%</b>	
	<b>Kraj</b>	<b>Liczebność artykułów w zbiorze uczącym</b>
<b>Canada</b>	26	8
<b>France</b>	6	6
<b>Japan</b>	18	5
<b>UK</b>	28	22
<b>USA</b>	339	134
<b>West Germany</b>	3	5

Tabela 16. Liczebność artykułów w zbiorze. Podział 60%/40%, ograniczony zbiór artykułów

<b>Podział</b>	<b>60%/40%</b>	
<b>Kraj</b>	<b>Liczebność artykułów w zbiorze uczącym</b>	<b>Liczebność artykułów w zbiorze testowym</b>
<b>Canada</b>	24	10
<b>France</b>	6	6
<b>Japan</b>	18	5
<b>UK</b>	27	23
<b>USA</b>	283	190
<b>West Germany</b>	2	6

Tabela 17. Liczebność artykułów w zbiorze. Podział 50%/50%, ograniczony zbiór artykułów

<b>Podział</b>	<b>50%/50%</b>	
<b>Kraj</b>	<b>Liczebność artykułów w zbiorze uczącym</b>	<b>Liczebność artykułów w zbiorze testowym</b>
<b>Canada</b>	23	11
<b>France</b>	6	6
<b>Japan</b>	17	6
<b>UK</b>	23	27
<b>USA</b>	230	243
<b>West Germany</b>	1	7

Tabela 18. Liczebność artykułów w zbiorze. Podział 40%/60%, ograniczony zbiór artykułów

<b>Podział</b>	<b>40%/60%</b>	
<b>Kraj</b>	<b>Liczebność artykułów w zbiorze uczącym</b>	<b>Liczebność artykułów w zbiorze testowym</b>
<b>Canada</b>	18	16
<b>France</b>	5	7
<b>Japan</b>	17	6
<b>UK</b>	17	33
<b>USA</b>	182	291
<b>West Germany</b>	1	7

Tabela 19. Liczebność artykułów w zbiorze. Podział 30%/70%, ograniczony zbiór artykułów

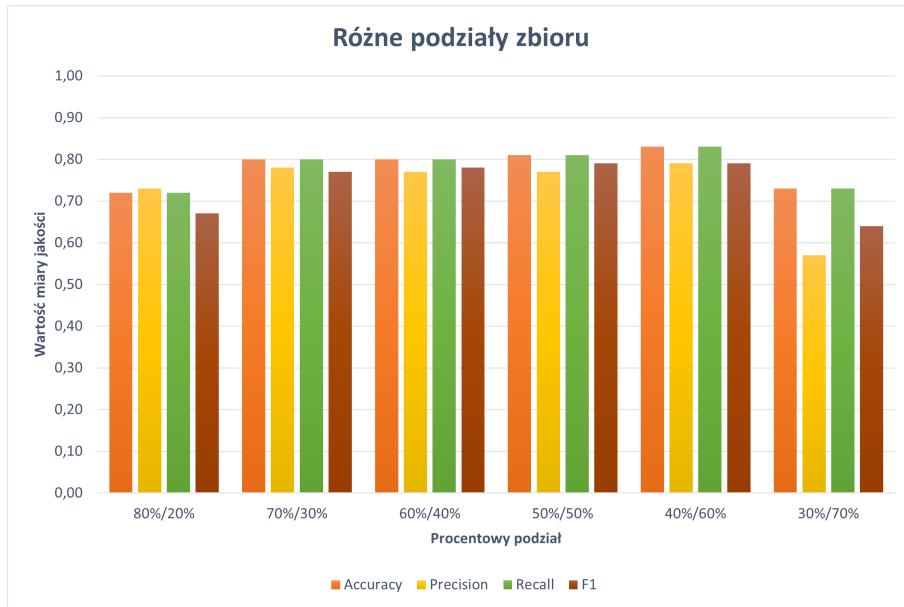
<b>Podział</b>	30%/70%	
	<b>Kraj</b>	<b>Liczebność artykułów w zbiorze uczącym</b>
<b>Canada</b>	18	16
<b>France</b>	1	11
<b>Japan</b>	0	23
<b>UK</b>	2	48
<b>USA</b>	159	314
<b>West Germany</b>	0	8

Tabela 20. Różne podziały zbioru – eksperyment 1–3, ograniczony zbiór artykułów

<b>Numer eksperymentu</b>	<b>1</b>	<b>2</b>	<b>3</b>
<b>Parametr <math>k</math></b>		5	
<b>Metryka</b>		<i>euklidesowa</i>	
<b>Miara podobieństwa</b>		<i>trigramów</i>	
<b>Zbiór cech</b>		$J$	
<b>Podział zbioru (uczący/testowy)</b>	80%/20%	70%/30%	60%/40%
<b>Accuracy</b>	0.72	0.80	0.80
<b>Precision</b>	0.73	0.78	0.77
<b>Recall</b>	0.72	0.80	0.80
<b>F1</b>	0.67	0.77	0.78
<b>Precision Canada</b>	0.14	0.14	0.18
<b>Precision France</b>	1.00	1.00	1.00
<b>Precision Japan</b>	0.67	0.75	0.60
<b>Precision UK</b>	1.00	0.86	0.50
<b>Precision USA</b>	0.72	0.82	0.85
<b>Precision West Germany</b>	NaN	NaN	NaN
<b>Recall Canada</b>	0.17	0.12	0.20
<b>Recall France</b>	1.00	1.00	1.00
<b>Recall Japan</b>	0.40	0.60	0.60
<b>Recall UK</b>	0.27	0.27	0.26
<b>Recall USA</b>	0.93	0.96	0.93
<b>Recall West Germany</b>	0.00	0.00	0.00
<b>F1 Canada</b>	0.15	0.13	0.19
<b>F1 France</b>	1.00	1.00	1.00
<b>F1 Japan</b>	0.50	0.67	0.60
<b>F1 UK</b>	0.43	0.41	0.34
<b>F1 USA</b>	0.82	0.88	0.89
<b>F1 West Germany</b>	NaN	NaN	NaN

Tabela 21. Różne podziały zbioru – eksperyment 4–6, ograniczony zbiór artykułów

Numer eksperymentu	4	5	6
Parametr $k$	5		
Metryka	<i>euklidesowa</i>		
Miara podobieństwa	<i>trigramów</i>		
Zbiór cech	$J$		
Podział zbioru (uczący/testowy)	50%/50%	40%/60%	30%/70%
Accuracy	0.81	0.83	0.73
Precision	0.77	0.79	0.57
Recall	0.81	0.83	0.73
F1	0.79	0.79	0.64
Precision Canada	0.20	0.12	0.18
Precision France	1.00	1.00	NaN
Precision Japan	0.67	0.40	NaN
Precision UK	0.44	0.80	NaN
Precision USA	0.86	0.85	0.75
Precision West Germany	NaN	NaN	NaN
Recall Canada	0.18	0.06	0.12
Recall France	1.00	1.00	0.00
Recall Japan	0.33	0.33	0.00
Recall UK	0.26	0.24	0.00
Recall USA	0.93	0.96	0.97
Recall West Germany	0.00	0.00	0.00
F1 Canada	0.19	0.08	0.15
F1 France	1.00	1.00	NaN
F1 Japan	0.44	0.36	NaN
F1 UK	0.33	0.37	NaN
F1 USA	0.89	0.90	0.85
F1 West Germany	NaN	NaN	NaN



Rysunek 23. Wykres przedstawiający wyniki eksperymentów 1–6 dla różnych podziałów zbioru, ograniczony zbiór artykułów

Tabela 22. Różne podziały zbioru, eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	1	0	0	1	5	0
France	0	6	0	0	0	0
Japan	0	0	2	1	0	0
UK	0	0	0	6	0	0
USA	5	0	3	14	71	5
West Germany	0	0	0	0	0	0

Tabela 23. Różne podziały zbioru, eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	1	0	0	1	5	0
France	0	6	0	0	0	0
Japan	0	0	3	1	0	0
UK	0	0	0	6	1	0
USA	7	0	2	14	128	5
West Germany	0	0	0	0	0	0

Tabela 24. Różne podziały zbioru, eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	1	8	0
France	0	6	0	0	0	0
Japan	0	0	3	1	1	0
UK	1	0	0	6	5	0
USA	7	0	2	15	176	6
West Germany	0	0	0	0	0	0

Tabela 25. Różne podziały zbioru, eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	0	8	0
France	0	6	0	0	0	0
Japan	0	0	2	1	0	0
UK	1	0	0	7	8	0
USA	8	0	4	19	227	7
West Germany	0	0	0	0	0	0

Tabela 26. Różne podziały zbioru, eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	1	0	0	0	7	0
France	0	7	0	0	0	0
Japan	0	0	2	1	2	0
UK	0	0	0	8	2	0
USA	15	0	4	24	280	7
West Germany	0	0	0	0	0	0

Tabela 27. Różne podziały zbioru, eksperyment 6, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	1	0	8	0
France	0	0	0	0	0	0
Japan	0	0	0	0	0	0
UK	0	0	0	0	0	0
USA	14	11	22	48	306	8
West Germany	0	0	0	0	0	0

Stopień podziału zbioru ma wpływ na wszystkie miary jakości klasyfikacji. Najlepsze wyniki uzyskano dla podziału 40% – zbiór uczący, 60% – zbiór testowy, z kolei najgorsze dla podziałów: 80% – zbiór uczący, 20% – zbiór

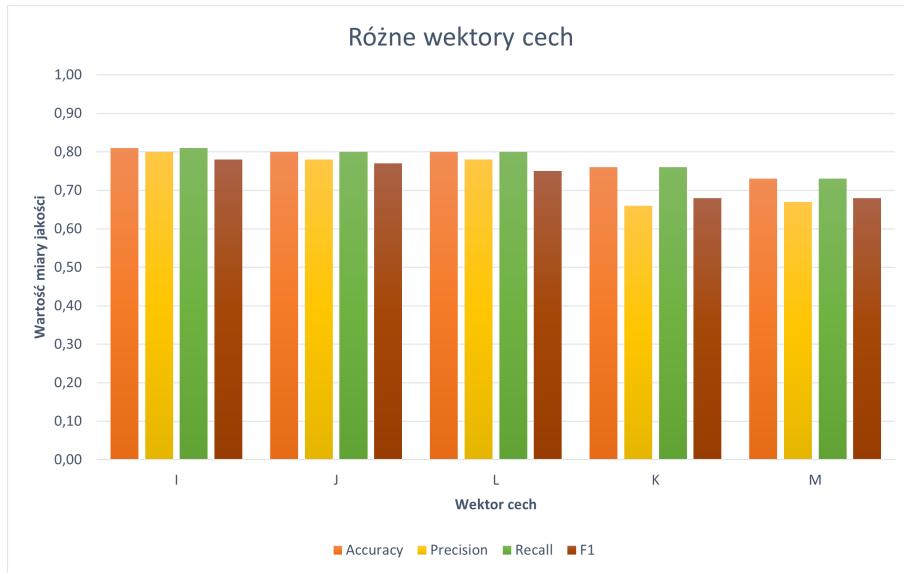
testowy i 30% – zbiór uczący, 70% – zbiór testowy. W przypadku, gdy jest mało elementów w zbiorze testowym, każdy błąd ma większy wpływ na wyniki miar jakości klasyfikacji. Z drugiej strony, zbyt mała liczba elementów w zbiorze uczącym sprawia, że spada liczba porównań wektorów, a tym samym zwiększa się prawdopodobieństwo nieodpowiedniego przyporządkowania. Im bardziej rośnie dysproporcja między zbiorom uczącym, a testowym, tym gorsze wyniki klasyfikacji uzyskujemy.

### 5.3. Wyniki dla różnych wektorów cech

Wykorzystane wektory cech opisano we wstępie do rozdziału 5.

Tabela 28. Różne wektory cech – eksperyment 1–5, ograniczony zbiór artykułów

Numer eksperymentu	1	2	3	4	5
Parametr $k$	5				
Metryka	euklidesowa				
Miara podobieństwa	trigramów				
Podział zbioru (uczący/testowy)	70%/30%				
Zbiór cech	$I$	$J$	$L$	$K$	$M$
Accuracy	0.81	0.80	0.80	0.76	0.73
Precision	0.80	0.78	0.78	0.66	0.67
Recall	0.81	0.80	0.80	0.76	0.73
F1	0.78	0.77	0.75	0.68	0.68
Precision Canada	0.33	0.14	0.33	NaN	0.29
Precision France	1.00	1.00	1.00	NaN	1.00
Precision Japan	0.40	0.75	0.67	0.20	0.14
Precision UK	1.00	0.86	1.00	0.60	0.33
Precision USA	0.83	0.82	0.80	0.78	0.33
Precision West Germany	NaN	NaN	NaN	NaN	0.00
Recall Canada	0.38	0.12	0.12	0.00	0.25
Recall France	1.00	1.00	0.83	0.00	0.50
Recall Japan	0.40	0.60	0.4	0.20	0.20
Recall UK	0.32	0.27	0.23	0.14	0.05
Recall USA	0.95	0.96	0.98	0.99	0.93
Recall West Germany	0.00	0.00	0.00	0.00	0.00
F1 Canada	0.35	0.13	0.18	NaN	0.27
F1 France	1.00	1.00	0.91	NaN	0.67
F1 Japan	0.40	0.67	0.50	0.20	0.17
F1 UK	0.48	0.41	0.37	0.22	0.08
F1 USA	0.89	0.88	0.88	0.87	0.85
F1 West Germany	NaN	NaN	NaN	NaN	NaN



Rysunek 24. Wykres przedstawiający wyniki eksperymentów 1–5 dla różnych cech, ograniczony zbiór artykułów

Tabela 29. Różne wektory cech, eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	3	0	0	1	5	0
France	0	6	0	0	0	0
Japan	0	0	2	1	2	0
UK	0	0	0	7	0	0
USA	5	0	3	13	127	5
West Germany	0	0	0	0	0	0

Tabela 30. Różne wektory cech, eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	1	0	0	1	5	0
France	0	6	0	0	0	0
Japan	0	0	3	1	0	0
UK	0	0	0	6	1	0
USA	7	0	2	14	128	5
West Germany	0	0	0	0	0	0

Tabela 31. Różne wektory cech, eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	1	0	0	0	2	0
France	0	5	0	0	0	0
Japan	0	0	2	0	1	0
UK	0	0	0	5	0	0
USA	7	1	3	17	131	5
West Germany	0	0	0	0	0	0

Tabela 32. Różne wektory cech, eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	0	0
France	0	0	0	0	0	0
Japan	2	1	1	0	1	0
UK	1	0	0	3	1	0
USA	5	5	4	19	132	5
West Germany	0	0	0	0	0	0

Tabela 33. Różne wektory cech, eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	0	5	0
France	0	3	0	0	0	0
Japan	0	1	1	1	4	0
UK	0	0	0	1	0	2
USA	6	2	4	20	124	3
West Germany	0	0	0	0	1	0

Wybór podzbiorów wektora cech ma wpływ na wszystkie miary jakości klasyfikacji. Najlepsze wyniki otrzymano po uwzględnieniu jedynie cech tekstowych, czyli dla wektora  $I$  (5.1). Różnica nie jest jednak zbyt duża w stosunku do wektorów  $J$  (5.2) i  $L$  (5.4) i może wynikać jedynie ze specyfiki pozostałych parametrów algorytmu. Częścią wspólną tych trzech wektorów są cechy  $C_4$  (2.1.4) i  $C_5$  (2.1.5). Na podstawie przedstawionych badań, można stwierdzić, że cechy te mają największy wpływ na polepszenie jakości klasyfikacji.

Zdecydowanie gorsze rezultaty otrzymano dla wektorów  $K$  (5.3) i  $M$  (5.5). Częścią wspólną tych wektorów jest cecha  $C_2$  (2.1.2), sprawdzającą czy nazwa państwa występuje w tytule artykułu. Wektory  $L$  oraz  $I$ , dla których otrzymano dużo lepsze rezultaty nie posiadają tej cechy. Możemy stwierdzić, że cecha  $C_2$  ma negatywny na jakość klasyfikacji.

#### 5.4. Wyniki dla różnych metryk i miar podobieństwa

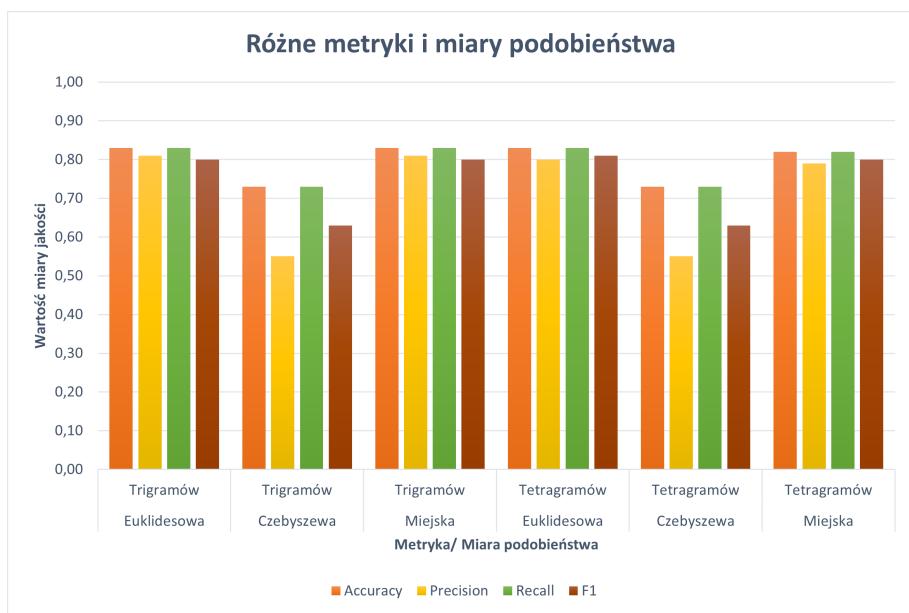
Eksperymenty przeprowadzane były dla 3 metryk i 2 miar.

Tabela 34. Różne metryki i miary podobieństwa – eksperyment 1–3, ograniczony zbiór artykułów

Numer eksperymentu	1	2	3
Parametr $k$	3		
Podział zbioru (uczący/testowy)	70%/30%		
Zbiór cech	$J$		
Miara podobieństwa	<i>trigramów</i>		
Metryka	<i>euklidesowa</i>	<i>Czebyszewa</i>	<i>miejska</i>
Accuracy	0.83	0.73	0.83
Precision	0.81	0.55	0.81
Recall	0.83	0.73	0.83
F1	0.80	0.63	0.80
Precision Canada	0.33	0.00	0.40
Precision France	1.00	NaN	1.00
Precision Japan	0.75	NaN	0.75
Precision UK	0.89	0.00	0.90
Precision USA	0.84	0.74	0.84
Precision West Germany	0.00	NaN	NaN
Recall Canada	0.25	0.00	0.25
Recall France	1.00	0.00	1.00
Recall Japan	0.60	0.00	0.60
Recall UK	0.36	0.00	0.41
Recall USA	0.97	0.98	0.97
Recall West Germany	0.00	0.00	0.00
F1 Canada	0.29	NaN	0.31
F1 France	1.00	NaN	1.00
F1 Japan	0.67	NaN	0.67
F1 UK	0.52	NaN	0.56
F1 USA	0.90	0.84	0.90
F1 West Germany	NaN	NaN	NaN

Tabela 35. Różne metryki i miary podobieństwa – eksperyment 4–6, ograniczony zbiór artykułów

Numer eksperymentu	4	5	6
Parametr $k$	3		
Podział zbioru (uczący/testowy)	70%/30%		
Zbiór cech	$J$		
Miara podobieństwa	tetragramów		
Metryka	<i>euklidesowa</i>	<i>Czebyszewa</i>	<i>miejska</i>
Accuracy	0.83	0.73	0.82
Precision	0.80	0.55	0.79
Recall	0.83	0.73	0.82
F1	0.81	0.63	0.80
Precision Canada	0.33	0.00	0.33
Precision France	1.00	NaN	1.00
Precision Japan	1.00	NaN	1.00
Precision UK	0.83	0.00	0.77
Precision USA	0.84	0.74	0.84
Precision West Germany	0.00	NaN	NaN
Recall Canada	0.25	0.00	0.25
Recall France	1.00	0.00	1.00
Recall Japan	0.60	0.00	0.60
Recall UK	0.45	0.00	0.45
Recall USA	0.96	0.98	0.95
Recall West Germany	0.00	0.00	0.00
F1 Canada	0.29	NaN	0.29
F1 France	1.00	NaN	1.00
F1 Japan	0.75	NaN	0.75
F1 UK	0.59	NaN	0.57
F1 USA	0.90	0.84	0.89
F1 West Germany	NaN	NaN	NaN



Rysunek 25. Wykres przedstawiający wyniki eksperymentów 1–6 dla metryk i miar podobieństwa, ograniczony zbiór artykułów

Tabela 36. Różne metryki i miary podobieństwa, eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	1	3	0
France	0	6	0	0	0	0
Japan	0	0	3	1	0	0
UK	0	0	0	8	1	0
USA	6	0	1	12	130	5
West Germany	0	0	1	0	0	0

Tabela 37. Różne metryki i miary podobieństwa, eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	2	0
France	0	0	0	0	0	0
Japan	0	0	0	0	0	0
UK	0	0	0	0	1	0
USA	8	6	5	22	131	5
West Germany	0	0	0	0	0	0

Tabela 38. Różne metryki i miary podobieństwa, eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	0	3	0
France	0	6	0	0	0	0
Japan	0	0	3	1	0	0
UK	0	0	0	9	1	0
USA	6	0	2	12	130	5
West Germany	0	0	0	0	0	0

Tabela 39. Różne metryki i miary podobieństwa, eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	0	4	0
France	0	6	0	0	0	0
Japan	0	0	3	0	0	0
UK	0	0	0	10	2	0
USA	6	0	1	12	128	5
West Germany	0	0	1	0	0	0

Tabela 40. Różne metryki i miary podobieństwa, eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	0	0	0	0	2	0
France	0	0	0	0	0	0
Japan	0	0	0	0	0	0
UK	0	0	0	0	1	0
USA	8	6	5	22	131	5
West Germany	0	0	0	0	0	0

Tabela 41. Różne metryki i miary podobieństwa, eksperyment 6, ograniczony zbiór artykułów – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	2	0	0	0	4	0
France	0	6	0	0	0	0
Japan	0	0	3	0	0	0
UK	0	0	0	10	3	0
USA	6	0	2	12	127	5
West Germany	0	0	0	0	0	0

Dla metody *trigramów* (3.1) i metody *tetragramów* (3.2) użytych jako miary podobieństwa, wyniki różnią się nieznacznie (34 i 35). Oznacza to, że wybrana miara podobieństwa tekstów nie ma dużego wpływu na jakość klasyfikacji.

Wyniki klasyfikacji dla metryk *eukliidesowej* (3.3.1) i metryki *miejskiej* 3.3.2 są bardzo zbliżone, a to która z metryk spisze się lepiej może wynikać z dobrania pozostałych parametrów algorytmu. Metryka *Czebyszewa* (3.3.3) sprawdza się dużo gorzej od pozostałych metryk. Dzieje się tak, gdyż maksymalna różnica między poszczególnymi współrzędnymi wektora wynosi 1 i pojawia się stosunkowo często, na przykład gdy wektory zawierają dwa niepodobne do siebie wyrazy. Na podstawie tabel (37 i 40) widzimy, że artykuły z pewnymi wyjątkami trafiły do klasy *usa*. Taka sytuacja ma miejsce, jedynie gdy wektor cech artykułu ze zbioru testowego znajdzie wektor o podobnych cechach tekstowych. Metryka *Czebyszewa* da lepsze wyniki dla wektorów składających się jedynie z cech liczbowych.

## 6. Dyskusja, wnioski, sprawozdanie końcowe

Tabela 42. Liczebność artykułów dla poszczególnych klas.

Kraj	Liczebność artykułów w zbiorze
Canada	849
France	273
Japan	498
UK	938
USA	10878
West Germany	330

### 6.1. Klasyfikacja dla różnych wartości parametru $k$

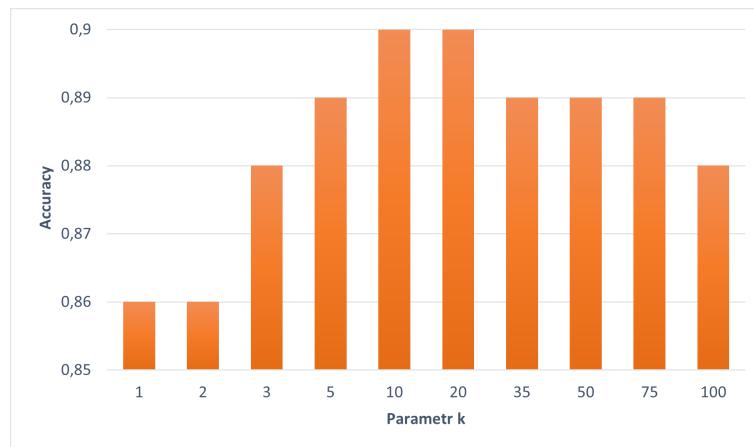
Podczas przeprowadzania eksperymentów użyliśmy 10 różnych wartości parametru  $k$ : 1, 2, 3, 5, 10, 20, 35, 50, 75, 100.

Tabela 43. Różne wartości parametru  $k$  – eksperyment 1–5

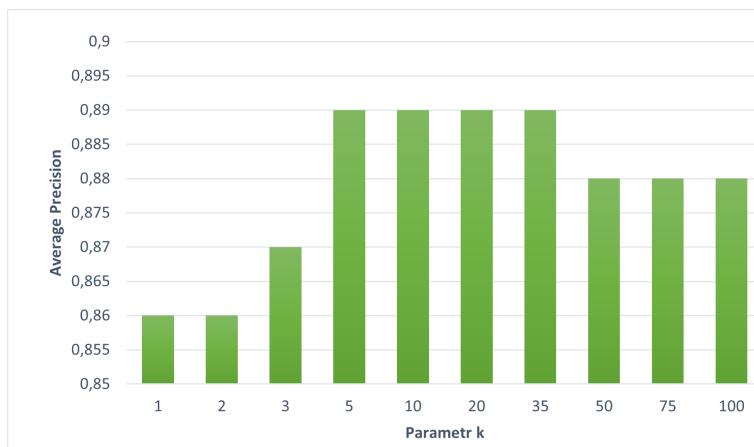
Numer eksperymentu	1	2	3	4	5
Parametr $k$	1	2	3	5	10
Metryka	<i>euklidesowa</i>				
Miara podobieństwa	<i>trigramów</i>				
Podział zbioru (uczący/testowy)	70%/30%				
Zbiór cech	$J$				
Accuracy	0.86	0.86	0.88	0.89	0.90
Precision	0.86	0.86	0.87	0.89	0.89
Recall	0.86	0.86	0.88	0.89	0.90
F1	0.86	0.86	0.88	0.89	0.89
Precision Canada	0.48	0.48	0.63	0.73	0.76
Precision France	0.76	0.76	0.78	0.84	0.87
Precision Japan	0.89	0.89	0.82	0.85	0.92
Precision UK	0.45	0.45	0.53	0.60	0.70
Precision USA	0.92	0.92	0.92	0.92	0.91
Precision West Germany	0.61	0.61	0.74	0.85	0.87
Recall Canada	0.56	0.56	0.54	0.56	0.53
Recall France	0.75	0.75	0.83	0.80	0.76
Recall Japan	0.68	0.68	0.66	0.67	0.69
Recall UK	0.50	0.50	0.47	0.46	0.43
Recall USA	0.92	0.92	0.95	0.97	0.98
Recall West Germany	0.62	0.62	0.58	0.62	0.56
F1 Canada	0.52	0.51	0.58	0.63	0.63
F1 France	0.75	0.75	0.80	0.82	0.81
F1 Japan	0.77	0.77	0.73	0.75	0.79
F1 UK	0.47	0.47	0.50	0.52	0.54
F1 USA	0.92	0.92	0.94	0.94	0.95
F1 West Germany	0.62	0.62	0.65	0.72	0.68

Tabela 44. Różne wartości parametru  $k$  – eksperyment 6–10

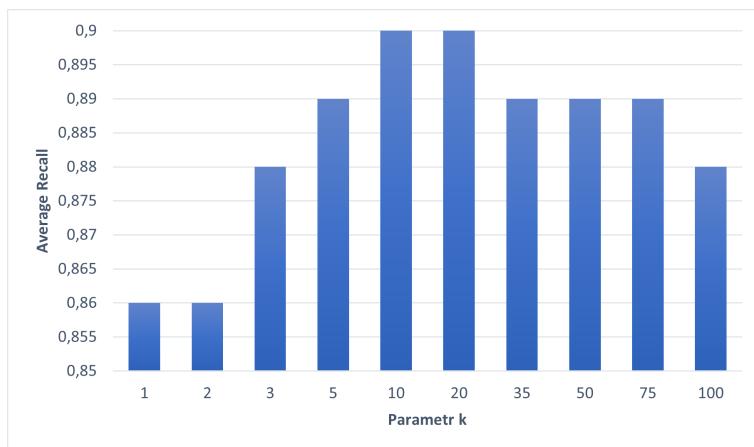
Numer eksperymentu	6	7	8	9	10
Parametr $k$	20	35	50	75	100
Metryka	euklidesowa				
Miara podobieństwa	trigramów				
Podział zbioru (uczący/testowy)	70%/30%				
Zbiór cech	$J$				
Accuracy	0.90	0.89	0.89	0.89	0.88
Precision	0.89	0.89	0.88	0.88	0.88
Recall	0.90	0.89	0.89	0.89	0.88
F1	0.89	0.88	0.87	0.87	0.86
Precision Canada	0.83	0.80	0.79	0.86	0.85
Precision France	0.84	0.84	0.85	0.90	0.90
Precision Japan	0.92	0.91	0.89	0.91	0.93
Precision UK	0.69	0.68	0.68	0.70	0.70
Precision USA	0.91	0.90	0.90	0.89	0.88
Precision West Germany	0.92	0.93	0.93	0.92	0.97
Recall Canada	0.52	0.49	0.47	0.45	0.40
Recall France	0.78	0.73	0.69	0.64	0.61
Recall Japan	0.68	0.61	0.56	0.56	0.52
Recall UK	0.41	0.38	0.35	0.34	0.33
Recall USA	0.98	0.99	0.99	0.99	0.99
Recall West Germany	0.47	0.38	0.38	0.35	0.30
F1 Canada	0.64	0.61	0.59	0.59	0.55
F1 France	0.81	0.78	0.77	0.75	0.73
F1 Japan	0.78	0.73	0.69	0.69	0.67
F1 UK	0.51	0.49	0.46	0.45	0.45
F1 USA	0.94	0.94	0.94	0.94	0.93
F1 West Germany	0.62	0.54	0.54	0.50	0.45



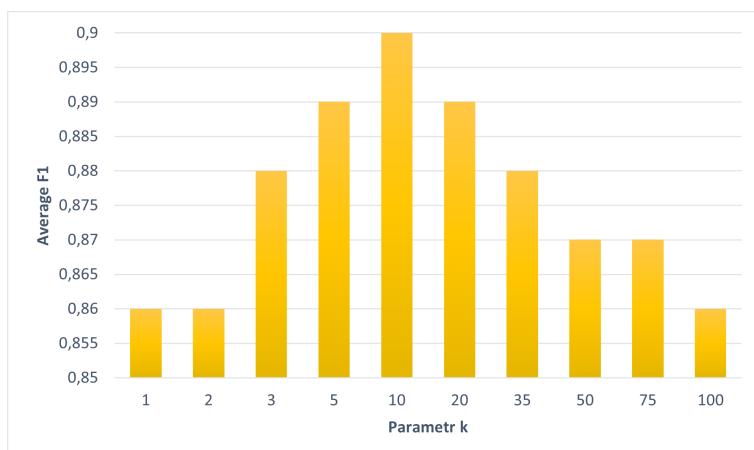
Rysunek 26. Wyniki Accuracy dla różnych wartości parametru  $k$



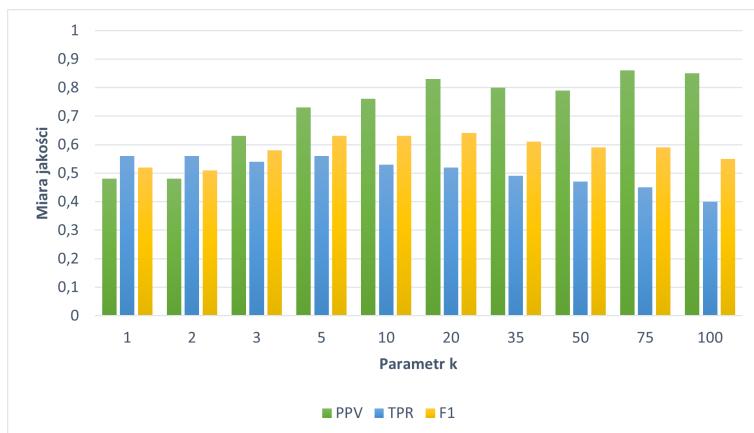
Rysunek 27. Średni wynik *Precision* dla różnych wartości parametru  $k$



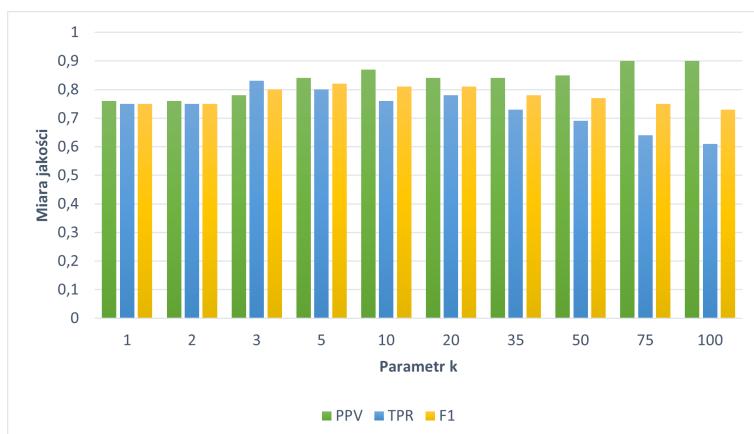
Rysunek 28. Średni wynik *Recall* dla różnych wartości parametru  $k$



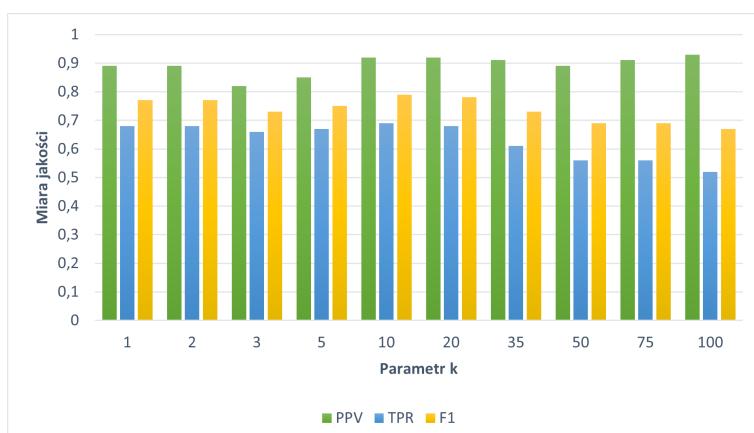
Rysunek 29. Średni wynik *F1* dla różnych wartości parametru  $k$



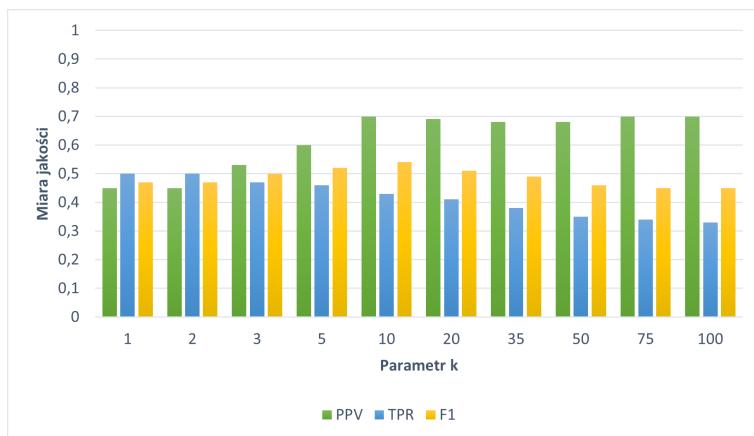
Rysunek 30. Wyniki miar jakości dla klasy *canada* dla różnych wartości parametru  $k$



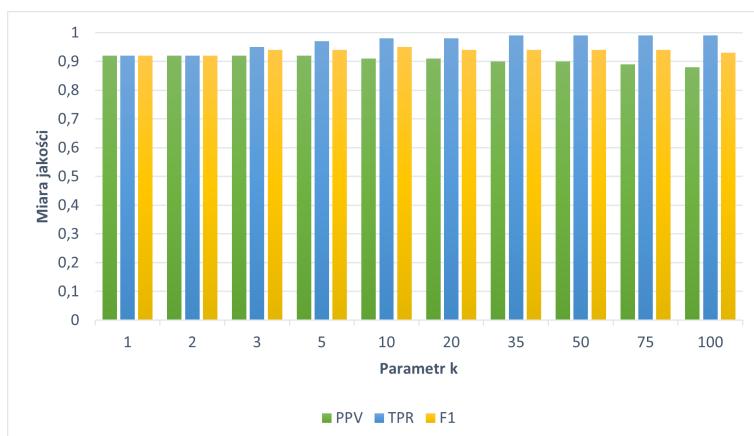
Rysunek 31. Wyniki miar jakości dla klasy *france* dla różnych wartości parametru  $k$



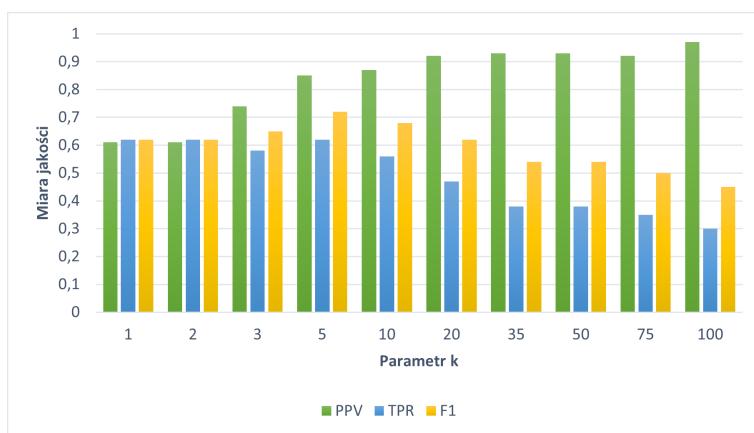
Rysunek 32. Wyniki miar jakości dla klasy *japan* dla różnych wartości parametru  $k$



Rysunek 33. Wyniki miar jakości dla klasy  $uk$  dla różnych wartości parametru  $k$



Rysunek 34. Wyniki miar jakości dla klasy  $usa$  dla różnych wartości parametru  $k$



Rysunek 35. Wyniki miar jakości dla klasy  $west-germany$  dla różnych wartości parametru  $k$

Tabela 45. Różne wartości parametru k, eksperyment 1 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	130	1	1	11	127	0
France	0	44	1	1	9	3
Japan	1	0	121	3	11	0
UK	11	5	13	103	93	6
USA	92	9	42	88	3076	28
West Germany	0	0	1	2	36	61

Tabela 46. Różne wartości parametru k, eksperyment 2 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	131	1	1	11	131	0
France	0	44	1	1	9	3
Japan	1	0	121	3	11	0
UK	11	5	13	103	93	6
USA	91	9	42	88	3072	28
West Germany	0	0	1	2	36	61

Tabela 47. Różne wartości parametru k, eksperyment 3 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	131	1	1	11	131	0
France	0	44	1	1	9	3
Japan	1	0	121	3	11	0
UK	11	5	13	103	93	6
USA	91	9	42	88	3072	28
West Germany	0	0	1	2	36	61

Tabela 48. Różne wartości parametru k, eksperyment 4 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	130	0	1	7	39	0
France	0	47	1	3	3	2
Japan	0	1	120	4	17	0
UK	4	3	8	95	44	4
USA	99	8	48	97	3242	31
West Germany	1	0	1	2	7	61

Tabela 49. Różne wartości parametru k, eksperyment 5 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	125	0	1	6	32	1
France	0	45	1	3	1	2
Japan	0	0	123	4	7	0
UK	5	2	4	90	25	2
USA	103	12	49	105	3281	38
West Germany	1	0	1	0	6	55

Tabela 50. Różne wartości parametru k, eksperyment 6 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	121	0	1	6	17	0
France	0	46	1	3	2	3
Japan	0	0	121	5	6	0
UK	5	2	2	85	28	2
USA	108	11	53	109	3296	47
West Germany	0	0	1	0	3	46

Tabela 51. Różne wartości parametru k, eksperyment 7 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	114	0	9	6	13	0
France	0	43	1	3	1	3
Japan	0	0	110	5	6	0
UK	5	2	3	79	23	5
USA	115	14	55	115	3307	53
West Germany	0	0	1	0	2	37

Tabela 52. Różne wartości parametru k, eksperyment 8 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	111	0	9	6	15	0
France	0	41	1	3	1	2
Japan	0	0	101	4	8	0
UK	5	2	4	73	20	3
USA	118	16	63	122	3306	56
West Germany	0	0	1	0	2	37

Tabela 53. Różne wartości parametru k, eksperyment 9 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	106	0	3	1	13	0
France	0	38	0	3	0	1
Japan	0	0	100	2	8	0
UK	3	3	2	70	19	3
USA	125	18	73	132	3310	60
West Germany	0	0	1	0	2	34

Tabela 54. Różne wartości parametru k, eksperyment 10 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	94	0	3	1	12	0
France	0	36	0	3	0	1
Japan	0	0	93	2	4	1
UK	1	3	2	69	20	3
USA	139	20	80	133	3316	64
West Germany	0	0	1	0	0	29

Wybór parametru k wpływa na wynik klasyfikacji. Najlepsze wyniki uzyskano dla  $k = 10$  oraz dla  $k = 20$ . Dalsze zwiększanie parametru  $k$  sprawi, że coraz więcej artykułów będzie klasyfikowanych jako *usa*, z uwagi, że klasa ta jest zdecydowanie najliczniej reprezentowana w zbiorze uczącym (56). Sprawia to, iż coraz więcej dalszych sąsiadów, będzie pochodziło z tej klasy. Efekt ten objawia się spadkiem miary *Precision* dla klasy *usa* (rys. 34). Niesie za sobą również spadki miary *Recall* dla innych klas (rys.30, rys.31, rys.32, rys.33, rys.35). Z kolei zbyt mała wartość  $k$  sprawia, że maleje liczba porównań, a tym samym mniej wektorów ze zbioru uczącego jest ostatecznie brana pod uwagę. Taka sytuacja objawia się niższymi wartościami miary *Precision* dla klas reprezentowanych mniej licznie, czyli wszystkich poza *usa* (rys.30, rys.31, rys.32, rys.33, rys.35).

## 6.2. Klasyfikacja dla różnych podziałów zbioru artykułów

Tabela 55. Liczebność artykułów w zbiorze. Podział 80%/20%

Podział	80%/20%	
	Kraj	Liczebność artykułów w zbiorze uczącym
Canada	691	158
France	242	31
Japan	396	102
UK	817	121
USA	8595	2283
West Germany	271	59

Tabela 56. Liczebność artykułów w zbiorze. Podział 70%/30%

Podział	70%/30%	
Kraj	Liczebność artykułów w zbiorze uczącym	Liczebność artykułów w zbiorze testowym
Canada	615	234
France	214	59
Japan	319	179
UK	730	208
USA	7526	3352
West Germany	232	98

Tabela 57. Liczebność artykułów w zbiorze. Podział 60%/40%

Podział	60%/40%	
Kraj	Liczebność artykułów w zbiorze uczącym	Liczebność artykułów w zbiorze testowym
Canada	518	331
France	177	96
Japan	267	231
UK	619	319
USA	6486	4392
West Germany	192	138

Tabela 58. Liczebność artykułów w zbiorze. Podział 50%/50%

Podział	50%/50%	
Kraj	Liczebność artykułów w zbiorze uczącym	Liczebność artykułów w zbiorze testowym
Canada	455	394
France	151	122
Japan	192	306
UK	518	420
USA	5417	5461
West Germany	150	180

Tabela 59. Liczebność artykułów w zbiorze. Podział 40%/60%

Podział	40%/60%	
Kraj	Liczebność artykułów w zbiorze uczącym	Liczebność artykułów w zbiorze testowym
Canada	387	462
France	118	155
Japan	141	357
UK	396	542
USA	4361	6517
West Germany	103	227

Tabela 60. Liczebność artykułów w zbiorze. Podział 30%/70%

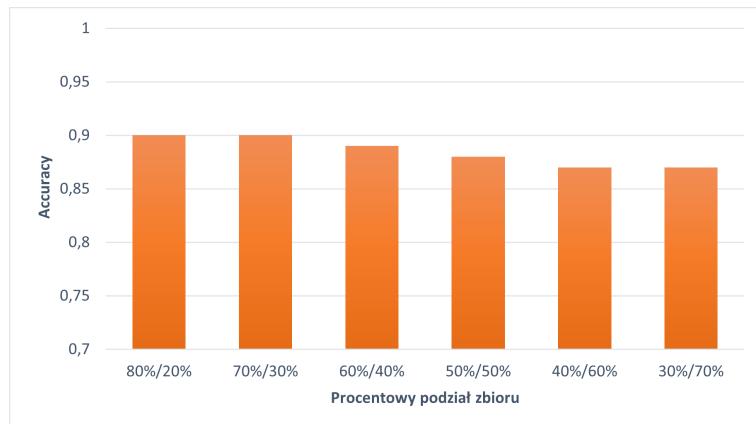
Podział	30%/70%	
	Kraj	Liczebność artykułów w zbiorze uczącym
Canada	296	553
France	94	179
Japan	121	377
UK	291	647
USA	3255	7623
West Germany	72	258

Tabela 61. Różne podziały zbioru – eksperyment 1–3

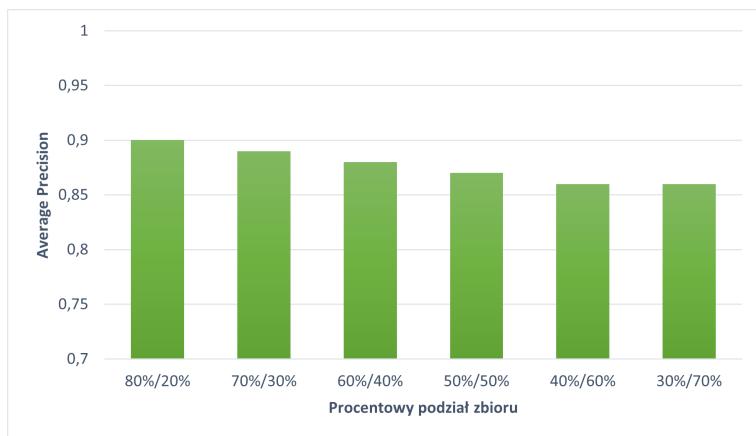
Numer eksperymentu	1	2	3
Parametr $k$	20		
Metryka	euklidesowa		
Miara podobieństwa	trigramów		
Zbiór cech	J		
Podział zbioru (uczący/testowy)	80%/20%	70%/30%	60%/40%
Accuracy	0.90	0.90	0.89
Precision	0.90	0.89	0.88
Recall	0.90	0.90	0.89
F1	0.89	0.89	0.88
Precision Canada	0.84	0.83	0.78
Precision France	0.78	0.84	0.87
Precision Japan	0.89	0.92	0.89
Precision UK	0.64	0.69	0.72
Precision USA	0.92	0.91	0.90
Precision West Germany	0.86	0.92	0.93
Recall Canada	0.51	0.52	0.49
Recall France	0.81	0.78	0.69
Recall Japan	0.65	0.68	0.63
Recall UK	0.40	0.41	0.42
Recall USA	0.98	0.98	0.98
Recall West Germany	0.42	0.47	0.45
F1 Canada	0.63	0.64	0.60
F1 France	0.79	0.81	0.77
F1 Japan	0.75	0.78	0.74
F1 UK	0.49	0.51	0.53
F1 USA	0.95	0.94	0.94
F1 West Germany	0.57	0.62	0.60

Tabela 62. Różne podziały zbioru – eksperyment 4–6

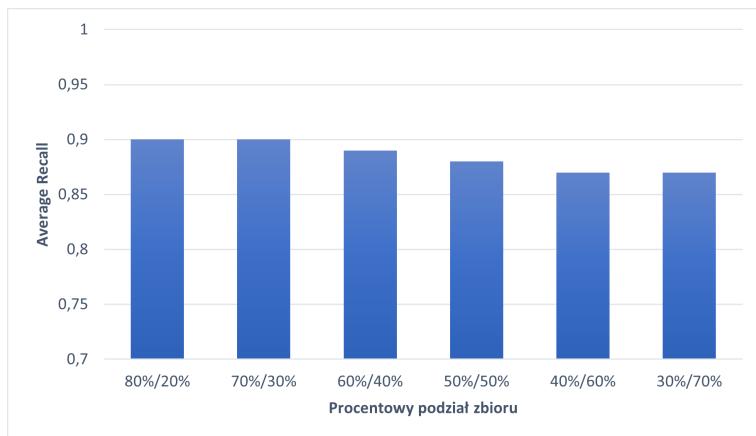
Numer eksperymentu	4	5	6
Parametr $k$	20		
Metryka	<i>euklidesowa</i>		
Miara podobieństwa	<i>trigramów</i>		
Zbiór cech	J		
Podział zbioru (uczący/testowy)	50%/50%	40%/60%	30%/70%
Accuracy	0.88	0.87	0.87
Precision	0.87	0.86	0.86
Recall	0.88	0.87	0.87
F1	0.87	0.85	0.85
Precision Canada	0.73	0.71	0.70
Precision France	0.86	0.85	0.85
Precision Japan	0.89	0.91	0.88
Precision UK	0.68	0.70	0.72
Precision USA	0.89	0.88	0.88
Precision West Germany	0.94	0.95	0.96
Recall Canada	0.48	0.43	0.41
Recall France	0.68	0.70	0.71
Recall Japan	0.63	0.63	0.63
Recall UK	0.41	0.36	0.36
Recall USA	0.98	0.98	0.98
Recall West Germany	0.38	0.33	0.31
F1 Canada	0.58	0.54	0.52
F1 France	0.76	0.77	0.77
F1 Japan	0.74	0.75	0.73
F1 UK	0.51	0.48	0.48
F1 USA	0.93	0.93	0.93
F1 West Germany	0.54	0.49	0.47



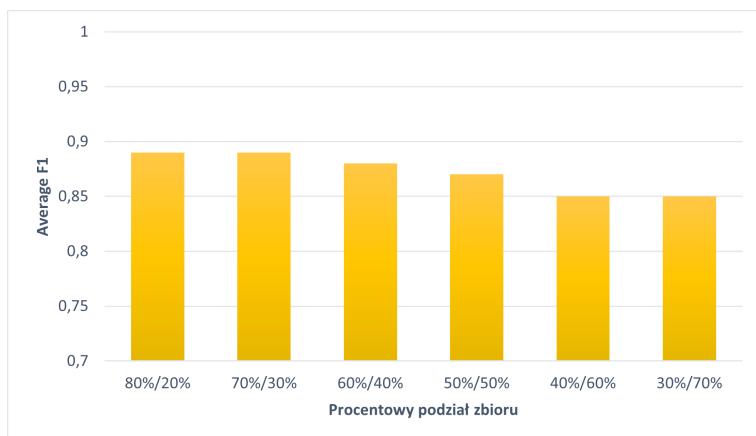
Rysunek 36. Wyniki *Accuracy* dla różnych podziałów zbioru



Rysunek 37. Średni wynik *Precision* dla różnych podziałów zbioru



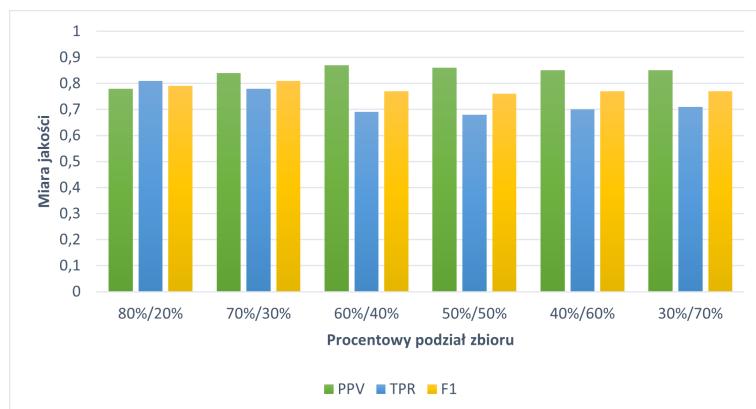
Rysunek 38. Średni wynik *Recall* dla różnych podziałów zbioru



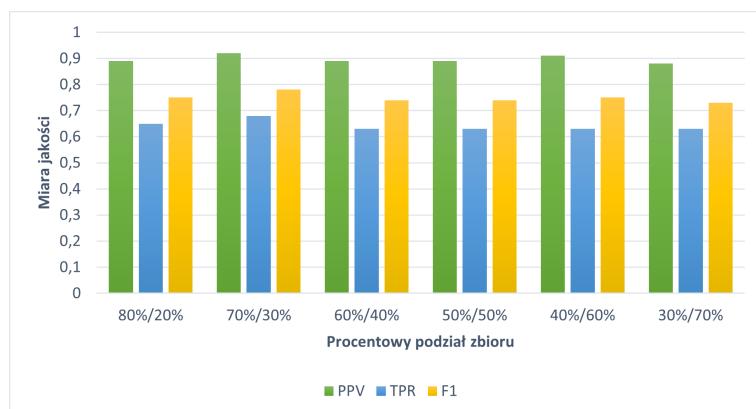
Rysunek 39. Średni wynik *F1* dla różnych podziałów zbioru



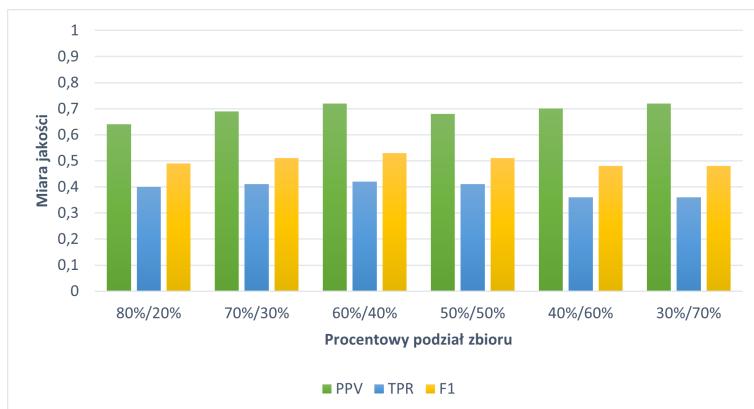
Rysunek 40. Wyniki miar jakości dla klasy *canada* dla różnych podziałów zbioru



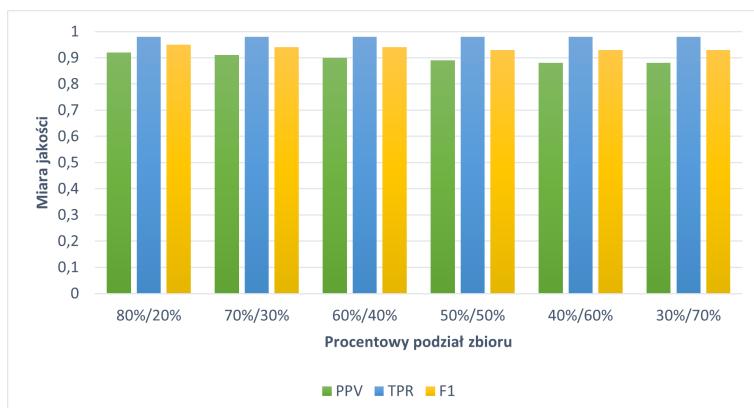
Rysunek 41. Wyniki miar jakości dla klasy *france* dla różnych podziałów zbioru



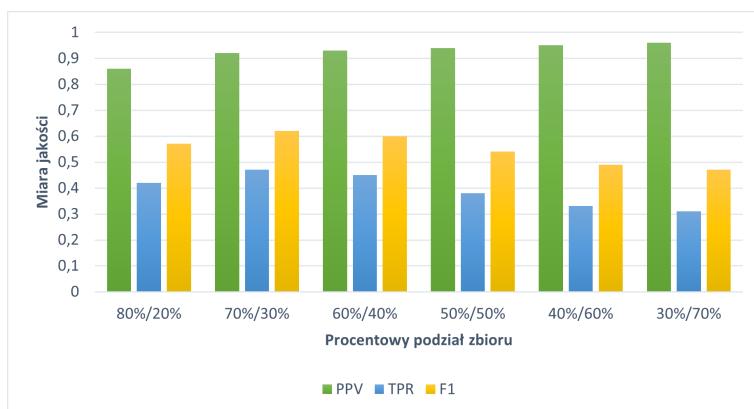
Rysunek 42. Wyniki miar jakości dla klasy *japan* dla różnych podziałów zbioru



Rysunek 43. Wyniki miar jakości dla klasy *uk* dla różnych podziałów zbioru



Rysunek 44. Wyniki miar jakości dla klasy *usa* dla różnych podziałów zbioru



Rysunek 45. Wyniki miar jakości dla klasy *west-germany* dla różnych podziałów zbioru

Tabela 63. Różne podziały zbioru, eksperyment 1 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	80	0	1	3	11	0
France	0	25	1	2	1	3
Japan	0	1	66	1	6	0
UK	4	2	2	48	18	1
USA	74	3	31	67	2244	30
West Germany	0	0	1	0	3	25

Tabela 64. Różne podziały zbioru, eksperyment 2 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	121	0	1	6	17	0
France	0	46	1	3	2	3
Japan	0	0	121	5	6	0
UK	5	2	2	85	28	2
USA	108	11	53	109	3296	47
West Germany	0	0	1	0	3	46

Tabela 65. Różne podziały zbioru, eksperyment 3 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	162	2	12	10	22	0
France	0	66	1	3	3	3
Japan	0	2	146	8	7	1
UK	6	4	3	135	36	3
USA	163	22	68	163	4320	69
West Germany	0	0	1	0	4	62

Tabela 66. Różne podziały zbioru, eksperyment 4 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	188	3	16	16	36	0
France	0	83	1	4	4	4
Japan	0	1	193	10	12	1
UK	8	5	6	174	58	5
USA	198	30	89	216	5348	102
West Germany	0	0	1	0	3	68

Tabela 67. Różne podziały zbioru, eksperyment 5 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	199	5	17	15	44	0
France	0	109	1	4	7	7
Japan	0	1	225	9	11	1
UK	10	5	9	197	55	6
USA	253	35	104	316	6398	139
West Germany	0	0	1	1	2	74

Tabela 68. Różne podziały zbioru, eksperyment 6 – tablica pomyłek

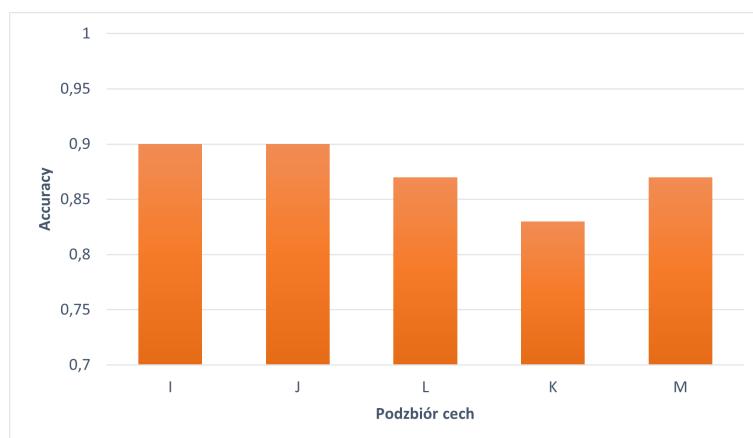
	Canada	France	Japan	UK	USA	West Germany
Canada	225	6	15	28	46	0
France	0	127	0	6	10	6
Japan	0	1	237	14	17	1
UK	8	5	10	233	62	7
USA	320	40	114	366	7486	164
West Germany	0	0	1	0	2	80

Współczynnik podziału zbioru wpływa na jakość klasyfikacji. Najlepsze wyniki uzyskano dla podziałów 80% - zbiór uczący i 20% - zbiór testowy, a także 70% - zbiór uczący i 30 % -zbiór testowy. Wraz ze wzrostem liczby artykułów w zbiorze uczącym, wyniki klasyfikacji poprawiają się. Po przeprowadzeniu eksperymentów na ograniczonym zbiorze danych, wiemy jednak, że zbyt mała liczba artykułów w zbiorze testowym będzie skutkować pogorszeniem wyników, z uwagi na zwiększenie wagi nieprawidłowo sklasyfikowanych artykułów w końcowym wyniku miary jakości klasyfikacji. Taką sytuację możemy zaobserwować dla miar dotyczących klas *france*, *japan* i *west-germany* (rys. 41, rys.42, rys. 45).

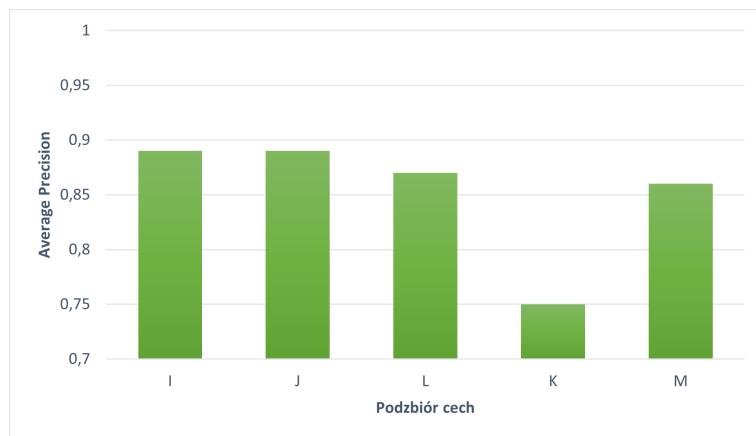
### 6.3. Klasyfikacja dla różnych podzbiorów cech.

Tabela 69. Różne wektory cech – eksperyment 1–5

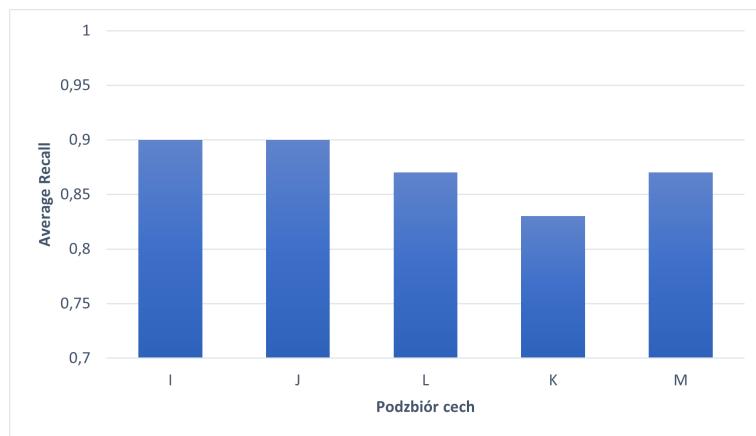
Numer eksperymentu	1	2	3	4	5
Parametr $k$		20			
Metryka		euklidesowa			
Miara podobieństwa		trigramów			
Podział zbioru (uczący/testowy)		70%/30%			
Zbiór cech	I	J	L	K	M
Accuracy	0.90	0.90	0.87	0.83	0.87
Precision	0.89	0.89	0.87	0.75	0.86
Recall	0.90	0.90	0.87	0.83	0.87
F1	0.89	0.89	0.85	0.77	0.85
Precision Canada	0.86	0.83	0.78	0.41	0.85
Precision France	0.82	0.84	0.76	0.00	0.92
Precision Japan	0.86	0.92	0.69	0.53	0.92
Precision UK	0.64	0.69	0.98	0.22	0.46
Precision USA	0.91	0.91	0.89	0.85	0.88
Precision West Germany	0.94	0.92	1.00	0.20	0.90
Recall Canada	0.56	0.52	0.48	0.11	0.31
Recall France	0.78	0.78	0.71	0.00	0.58
Recall Japan	0.70	0.68	0.53	0.34	0.60
Recall UK	0.43	0.41	0.36	0.06	0.23
Recall USA	0.98	0.98	0.98	0.99	0.99
Recall West Germany	0.48	0.47	0.08	0.01	0.44
F1 Canada	0.67	0.64	0.59	0.18	0.45
F1 France	0.80	0.81	0.74	NaN	0.71
F1 Japan	0.77	0.78	0.60	0.41	0.73
F1 UK	0.51	0.51	0.47	0.10	0.30
F1 USA	0.95	0.94	0.93	0.91	0.93
F1 West Germany	0.64	0.62	0.15	0.02	0.59



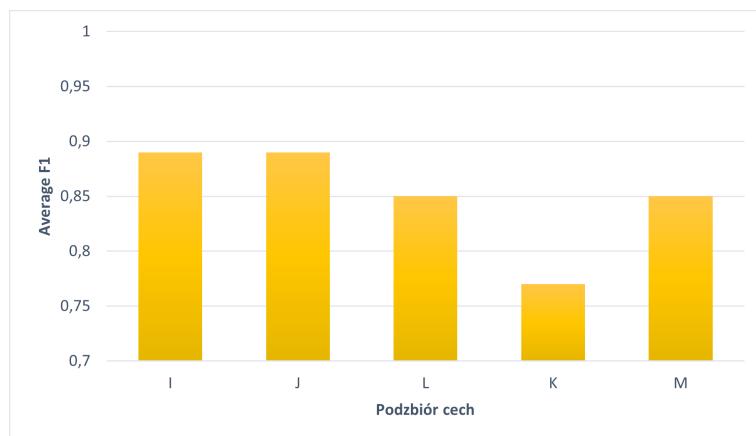
Rysunek 46. Wyniki Accuracy dla różnych wektorów cech



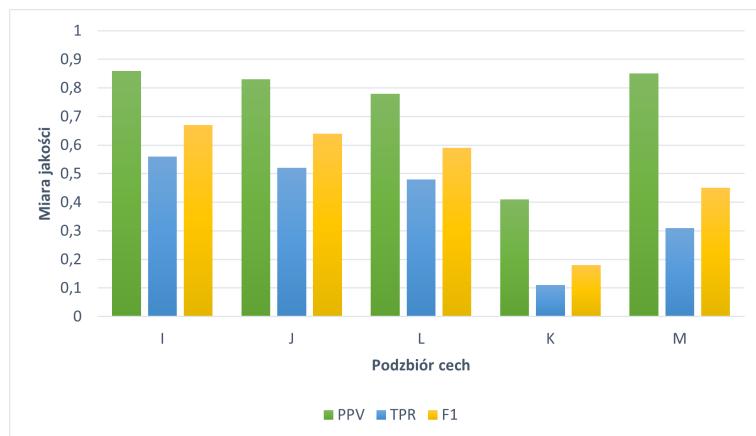
Rysunek 47. Średni wynik *Precision* dla różnych wektorów cech



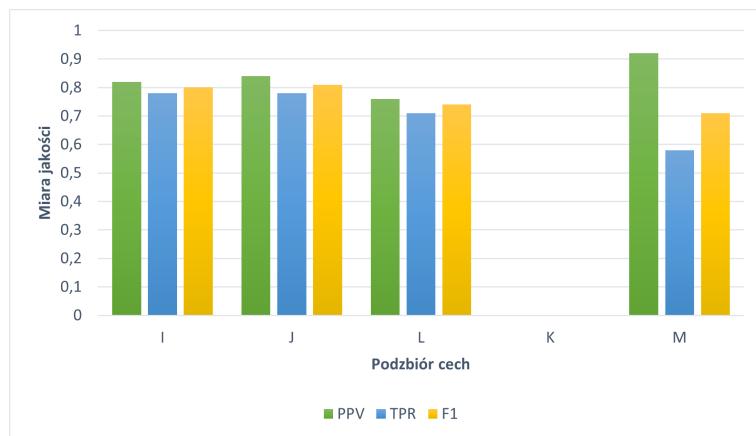
Rysunek 48. Średni wynik *Recall* dla różnych wektorów cech



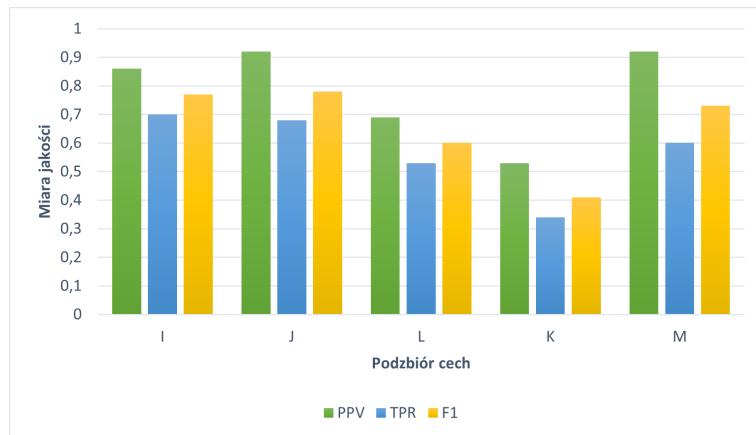
Rysunek 49. Średni wynik *F1* dla różnych wektorów cech



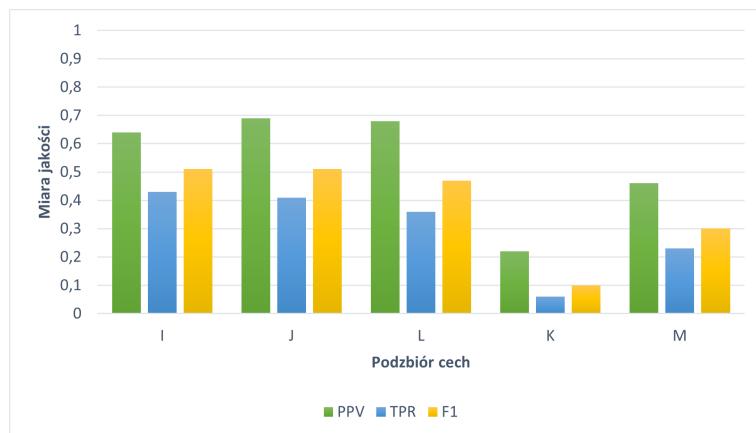
Rysunek 50. Wyniki miar jakości dla klasy *canada* dla różnych wektorów cech



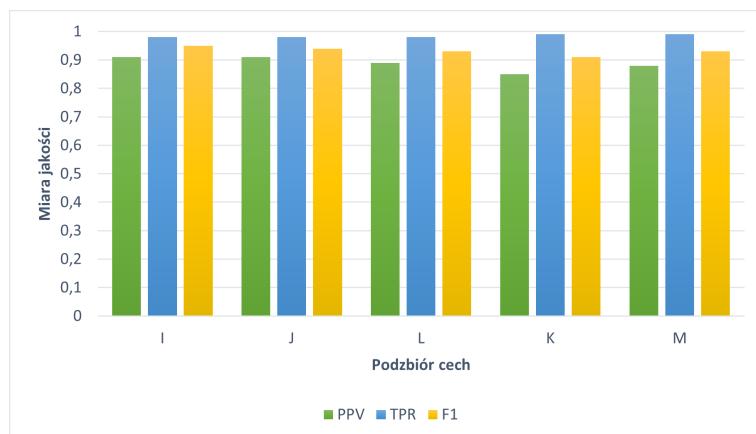
Rysunek 51. Wyniki miar jakości dla klasy *france* dla różnych wektorów cech



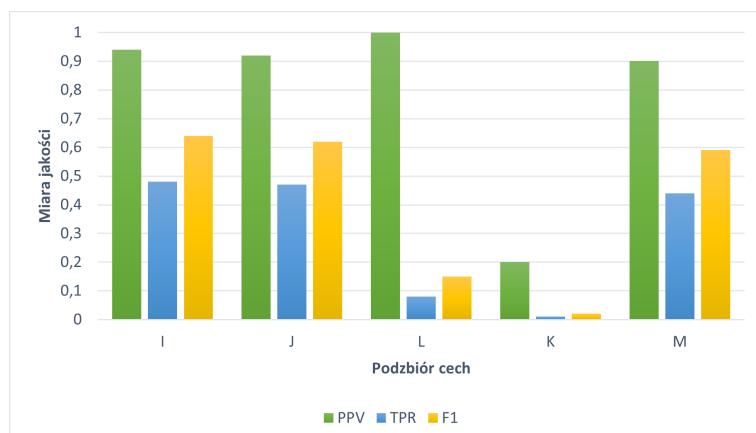
Rysunek 52. Wyniki miar jakości dla klasy *japan* dla różnych wektorów cech



Rysunek 53. Wyniki miar jakości dla klasy *uk* dla różnych wektorów cech



Rysunek 54. Wyniki miar jakości dla klasy *usa* dla różnych wektorów cech



Rysunek 55. Wyniki miar jakości dla klasy *west-germany* dla różnych wektorów cech

Tabela 70. Różne wektory cech, eksperyment 1 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	130	1	1	3	17	0
France	0	46	1	3	3	3
Japan	0	1	126	6	14	0
UK	7	2	2	89	35	3
USA	97	9	48	107	3281	45
West Germany	0	0	1	0	2	47

Tabela 71. Różne wektory cech, eksperyment 2 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	121	0	1	6	17	0
France	0	46	1	3	2	3
Japan	0	0	121	5	6	0
UK	5	2	2	85	28	2
USA	108	11	53	109	3296	47
West Germany	0	0	1	0	3	46

Tabela 72. Różne wektory cech, eksperyment 3 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	112	1	0	2	25	4
France	0	42	0	2	4	7
Japan	1	2	95	6	31	3
UK	8	3	2	74	17	5
USA	113	11	82	124	3275	71
West Germany	0	0	0	0	0	8

Tabela 73. Różne wektory cech, eksperyment 4 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	26	1	22	5	9	0
France	1	0	0	1	2	0
Japan	29	7	60	6	11	0
UK	4	10	6	13	20	6
USA	174	41	90	182	3308	91
West Germany	0	0	1	1	2	1

Tabela 74. Różne wektory cech, eksperyment 5 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	72	1	0	8	4	0
France	0	34	0	1	1	1
Japan	2	0	108	4	3	0
UK	7	3	3	47	37	5
USA	153	21	66	147	3305	49
West Germany	0	0	2	1	2	43

Eksperymenty pozwalają ocenić przydatność poszczególnych cech. Najlepsze wyniki uzyskano dla podzbiorów  $I$  (5.1) i wektorów  $J$  (5.2). Podzbiór  $I$  składa się jedynie z cech tekstowych i to właśnie te cechy najmocniej poprawiają jakość klasyfikacji. Z kolei najlepsze wyniki otrzymano dla podzbioru  $K$  (5.3), składającego się jedynie z cech tekstowych. Dla mniej licznych klas podzbiór  $L$  (5.4) spisał się dużo lepiej niż podzbiór  $M$  (5.5) (rys.50, rys.51, rys.52, rys.53, rys.55). Ponieważ wektory te składają się zarówno z cech liczbowych, jak i tekstowych, możemy ocenić, że najlepszymi cechami są cechy dotyczące najliczniej występującej nazwy państwa w tekście (2.1.4) i najliczniej występującej nazwy geograficznej w tekście (2.1.5). Z kolei najmniej użyteczną cechą jest cecha określająca, czy w tytule znajduje się nazwa jednego z państw (2.1.2).

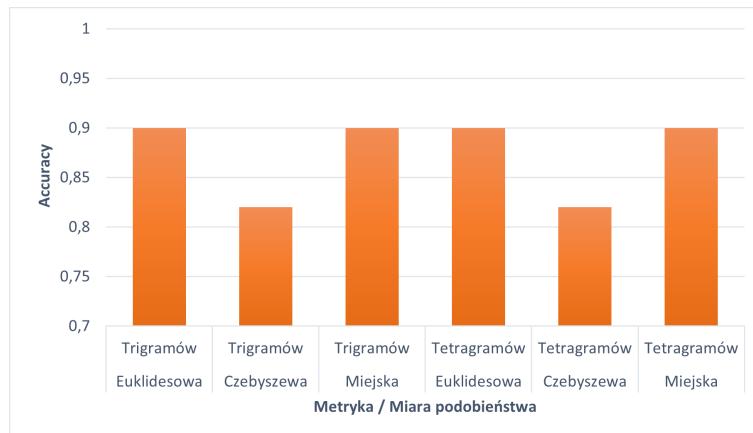
#### 6.4. Klasyfikacja dla różnych metryki i miar podobieństwa.

Tabela 75. Różne metryki i miary podobieństwa – eksperyment 1–3

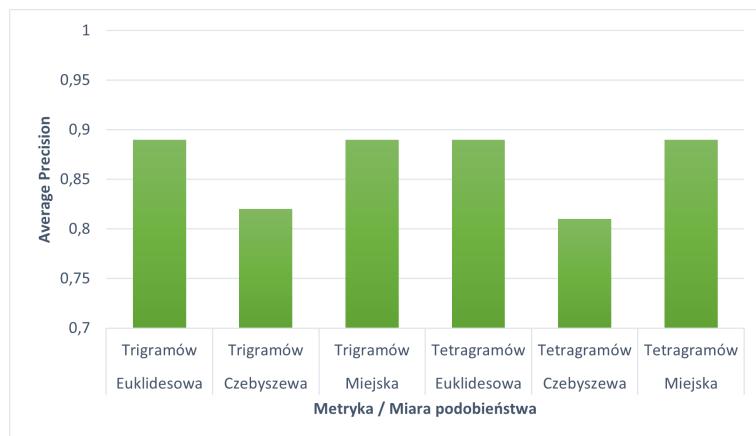
Numer eksperymentu	1	2	3
Parametr $k$	20		
Podział zbioru (uczący/testowy)	70%/30%		
Zbiór cech	$J$		
Miara podobieństwa	<i>trigramów</i>		
Metryk	<i>euklidesowa</i>	<i>Czebyszewa</i>	<i>miejska</i>
Accuracy	0.90	0.82	0.90
Precision	0.89	0.82	0.89
Recall	0.90	0.82	0.90
F1	0.89	0.75	0.89
Precision Canada	0.83	1.00	0.84
Precision France	0.84	1.00	0.85
Precision Japan	0.92	1.00	0.91
Precision UK	0.69	0.71	0.69
Precision USA	0.91	0.82	0.91
Precision West Germany	0.92	NaN	0.92
Recall Canada	0.52	0.03	0.53
Recall France	0.78	0.07	0.78
Recall Japan	0.68	0.04	0.68
Recall UK	0.41	0.14	0.40
Recall USA	0.98	1.00	0.98
Recall West Germany	0.47	0.00	0.47
F1 Canada	0.64	0.06	0.65
F1 France	0.81	0.13	0.81
F1 Japan	0.78	0.09	0.78
F1 UK	0.51	0.23	0.51
F1 USA	0.94	0.90	0.95
F1 West Germany	0.62	NaN	0.62

Tabela 76. Różne metryki i miary podobieństwa – eksperyment 4–6

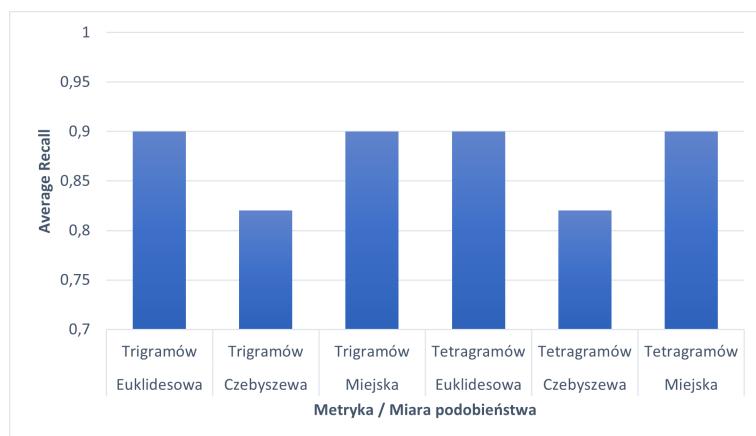
Numer eksperymentu	4	5	6
Parametr $k$	20		
Podział zbioru (uczący/testowy)	70%/30%		
Zbiór cech	$J$		
Miara podobieństwa	tetagramów		
Metryk	euklidesowa	Czebyszewa	miejska
Accuracy	0.90	0.82	0.90
Precision	0.89	0.81	0.89
Recall	0.90	0.82	0.90
F1	0.89	0.75	0.89
Precision Canada	0.84	1.00	0.85
Precision France	0.85	1.00	0.85
Precision Japan	0.94	1.00	0.93
Precision UK	0.70	0.67	0.63
Precision USA	0.90	0.82	0.90
Precision West Germany	0.94	NaN	0.93
Recall Canada	0.53	0.03	0.53
Recall France	0.78	0.03	0.76
Recall Japan	0.58	0.04	0.60
Recall UK	0.40	0.11	0.40
Recall USA	0.98	1.00	0.98
Recall West Germany	0.46	0.00	0.44
F1 Canada	0.65	0.05	0.65
F1 France	0.81	0.07	0.80
F1 Japan	0.72	0.09	0.73
F1 UK	0.51	0.18	0.51
F1 USA	0.94	0.90	0.94
F1 West Germany	0.62	NaN	0.60



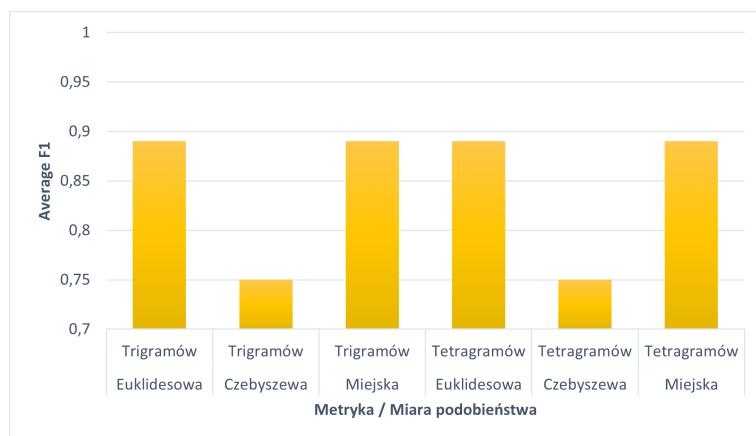
Rysunek 56. Wyniki *Accuracy* dla różnych metryk i miar podobieństwa



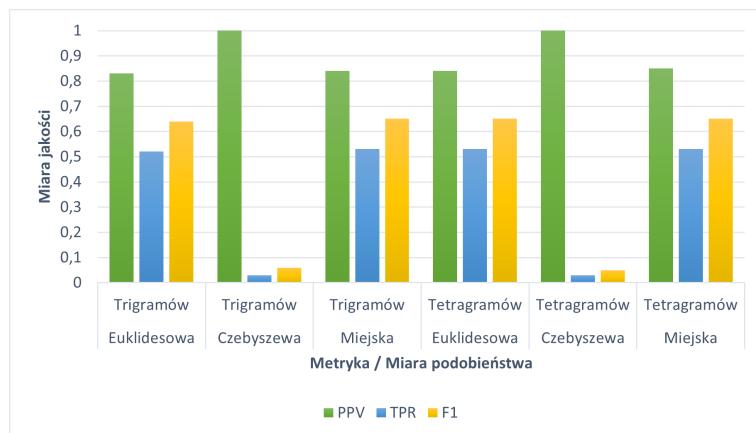
Rysunek 57. Średni wynik *Precision* dla różnych metryk i miar podobieństwa



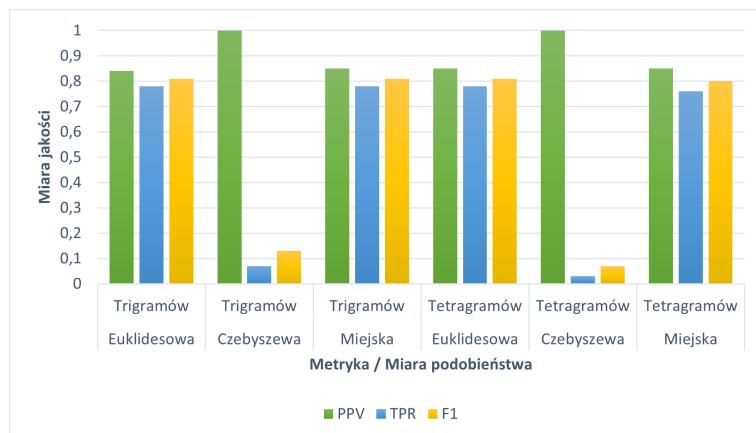
Rysunek 58. Średni wynik *Recall* dla różnych metryk i miar podobieństwa



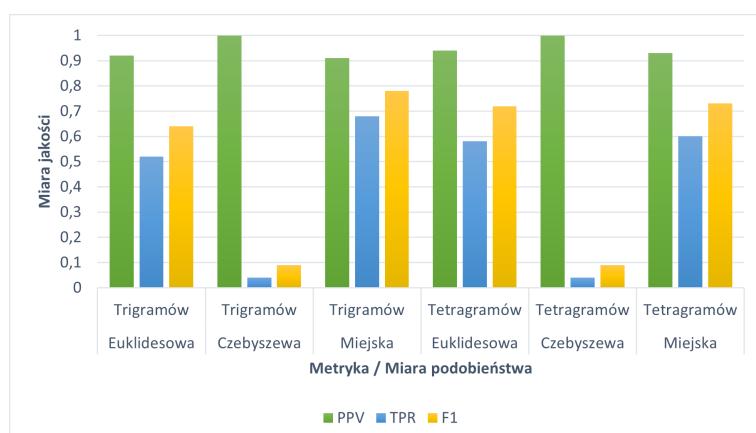
Rysunek 59. Średni wynik *F1* dla różnych metryk i miar podobieństwa



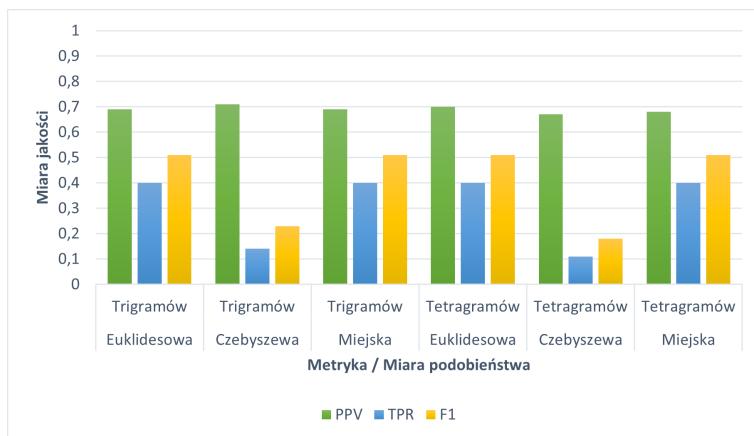
Rysunek 60. Wyniki miar jakości dla klasy *canada* dla różnych metryk i miar podobieństwa



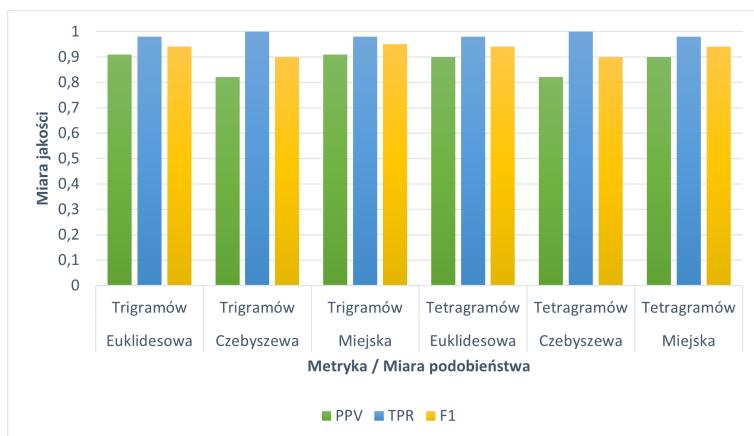
Rysunek 61. Wyniki miar jakości dla klasy *france* dla różnych metryk i miar podobieństwa



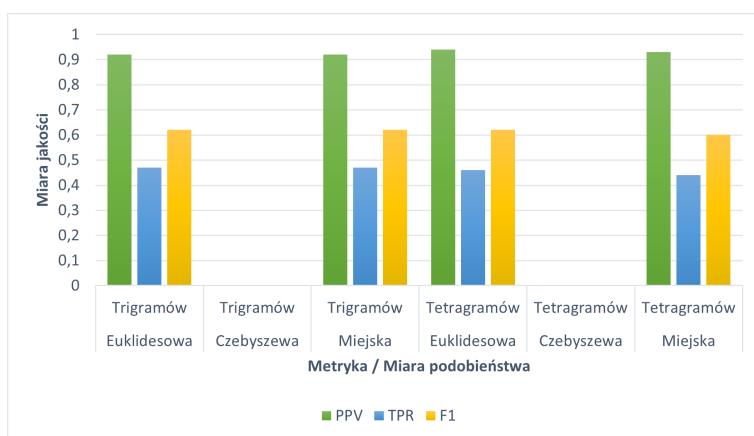
Rysunek 62. Wyniki miar jakości dla klasy *japan* dla różnych metryk i miar podobieństwa



Rysunek 63. Wyniki miar jakości dla klasy *uk* dla różnych metryk i miar podobieństwa



Rysunek 64. Wyniki miar jakości dla klasy *usa* dla różnych metryk i miar podobieństwa



Rysunek 65. Wyniki miar jakości dla klasy *west-germany* dla różnych metryk i miar podobieństwa

Tabela 77. Różne metryki i miary podobieństwa, eksperyment 1 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	121	0	1	6	17	0
France	0	46	1	3	2	3
Japan	0	0	121	5	6	0
UK	5	2	2	85	28	2
USA	108	11	53	109	3296	47
West Germany	0	0	1	0	3	46

Tabela 78. Różne metryki i miary podobieństwa, eksperyment 2 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	7	0	0	0	0	0
France	0	4	0	0	0	0
Japan	0	0	8	0	0	0
UK	1	0	0	29	10	1
USA	226	55	171	179	3342	97
West Germany	0	0	0	0	0	0

Tabela 79. Różne metryki i miary podobieństwa, eksperyment 3 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	124	0	1	6	17	0
France	0	46	1	3	1	3
Japan	0	1	122	5	6	0
UK	4	2	2	83	27	2
USA	106	10	52	111	3298	47
West Germany	0	0	1	0	3	46

Tabela 80. Różne metryki i miary podobieństwa, eksperyment 4 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	125	0	1	5	17	0
France	0	46	1	3	1	3
Japan	0	0	104	2	5	0
UK	4	2	2	83	26	2
USA	105	11	70	115	3301	48
West Germany	0	0	1	0	2	45

Tabela 81. Różne metryki i miary podobieństwa, eksperyment 5 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	125	0	1	5	17	0
France	0	46	1	3	1	3
Japan	0	0	104	2	5	0
UK	4	2	2	83	26	2
USA	105	11	70	115	3301	48
West Germany	0	0	1	0	2	45

Tabela 82. Różne metryki i miary podobieństwa, eksperyment 6 – tablica pomyłek

	Canada	France	Japan	UK	USA	West Germany
Canada	124	0	1	5	16	0
France	0	45	1	3	1	3
Japan	0	1	107	2	5	0
UK	4	3	2	84	28	2
USA	106	10	67	114	3300	50
West Germany	0	0	1	0	2	43

Metryka *euklidesowa* i metryka *miejska* dają zbliżone wyniki. Z kolei metryka *Czebyszewa* znacznie odstaje od pozostałych dwóch. Jej wynik jest zbliżony do wyniku dla podzbioru cech K (5.3) otrzymanego w trakcie badania wpływu cech na jakość klasyfikacji (69). Metryka *Czebyszewa* nie sprawdza się dla cech tekstowych, gdyż brak podobieństwa choćby jednej cechy tekstuowej da zawsze wartość 1, czyli maksymalną jaką w przypadku tego zadania możemy uzyskać.

Metoda *trigramów* i *tetragramów* dają zbliżone rezultaty. Oznacza to, iż wybór miary podobieństwa tekstu nie ma wpływu na jakość klasyfikacji.

## 6.5. Wnioski

Wartości miar jakości klasyfikacji z powyższych eksperymentów pozwalają na ocenę przeprowadzonego badania. Wartości miary *Accuracy* wynoszą około 90%, co można uznać za dobry wynik. Wyniki F1 i Recall dla klasy *usa* są jeszcze wyższe i oscylują w granicach 95% w przypadku F1 i 98% w przypadku Recall. Dla klas mniej licznie reprezentowanych w zbiorze testowym wartości miar jakości klasyfikacji osiągają średnio między 60%, a 70%. Wpływ na taką sytuację ma dominująca pozycja klasy *usa* w zbiorze, gdyż można przypuszczać, że często nieprawidłowo będzie również dominująca wśród  $k$  najbliższych sąsiadów, a tym samym zaburzy wynik klasyfikacji dla pozostałych państw. Artykuły, należące do tej klasy, stanowią niecałe 80% wszystkich artykułów. Liczba ta najlepiej pokazuje wpływ poszczególnych parametrów algorytmu na końcową klasyfikację, gdyż przypisując wszystkie artykuły do tej klasy otrzymalibyśmy wartość *Accuracy* zbliżoną do 80%. Dzięki temu, możemy stwierdzić, że dobrane przez nas cechy, w każdym przypadku podnoszą jakość klasyfikacji. Szczególnie duże znaczenie mają cechy tekstowe, dla których rezultaty są bardzo zbliżone jak dla wektorów złożonych

ze wszystkich cech. Badanie można by powtórzyć dla artykułów należących jedynie do klas *canada*, *france*, *japan*, *uk*, *west-germany*.

Dla zbioru, w którym żadna klasa nie miałaby aż tak dominującej pozycji lepiej można by zbadać wpływ konkretnych cech na jakość klasyfikacji.

## Spis rysunków

1. Diagram pakietów . . . . .	21
2. Diagram klas - pakiet <i>extractor</i> . . . . .	21
3. Diagram klas - pakiet <i>data</i> . . . . .	23
4. Diagram klas - pakiet <i>exception</i> . . . . .	24
5. Diagram klas - pakiet <i>algorithm</i> . . . . .	25
6. Diagram klas - pakiet <i>metrics</i> . . . . .	25
7. Diagram klas - pakiet <i>similarity</i> . . . . .	26
8. Diagram klas - pakiet <i>view</i> . . . . .	27
9. Interfejs użytkownika - widok startowy . . . . .	28
10. Interfejs użytkownika - okno zapisu do pliku . . . . .	29
11. Interfejs użytkownika - okno wczytania pliku wektorów . . . . .	29
12. Interfejs użytkownika - w trakcie przygotowania wektorów . . . . .	30
13. Interfejs użytkownika - widok po wczytaniu wektorów . . . . .	30
14. Interfejs użytkownika - informacja o błędny parametrze <i>k</i> . . . . .	31
15. Interfejs użytkownika - metryki do wyboru . . . . .	31
16. Interfejs użytkownika - informacja o braku wyboru metryki . . . . .	31
17. Interfejs użytkownika - miary podobieństwa do wyboru . . . . .	32
18. Interfejs użytkownika - informacja o braku wyboru podobieństwa . . . . .	32
19. Interfejs użytkownika - informacja o błędny podziale zbioru . . . . .	33
20. Interfejs użytkownika - w trakcie klasyfikacji . . . . .	33
21. Interfejs użytkownika - wyświetlanie wyników . . . . .	33
22. Wykres przedstawiający wyniki eksperymentów 1–10 dla różnych wartości parametru <i>k</i> , ograniczony zbiór artykułów . . . . .	37
23. Wykres przedstawiający wyniki eksperymentów 1–6 dla różnych podziałów zbioru, ograniczony zbiór artykułów . . . . .	44
24. Wykres przedstawiający wyniki eksperymentów 1–5 dla różnych cech, ograniczony zbiór artykułów . . . . .	47
25. Wykres przedstawiający wyniki eksperymentów 1–6 dla metryk i miar podobieństwa, ograniczony zbiór artykułów . . . . .	51
26. Wyniki <i>Accuracy</i> dla różnych wartości parametru <i>k</i> . . . . .	55
27. Średni wynik <i>Precision</i> dla różnych wartości parametru <i>k</i> . . . . .	56
28. Średni wynik <i>Recall</i> dla różnych wartości parametru <i>k</i> . . . . .	56
29. Średni wynik <i>F1</i> dla różnych wartości parametru <i>k</i> . . . . .	56
30. Wyniki miar jakości dla klasy <i>canada</i> dla różnych wartości parametru <i>k</i> . . . . .	57
31. Wyniki miar jakości dla klasy <i>france</i> dla różnych wartości parametru <i>k</i> . . . . .	57
32. Wyniki miar jakości dla klasy <i>japan</i> dla różnych wartości parametru <i>k</i> . . . . .	57
33. Wyniki miar jakości dla klasy <i>uk</i> dla różnych wartości parametru <i>k</i> . . . . .	58
34. Wyniki miar jakości dla klasy <i>usa</i> dla różnych wartości parametru <i>k</i> . . . . .	58
35. Wyniki miar jakości dla klasy <i>west-germany</i> dla różnych wartości parametru <i>k</i> . . . . .	58
36. Wyniki <i>Accuracy</i> dla różnych podziałów zbioru . . . . .	64
37. Średni wynik <i>Precision</i> dla różnych podziałów zbioru . . . . .	65
38. Średni wynik <i>Recall</i> dla różnych podziałów zbioru . . . . .	65

39. Średni wynik <i>F1</i> dla różnych podziałów zbioru . . . . .	65
40. Wyniki miar jakości dla klasy <i>canada</i> dla różnych podziałów zbioru . . . . .	66
41. Wyniki miar jakości dla klasy <i>france</i> dla różnych podziałów zbioru . . . . .	66
42. Wyniki miar jakości dla klasy <i>japan</i> dla różnych podziałów zbioru . . . . .	66
43. Wyniki miar jakości dla klasy <i>uk</i> dla różnych podziałów zbioru . . . . .	67
44. Wyniki miar jakości dla klasy <i>usa</i> dla różnych podziałów zbioru . . . . .	67
45. Wyniki miar jakości dla klasy <i>west-germany</i> dla różnych podziałów zbioru . . . . .	67
46. Wyniki <i>Accuracy</i> dla różnych wektorów cech . . . . .	70
47. Średni wynik <i>Precision</i> dla różnych wektorów cech . . . . .	71
48. Średni wynik <i>Recall</i> dla różnych wektorów cech . . . . .	71
49. Średni wynik <i>F1</i> dla różnych wektorów cech . . . . .	71
50. Wyniki miar jakości dla klasy <i>canada</i> dla różnych wektorów cech . . . . .	72
51. Wyniki miar jakości dla klasy <i>france</i> dla różnych wektorów cech . . . . .	72
52. Wyniki miar jakości dla klasy <i>japan</i> dla różnych wektorów cech . . . . .	72
53. Wyniki miar jakości dla klasy <i>uk</i> dla różnych wektorów cech . . . . .	73
54. Wyniki miar jakości dla klasy <i>usa</i> dla różnych wektorów cech . . . . .	73
55. Wyniki miar jakości dla klasy <i>west-germany</i> dla różnych wektorów cech . . . . .	73
56. Wyniki <i>Accuracy</i> dla różnych metryk i miar podobieństwa . . . . .	77
57. Średni wynik <i>Precision</i> dla różnych metryk i miar podobieństwa . . . . .	78
58. Średni wynik <i>Recall</i> dla różnych metryk i miar podobieństwa . . . . .	78
59. Średni wynik <i>F1</i> dla różnych metryk i miar podobieństwa . . . . .	78
60. Wyniki miar jakości dla klasy <i>canada</i> dla różnych metryk i miar podobieństwa . . . . .	79
61. Wyniki miar jakości dla klasy <i>france</i> dla różnych metryk i miar podobieństwa . . . . .	79
62. Wyniki miar jakości dla klasy <i>japan</i> dla różnych metryk i miar podobieństwa . . . . .	79
63. Wyniki miar jakości dla klasy <i>uk</i> dla różnych metryk i miar podobieństwa . . . . .	80
64. Wyniki miar jakości dla klasy <i>usa</i> dla różnych metryk i miar podobieństwa . . . . .	80
65. Wyniki miar jakości dla klasy <i>west-germany</i> dla różnych metryk i miar podobieństwa . . . . .	80

## Spis tabel

1. Liczebność danej klasy w ograniczonym zbiorze wykorzystanym w eksperymetach . . . . .	34
2. Różne wartości parametru $k$ – eksperyment 1–5, ograniczony zbiór artykułów . . . . .	35
3. Różne wartości parametru $k$ – eksperyment 6–10, ograniczony zbiór artykułów . . . . .	36
4. Różne wartości parametru $k$ , eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek . . . . .	37
5. Różne wartości parametru $k$ , eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek . . . . .	37
6. Różne wartości parametru $k$ , eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek . . . . .	38
7. Różne wartości parametru $k$ , eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek . . . . .	38
8. Różne wartości parametru $k$ , eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek . . . . .	38

9. Różne wartości parametru $k$ , eksperyment 6, ograniczony zbiór artykułów – tablica pomyłek . . . . .	38
10. Różne wartości parametru $k$ , eksperyment 7, ograniczony zbiór artykułów – tablica pomyłek . . . . .	39
11. Różne wartości parametru $k$ , eksperyment 8, ograniczony zbiór artykułów – tablica pomyłek . . . . .	39
12. Różne wartości parametru $k$ , eksperyment 9, ograniczony zbiór artykułów – tablica pomyłek . . . . .	39
13. Różne wartości parametru $k$ , eksperyment 10, ograniczony zbiór artykułów – tablica pomyłek . . . . .	39
14. Liczebność artykułów w zbiorze. Podział 80%/20%, ograniczony zbiór artykułów . . . . .	40
15. Liczebność artykułów w zbiorze. Podział 70%/30%, ograniczony zbiór artykułów . . . . .	40
16. Liczebność artykułów w zbiorze. Podział 60%/40%, ograniczony zbiór artykułów . . . . .	41
17. Liczebność artykułów w zbiorze. Podział 50%/50%, ograniczony zbiór artykułów . . . . .	41
18. Liczebność artykułów w zbiorze. Podział 40%/60%, ograniczony zbiór artykułów . . . . .	41
19. Liczebność artykułów w zbiorze. Podział 30%/70%, ograniczony zbiór artykułów . . . . .	42
20. Różne podziały zbioru – eksperyment 1–3, ograniczony zbiór artykułów . . . . .	42
21. Różne podziały zbioru – eksperyment 4–6, ograniczony zbiór artykułów . . . . .	43
22. Różne podziały zbioru, eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek . . . . .	44
23. Różne podziały zbioru, eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek . . . . .	44
24. Różne podziały zbioru, eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek . . . . .	45
25. Różne podziały zbioru, eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek . . . . .	45
26. Różne podziały zbioru, eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek . . . . .	45
27. Różne podziały zbioru, eksperyment 6, ograniczony zbiór artykułów – tablica pomyłek . . . . .	45
28. Różne wektory cech – eksperyment 1–5, ograniczony zbiór artykułów . . . . .	46
29. Różne wektory cech, eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek . . . . .	47
30. Różne wektory cech, eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek . . . . .	47
31. Różne wektory cech, eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek . . . . .	48
32. Różne wektory cech, eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek . . . . .	48
33. Różne wektory cech, eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek . . . . .	48
34. Różne metryki i miary podobieństwa – eksperyment 1–3, ograniczony zbiór artykułów . . . . .	49
35. Różne metryki i miary podobieństwa – eksperyment 4–6, ograniczony zbiór artykułów . . . . .	50

36. Różne metryki i miary podobieństwa, eksperyment 1, ograniczony zbiór artykułów – tablica pomyłek	51
37. Różne metryki i miary podobieństwa, eksperyment 2, ograniczony zbiór artykułów – tablica pomyłek	51
38. Różne metryki i miary podobieństwa, eksperyment 3, ograniczony zbiór artykułów – tablica pomyłek	52
39. Różne metryki i miary podobieństwa, eksperyment 4, ograniczony zbiór artykułów – tablica pomyłek	52
40. Różne metryki i miary podobieństwa, eksperyment 5, ograniczony zbiór artykułów – tablica pomyłek	52
41. Różne metryki i miary podobieństwa, eksperyment 6, ograniczony zbiór artykułów – tablica pomyłek	52
42. Liczebność artykułów dla poszczególnych klas.	53
43. Różne wartości parametru $k$ – eksperyment 1–5	54
44. Różne wartości parametru $k$ – eksperyment 6–10	55
45. Różne wartości parametru $k$ , eksperyment 1 – tablica pomyłek	59
46. Różne wartości parametru $k$ , eksperyment 2 – tablica pomyłek	59
47. Różne wartości parametru $k$ , eksperyment 3 – tablica pomyłek	59
48. Różne wartości parametru $k$ , eksperyment 4 – tablica pomyłek	59
49. Różne wartości parametru $k$ , eksperyment 5 – tablica pomyłek	60
50. Różne wartości parametru $k$ , eksperyment 6 – tablica pomyłek	60
51. Różne wartości parametru $k$ , eksperyment 7 – tablica pomyłek	60
52. Różne wartości parametru $k$ , eksperyment 8 – tablica pomyłek	60
53. Różne wartości parametru $k$ , eksperyment 9 – tablica pomyłek	61
54. Różne wartości parametru $k$ , eksperyment 10 – tablica pomyłek	61
55. Liczebność artykułów w zbiorze. Podział 80%/20%	61
56. Liczebność artykułów w zbiorze. Podział 70%/30%	62
57. Liczebność artykułów w zbiorze. Podział 60%/40%	62
58. Liczebność artykułów w zbiorze. Podział 50%/50%	62
59. Liczebność artykułów w zbiorze. Podział 40%/60%	62
60. Liczebność artykułów w zbiorze. Podział 30%/70%	63
61. Różne podziały zbioru – eksperyment 1–3	63
62. Różne podziały zbioru – eksperyment 4–6	64
63. Różne podziały zbioru, eksperyment 1 – tablica pomyłek	68
64. Różne podziały zbioru, eksperyment 2 – tablica pomyłek	68
65. Różne podziały zbioru, eksperyment 3 – tablica pomyłek	68
66. Różne podziały zbioru, eksperyment 4 – tablica pomyłek	68
67. Różne podziały zbioru, eksperyment 5 – tablica pomyłek	69
68. Różne podziały zbioru, eksperyment 6 – tablica pomyłek	69
69. Różne wektory cech – eksperyment 1–5	70
70. Różne wektory cech, eksperyment 1 – tablica pomyłek	74
71. Różne wektory cech, eksperyment 2 – tablica pomyłek	74
72. Różne wektory cech, eksperyment 3 – tablica pomyłek	74
73. Różne wektory cech, eksperyment 4 – tablica pomyłek	74
74. Różne wektory cech, eksperyment 5 – tablica pomyłek	75
75. Różne metryki i miary podobieństwa – eksperyment 1–3	76
76. Różne metryki i miary podobieństwa – eksperyment 4–6	77
77. Różne metryki i miary podobieństwa, eksperyment 1 – tablica pomyłek	81
78. Różne metryki i miary podobieństwa, eksperyment 2 – tablica pomyłek	81
79. Różne metryki i miary podobieństwa, eksperyment 3 – tablica pomyłek	81
80. Różne metryki i miary podobieństwa, eksperyment 4 – tablica pomyłek	81

81. Różne metryki i miary podobieństwa, eksperyment 5 – tablica pomylek	82
82. Różne metryki i miary podobieństwa, eksperyment 6 – tablica pomylek	82

## Literatura

- [1] R. Tadeusiewicz: Rozpoznawanie obrazów, PWN, Warszawa, 1991. Rozdział 4.3
- [2] A. Niewiadomski, Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2008.
- [3] ToniCifre, (2016), all-countries-and-cities-json, data dostępu 11.03.2023, dostępne w Github <https://github.com/ToniCifre/all-countries-and-cities-json>
- [4] "Podział administracyjny Stanów Zjednoczonych." W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 30.11.2021, data dostępu 11.03.2023, [https://pl.wikipedia.org/wiki/Podzia%C5%82\\_administracyjny\\_Stan%C3%B3w\\_Zjednoczonych](https://pl.wikipedia.org/wiki/Podzia%C5%82_administracyjny_Stan%C3%B3w_Zjednoczonych)
- [5] "List of regions of the United States", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 12.03.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/List\\_of\\_regions\\_of\\_the\\_United\\_States](https://en.wikipedia.org/wiki/List_of_regions_of_the_United_States)
- [6] "List of English people", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 12.03.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/List\\_of\\_English\\_people](https://en.wikipedia.org/wiki/List_of_English_people)
- [7] "List of French people", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 12.03.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/List\\_of\\_French\\_people](https://en.wikipedia.org/wiki/List_of_French_people)
- [8] "List of Japanese people", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 12.03.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/List\\_of\\_Japanese\\_people](https://en.wikipedia.org/wiki/List_of_Japanese_people)
- [9] "Lists of Americans", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 12.01.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/Lists\\_of\\_Americans](https://en.wikipedia.org/wiki/Lists_of_Americans)
- [10] "Lists of Canadians", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 6.03.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/Lists\\_of\\_Canadians](https://en.wikipedia.org/wiki/Lists_of_Canadians)
- [11] "List of Germans", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 11.03.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/List\\_of\\_Germans](https://en.wikipedia.org/wiki/List_of_Germans)
- [12] "Lists of companies by country", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 3.05.2020, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/Category:Lists\\_of\\_companies\\_by\\_country](https://en.wikipedia.org/wiki/Category:Lists_of_companies_by_country)
- [13] Reuters Ltd., Reuters-21578 Text Categorization Collection Data Set, W: UCI Machine Learning Repository, data ostatniej modyfikacji: 1987, data dostępu 11.03.2023, <http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection>
- [14] "Confusion matrix", W: Wikipedia: wolna encyklopedia, data ostatniej modyfikacji: 7.04.2023, data dostępu 12.03.2023, [https://en.wikipedia.org/wiki/Confusion\\_matrix](https://en.wikipedia.org/wiki/Confusion_matrix)
- [15] Komorowski Rafał. Presler Kamil, Plik "dictionary.zip" (archiwum zawiera wszystkie wykorzystywane słowniki)

- [16] Komorowski Rafał. Presler Kamil , "currencies.txt", plik znajduje się w archiwum "dictionary.zip" , format pliku (txt)
- [17] Komorowski Rafał. Presler Kamil , "countries.txt", plik znajduje się w archiwum "dictionary.zip" , format pliku (txt)
- [18] Komorowski Rafał. Presler Kamil , "dictionary.txt", plik znajduje się w archiwum "dictionary.zip" , format pliku (txt)
- [19] Komorowski Rafał. Presler Kamil , "lands.txt", plik znajduje się w archiwum "dictionary.zip" , format pliku (txt)
- [20] Komorowski Rafał. Presler Kamil , "people.txt", plik znajduje się w archiwum "dictionary.zip" , format pliku (txt)
- [21] Komorowski Rafał. Presler Kamil , "companies.txt", plik znajduje się w archiwum "dictionary.zip" , format pliku (txt)
- [22] Charu C. Aggarwal, Machine Learning for Text, Springer, Yorktown Heights, NY, USA, 2018, Rozdział 2.5, strona 26,
- [23] Uday Kamath, John Liu, James Whitaker, Deep Learning for NLP and Speech Recognition, Springer Nature Switzerland AG 2019, Rozdział 3.3.3.
- [24] Scott Krig, Computer Vision Metrics, Springer, Apress Berkeley, CA, 2014, Rozdział 4, strony 139-145.
- [25] Oracle, 2023, data dostępu 2.04.2023, <https://www.oracle.com/pl/java/technologies/downloads/>
- [26] The Apache Software Foundation, 2023 data dostępu 2.04.2023, <https://maven.apache.org/download.cgi>