Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2018/2019**

Prowadzący: dr inż. Marcin Kacprowicz wtorek, 16:00

Data oddania: _____ Ocena: _____

Damian Salata 210311
Aleksandra Kowalczyk 210228

Zadanie 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja

1. Cel

Celem zadania było napisanie uniwersalnej aplikacji pozwalającej na klasyfikację metodą k-NN niezależnie od typu klasyfikowanych obiektów. Implementacja powinna obejmować dwa istniejące sposoby ekstrakcji wektorów cech, trzy metryki (euklidesowa, uliczna, Czebyszewa) oraz dwie istniejące miary podobieństwa. Ponadto, naszym zadaniem było porównanie wyników klasyfikacji metody k-NN dla wyżej wymienionych miar podobieństwa i różnych wartości parametru k.

2. Wprowadzenie

2.1. Klasyfikacja

Klasyfikacja to technika pozwalająca określić do jakiej z dostępnych klas należy rozważany obiekt [1].

2.2. Algorytm k-NN

Algorytm k najbliższych sąsiadów (ang. k nearest neighbours) jest jednym z algorytmów używanych do klasyfikacji obiektów. Algorytm ten wykorzystuje podobieństwo wektorów cech do wyznaczenia przewidywanej przynależności do klasy dla nowych, wprowadzanych obiektów (w przypadku realizowanego przez nas zadania są to artykuły). Oznacza to, że nowemu obiektowi przypisana zostaje klasa wyznaczona na podstawie odległości jego wektora cech od k najbliższych sąsiednich wektorów [2]. Odległość ta wyliczana jest na podstawie wybranej metryki (w naszym przypadku są to metryki: euklidesowa, uliczna, Czebyszewa).

2.3. Metryki odległości pomiędzy wektorami

Metryka jest funkcją definiującą odległość pomiędzy parą wektorów [3]. Metryki zaimplementowane w programie przedstawione są poniżej [4].

2.3.1. Metryka euklidesowa

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

2.3.2. Metryka uliczna

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

2.3.3. Metryka Czebyszewa

$$d(x, y) = \max_i(|x_i - y_i|) \tag{3}$$

Oznaczenia dla wszystkich powyższych wzorów:
x_i - i-ta waga pierwszego wektora,
y_i - i-ta waga drugiego wektora,
d - odległość pomiędzy wektorami

2.4. Miary podobieństwa tekstów [4]

Miary te służą określeniu stopnia, w jakim poszczególne słowa porównywanych tekstów są do siebie podobne.

2.4.1. Miara binarna

$$sim_n(s_1, s_2) = \begin{cases} 1 & : s_1 = s_2 \\ 0 & : s_1 \neq s_2 \end{cases} \tag{4}$$

s_1 - pierwsze porównywane słowo,
s_2 - drugie porównywane słowo

2.4.2. Miara n-gram

Metoda ta wykorzystywana jest w naszej aplikacji dla $n = 3$.

$$sim_n(s_1, s_2) = \frac{1}{N - n + 1} \sum_{i=1}^{N-n+1} h(i) \tag{5}$$

$h(i) = 1$ jeśli n -elementowy podciąg zaczynający się od i - tej pozycji w s_1 występuje przynajmniej raz w s_2 (w przeciwnym przypadku $h(i) = 0$),
$N - n + 1$ - ilość możliwych n - elementowych podciągów w s_1,
s_1 - pierwsze porównywane słowo,
s_2 - drugie porównywane słowo

2.4.3. Miara Levenshteina

$$lev_{a,b}(i, j) = \begin{cases} \max(i, j) & : \min(i, j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1, j) + 1 \\ lev_{a,b}(i, j-1) + 1 \\ lev_{a,b}(i-1, j-1) + 1_{(a_i \neq b_j)} \end{cases} & : \min(i, j) \neq 0 \end{cases} \tag{6}$$

$1_{(a_i \neq b_j)}$ - funkcja przyjmująca wartość 0, jeśli $a_i = b_j$, i 1, jeśli $a_i \neq b_j$
$lev_{a,b}(i, j)$ - odległość pomiędzy pierwszymi i literami a i pierwszymi j literami b

2.5. Ekstrakcja cech

W naszym programie wydobywamy słowa kluczowe, które potem są wykorzystywane do wyznaczenia wektora cech. W tym celu używamy Term Frequency oraz Document Frequency.

- Term Frequency - oznacza jak często dane słowo pojawia się w tekście.
- Document Frequency - oznacza w ilu dokumentach ze wszystkich możliwych wystąpiło dane słowo.

W napisanej przez nas aplikacji w pierwszej kolejności określamy częstotliwość występowania słowa we wszystkich artykułach danej klasy. Następnie występuje sortowanie niniejszych słów, aby wyłonić z nich te występujące najczęściej. Posiadając wektory najczęściej występujących słów dla każdej klasy usuwamy z każdego wektora słowa występujące najczęściej w innych klasach, przez co zapewniamy że otrzymamy słowa najczęściej występujące tylko w rozważanej klasie.

Wektor cech w naszym programie składa się z:

- ilość słów kluczowych występujących w danym artykule
  rodzaj danych - liczby naturalne
  metoda w programie - CountOfKeyWordsExtractor
  parametry metody - allWords=true
- ilość słów kluczowych występujących w pierwszych 30% danego artykułu
  rodzaj danych - liczby naturalne
  metoda w programie - CountOfKeyWordsExtractor
  parametry metody - allWords=false
- suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
  rodzaj danych - liczby rzeczywiste
  metoda w programie - SumOfSimilarityArticle_KeyWords
  parametry metody - allWords=true, similarityMethod=Binary
- suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
  rodzaj danych - liczby rzeczywiste
  metoda w programie - SumOfSimilarityArticle_KeyWords
  parametry metody - allWords=false, similarityMethod=Binary
- suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
  rodzaj danych - liczby rzeczywiste
  metoda w programie - SumOfSimilarityArticle_KeyWords
  parametry metody - allWords=true, similarityMethod=N-Gram
- suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
  rodzaj danych - liczby rzeczywiste
  metoda w programie - SumOfSimilarityArticle_KeyWords
  parametry metody - allWords=false, similarityMethod=N-Gram
- suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów z artykułu
  rodzaj danych - liczby rzeczywiste

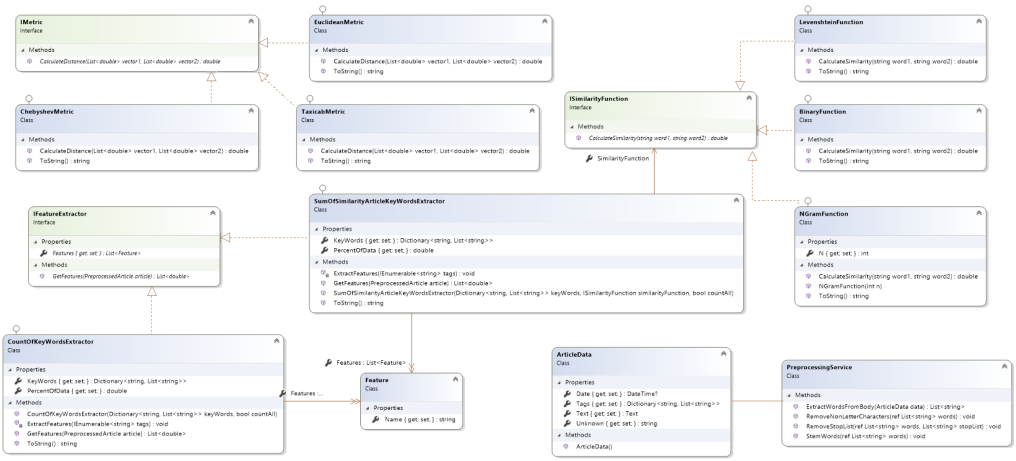
metoda w programie - SumOfSimilarityArticle_KeyWords
parametry metody - allWords=true, similarityMethod=Levenshtein

- suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
rodzaj danych - liczby naturalne
metoda w programie - SumOfSimilarityArticle_KeyWords
parametry metody - allWords=true, similarityMethod=Levenshtein

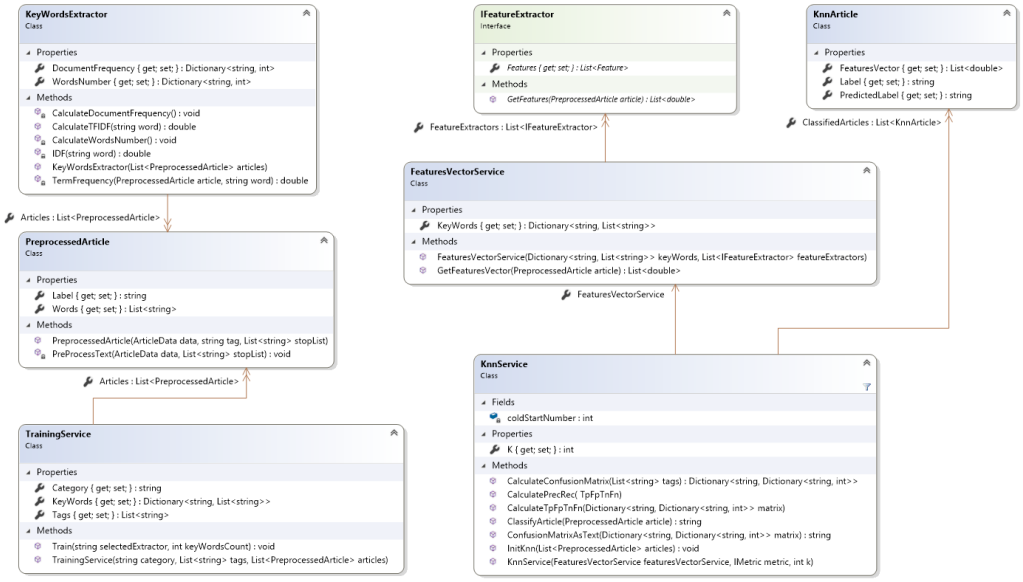
Liczba wartości w każdej z cech wektora jest równa liczbie możliwych tagów. Każda z cech jest znormalizowana do wartości z przedziału [0,1].

3. Opis implementacji

Program do klasyfikacji metodą k-NN został napisany w języku C#. Poniżej znajuje się opis poszczególnych klas.



Rys. 1. Diagram klas.



Rys. 2. Diagram klas.

3.1. Odczyt danych

Odczyt danych z pliku i przetworzenie ich do postaci obiektowej zostało zaimplementowane w klasie SGMParser. Została tutaj użyta biblioterka HtmlAgilityPack pomagająca odczytać dokumenty w postaci SGML/HTML/XML. Każdy dokument zamieniony jest na obiekt klasy ArticleData.

3.2. Przetwarzanie danych (Preprocessing)

Dane z postaci ArticleData zamieniane są na PreprocessedArticle, preprocessing został zaimplementowany w klasie PreprocessingService, w którym znajdują się metody do usuwania znaków nieliterowych, usuwania słów ze stop listy oraz stemmizacja. Do stemmizacji została użyta biblioteka Iveonik.

3.3. Uczenie (wyznaczenie słów kluczowych)

Uczenie czyli wyznaczenie słów kluczowych jest realizowane przez klasę TrainingService oraz KeyWordsExtractor. W klasie KeyWordsExtractor wyznaczane są tf i df dla słów z artykułów dla danej etykiety. W klasie TrainingService znajduje się algorytm, który wyznacza słowa kluczowe na podstawie tf lub df dla każdej etykiety.

3.4. Miary podobieństwa słów

Miara podobieństwa słów realizowane są przez klasy BinaryFunction, LevenshteinFunction oraz NGramFunction, które implementują interfejs ISimilarityFunction

3.5. Tworzenie wektorów cech

Wektory cech są tworzone w klasie FeaturesVectorService. Klasa ta posiada listę ekstraktorów cech czyli IFeatureExtractor. Interfejs IFeatureExtrac-

tor implementowany jest przez klasy CountOfKeyWordsExtractor oraz SumOfSimilarityArticle_KeyWords. Klasa SumOfSimilarityArticle_KeyWords jest zależna od miar podobieństwa słów czyli ISimilarityFunction.

3.6. Metryki

Metryki realizowane są przez klasy ChebyshevMetric, EuclideanMetric, TaxicabMetric, które implementują interfejs IMetric.

3.7. Klasyfikacja metodą k-NN

Za klasyfikację tekstów odpowiedzialna jest klasa KnnService. Wykorzystuje ona FeaturesVectorService do wydobycia z pojedynczego artykułu wektoru cech. Następnie wektory porównywane są ze sobą na podstawie odpowiednich metryk (IMetric). Klasa ta jest również odpowiedzialna za obliczanie wyników tj. Macierz pomyłek, Precision, Recall i procent poprawnie sklasyfikowanych.

4. Materiały i metody

Badania były przeprowadzone na zbiorze artykułów Reuters-21578 Text Categorization Collection Data Set [5] oraz zbiorze przygotowanym przez nas zawierającym 102 artykuły o tematyce sportowej.

Pierwszy zestaw składa się z artykułów o kategorii PLACES zawierające tagi [west-germany, usa, france, uk, canada, japan].

Drugi zestaw danych składa się z artykułów o kategorii TOPICS, zawierające jeden z tagów [earn, acq, money-fx, grain].

Trzeci zestaw danych składa się ze 102 artykułów o kategorii LEAGUE, zawierające jeden z tagów [PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga].

Wszystkie badania przeprowadzone zostały na zbiorze danych podzielonym na 60% danych treningowych i 40% danych testowych.

4.1. Przeprowadzone badania

4.1.1. Porównanie metody wyboru słów kluczowych oraz ich ilości

Cel testu: Celem testu jest porównanie metod wyboru słów kluczowych dla różnej ilości słów kluczowych.

Wartości stałe:
— Metryka - Taxicab
— Parametr K - 15 (dla zestawu FOOTBALL K=4)
— Ilość danych do cold startu - 20 (dla zestawu FOOTBALL cold start=2)
— Cechy:
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu

— suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
— suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
— suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

Testowane wartości:
Zestaw danych:
1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Wybór słów kluczowych:
   a) Document Frequency
      Ilość słów:
      — 5
      — 15
      — 30
   b) Term Frequency
      Ilość słów:
      — 5
      — 15
      — 30
2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Wybór słów kluczowych:
   a) Document Frequency
      Ilość słów:
      — 5
      — 15
      — 30
   b) Term Frequency
      Ilość słów:
      — 5
      — 15
      — 30
3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Wybór słów kluczowych:
   a) Document Frequency
      Ilość słów:
      — 5
      — 15
      — 30
   b) Term Frequency
      Ilość słów:
      — 5
      — 15

— 30

4.1.2. Porównanie metryk odległości

Cel testu: Celem testu jest porównanie wyników dla różnych metryk odległości: Euklides, Czebyszew oraz Uliczna.

Wartości stałe:
— Wybór słów kluczowych - Term Frequency
— Ilość słów kluczowych - 15
— Parametr K - 15 (dla zestawu FOOTBALL K=4)
— Ilość danych do cold startu - 20 (dla zestawu FOOTBALL cold start=2)
— Cechy:
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

Testowane wartości:
Zestaw danych:
1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Metryka:
   a) Euklides
   b) Czebyszew
   c) Uliczna
2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Metryka:
   a) Euklides
   b) Czebyszew
   c) Uliczna
3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Metryka:
   a) Euklides
   b) Czebyszew
   c) Uliczna

4.1.3. Dokładność a wybór parametru k

Cel testu: Celem testu jest porównanie wyników dla różnych wartości parametru k.

Wartości stałe:

— Wybór słów kluczowych - Term Frequency
— Ilość słów kluczowych - 15
— Metryka - Uliczna
— Ilość danych do cold startu - 20 (dla zestawu FOOTBALL cold start=2)
— Cechy:
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

Testowane wartości:
Zestaw danych:
1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Wartość parametru k:
   — 1
   — 5
   — 10
   — 15
   — 30
   — 50
2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Wartość parametru k:
   — 1
   — 5
   — 10
   — 15
   — 30
   — 50
3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Wartość parametru k:
   — 1
   — 2
   — 3
   — 4

4.1.4. Dokładność a wybór ilość danych do cold startu

Cel testu: Celem testu jest porównanie wyników dla różnej danych wybranych do cold startu.

Wartości stałe:

— Wybór słów kluczowych - Term Frequency
— Ilość słów kluczowych - 15
— Metryka - Uliczna
— Parametr K - 15 (dla zestawu FOOTBALL k=4)
— Cechy:
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

Testowane wartości:
Zestaw danych:
1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Ilość danych do cold startu:
   — 5
   — 10
   — 15
   — 30
   — 50
   — 100
2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Ilość danych do cold startu:
   — 5
   — 10
   — 15
   — 30
   — 50
   — 100
3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Ilość danych do cold startu:
   — 1
   — 2
   — 3
   — 4

4.1.5. Porównanie dokładności dla cech wykorzystujących miary podobieństwa słów

Cel testu: Celem testu jest porównanie dokładności dla cech, które obliczane są za pomocą miar podobieństwa słów - N-Gram, Levenshtein, Binarna.

Wartości stałe:
— Wybór słów kluczowych - Term Frequency
— Ilość słów kluczowych - 15
— Metryka - Uliczna
— Parametr K - 15 (dla zestawu FOOTBALL K=4)
— Ilość danych do cold startu - 20 (dla zestawu FOOTBALL cold start=2)

Testowane wartości:
Zestaw danych:
1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Cechy:
   — suma wartości miary podobieństwa N-Gram słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Binarnej słów z wektora słów kluczowych oraz słów z artykułu
2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Cechy:
   — suma wartości miary podobieństwa N-Gram słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Binarnej słów z wektora słów kluczowych oraz słów z artykułu
3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Cechy:
   — suma wartości miary podobieństwa N-Gram słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Binarnej słów z wektora słów kluczowych oraz słów z artykułu

4.1.6. Porównanie tekstowych i liczbowych metod obliczania cech

Cel testu: Celem testu jest porównanie tekstowych i liczbowych metod obliczania cech. Porównywane cechy to tekstowe: N-Gram, Levenshtein oraz liczbowe: ilość słów kluczowych występujących w artykule, Binarna miara podobieństwa (częstość występowania słowa).

Wartości stałe:
— Wybór słów kluczowych - Term Frequency
— Ilość słów kluczowych - 15

— Metryka - Uliczna
— Parametr K - 15 (dla zestawu FOOTBALL K=4)
— Ilość danych do cold startu - 20 (dla zestawu FOOTBALL cold start=2)

Testowane wartości:
Zestaw danych:
1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Cechy:
   — suma wartości miary podobieństwa N-Gram słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Levenshtein słów z wektora słów kluczowych oraz słów z artykułu
   — ilość słów kluczowych występujących w danym artykule
   — suma wartości miary podobieństwa Binarna słów z wektora słów kluczowych oraz słów z artykułu
2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Cechy:
   — suma wartości miary podobieństwa N-Gram słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Levenshtein słów z wektora słów kluczowych oraz słów z artykułu
   — ilość słów kluczowych występujących w danym artykule
   — suma wartości miary podobieństwa Binarna słów z wektora słów kluczowych oraz słów z artykułu
3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Cechy:
   — suma wartości miary podobieństwa N-Gram słów z wektora słów kluczowych oraz słów z artykułu
   — suma wartości miary podobieństwa Levenshtein słów z wektora słów kluczowych oraz słów z artykułu
   — ilość słów kluczowych występujących w danym artykule
   — suma wartości miary podobieństwa Binarna słów z wektora słów kluczowych oraz słów z artykułu

4.1.7. Porównanie różnych połączeń cech

Cel testu: Celem testu jest porównanie różnych cech połączonych ze sobą.

Wartości stałe:
— Wybór słów kluczowych - Term Frequency
— Ilość słów kluczowych - 15
— Metryka - Uliczna
— Parametr K - 15 (dla zestawu FOOTBALL K=4)
— Ilość danych do cold startu - 20 (dla zestawu FOOTBALL cold start=2)

Testowane wartości:
Zestaw danych:

1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Cechy:
   a) Wektor1
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   b) Wektor2
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   c) Wektor3
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   d) Wektor4
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu

— suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Cechy:
   a) Wektor1
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   b) Wektor2
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   c) Wektor3
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   d) Wektor4
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu

— suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Cechy:
   a) Wektor1
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
      — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa Levenshteina słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   b) Wektor2
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   c) Wektor3
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
   d) Wektor4
      — ilość słów kluczowych występujących w danym artykule
      — ilość słów kluczowych występujących w pierwszych 30% danego artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
      — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

— suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
— suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

4.1.8. Porównanie różnego stosunku danych treningowych do testowych

Cel testu: Celem testu jest porównanie wyników dla różnego stosunku danych trenigowych do testowych.

Wartości stałe:
— Wybór słów kluczowych - Term Frequency
— Ilość słów kluczowych - 15
— Metryka - Uliczna
— Parametr K - 15 (dla zestawu FOOTBALL K=4)
— Ilość danych do cold startu - 20 (dla zestawu FOOTBALL cold start=2)
— Cechy:
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości binarnej miary podobieństwa słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów z artykułu
  — suma wartości miary podobieństwa n-gram słów z wektora słów kluczowych oraz słów występujących w pierwszych 30% artykułu

Testowane wartości:
Zestaw danych:
1. Kategoria - PLACES
   Etykiety - USA, UK, Japan, West-Germany, Canada, France
   Procent danych treningowych/testowych:
   a) 30%/70%
   b) 50%/50%
   c) 70%/30%
2. Kategoria - TOPICS
   Etykiety - earn, acq, money-fx, grain
   Procent danych treningowych/testowych:
   a) 30%/70%
   b) 50%/50%
   c) 70%/30%
3. Kategoria - LEAGUE
   Etykiety - PremierLeague, Bundesliga, SerieA, PrimeraDivision, Ligue1, PrimeiraLiga
   Procent danych treningowych/testowych:
   a) 30%/70%
   b) 50%/50%
   c) 70%/30%

5. Wyniki

5.1. Porównanie metody wyboru słów kluczowych oraz ich ilości

	Term Frequency			Document Frequency		
	5	15	30	5	15	30
Overall Accuracy	38,33%	89,31%	87,96%	54,90%	84,69%	74,99%
Time	21,33s	44,86s	70,29s	24,66s	46,69s	70,47s

Tabela 1. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji z kategorii PLACES.

	Term Frequency			Document Frequency		
	5	15	30	5	15	30
Recall USA	32,68%	96,21%	95,12%	52,20%	89,93%	77,16%
Recall UK	46,16%	70,63%	61,19%	55,59%	66,78%	66,43%
Recall Japan	39,06%	38,54%	39,06%	52,60%	46,35%	47,40%
Recall Canada	82,20%	45,79%	45,79%	76,09%	53,54%	66,67%
Recall West-Germany	78,07%	77,19%	78,85%	79,82%	82,16%	77,19%
Recall France	84,06%	85,51%	85,51%	72,46%	84,69%	84,06%
Precision USA	93,98%	92,69%	92,20%	91,34%	92,89%	93,59%
Precision UK	82,50%	82,45%	78,83%	84,57%	81,62%	68,35%
Precision Japan	68,18%	66,67%	50,00%	63,52%	58,55%	52,30%
Precision Canada	7,77%	75,14%	75,56%	9,97%	34,95%	20,00%
Precision West-Germany	86,41%	60,27%	62,07%	80,87%	65,00%	64,23%
Precision France	60,42%	52,68%	47,20%	87,72%	49,12%	43,28%

Tabela 2. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	Term Frequency			Document Frequency		
	5	15	30	5	15	30
Overall Accuracy	89,07%	90,91%	94,5% 1	88,74%	93,47%	93,93%
Time	4,96s	10,76s	19,23s	4,91s	9,02s	18,72s

Tabela 3. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	Term Frequency			Document Frequency		
	5	15	30	5	15	30
Recall earn	86,4%	88,42%	96,09%	91,46%	93,85%	94,93%
Recall acq	97,15%	95,72%	92,76%	87,27%	94,62%	93,3%
Recall money-fx	50,53%	81,05%	88,42%	63,16%	76,84%	85,26%
Recall grain	0%	0%	0%	0%	0%	0%
Precision earn	98,68%	98,95%	97,01%	92,4%	98,26%	97,91%
Precision acq	79,87%	84,5%	93,89%	4,94%	90,64%	92,59%
Precision money-fx	68,57%	71,96%	75%	2,29%	69,52%	70,43%
Precision grain	0%	0%	0%	0%	0%	0%

Tabela 4. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	Term Frequency			Document Frequency		
	5	15	30	5	15	30
Overall Accuracy	93,33%	93,33%	90,00%	93,33%	90,00%	86,67%
Time	0,38s	0,95s	1,70s	0,21s	0,59s	1,15s

Tabela 5. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	Term Frequency			Document Frequency		
	5	15	30	5	15	30
Recall PremiereLeague	100%	100%	100%	66,67%	66,67%	66,67%
Recall BundesLiga	100%	100%	0%	100%	100%	0%
Recall SerieA	100%	50%	50%	100%	100%	100%
Recall PrimeraDivision	87,5%	100%	100%	100%	87,5%	100%
Recall Ligue1	100%	100%	100%	85,71%	100%	85,71%
Recall PrimeiraLiga	88,89%	88,89%	88,89%	100%	88,89%	88,89%
Precision PremiereLeague	75%	75%	60%	100%	50%	66,67%
Precision BundesLiga	50%	100%	0%	100%	100%	0%
Precision SerieA	100%	100%	100%	100%	100%	100%
Precision PrimeraDivision	100%	100%	100%	88,89%	100%	100%
Precision Ligue1	100%	100%	100%	85,71%	87,5%	85,71%
Precision PrimeiraLiga	100%	88,89%	88,89%	100%	100%	80%

Tabela 6. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

5.2. Porównanie metryk odległości

	Euklides	Czebyszew	Uliczna
Overall Accuracy	88,8%	87,81%	89,31%
Time	58,78s	51,80s	40,69s

Tabela 7. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii PLACES.

	Euklides	Czebyszew	Uliczna
Recall USA	95,18%	96,32%	96,21%
Recall UK	72,38%	55,94%	70,63%
Recall Japan	43,23%	36,98%	38,54%
Recall Canada	46,46%	36,03%	45,79%
Recall West-Germany	78,07%	71,93%	77,19%
Recall France	85,51%	79,71%	85,51%
Precision USA	93,06%	90,67%	92,69%
Precision UK	81,50%	82,05%	82,45%
Precision Japan	55,70%	59,66%	66,67%
Precision Canada	73,80%	79,26%	75,14%
Precision West-Germany	57,79%	56,94%	60,27%
Precision France	50,86%	56,70%	52,68%

Tabela 8. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	Euklides	Czebyszew	Uliczna
Overall Accuracy	90,45%	79,15%	90,91%
Time	12,70s	12,22s	11,64s

Tabela 9. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	Euklides	Czebyszew	Uliczna
Recall earn	87,12%	98,63%	88,42%
Recall acq	96,6%	49,73%	95,72%
Recall money-fx	80%	77,89%	81,05%
Recall grain	0%	0%	0%
Precision earn	99,09%	74,56%	98,95%
Precision acq	83,1%	97%	84,5%
Precision money-fx	75,25%	84,09%	71,96%
Precision grain	0%	0%	0%

Tabela 10. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	Euklides	Czebyszew	Uliczna
Overall Accuracy	90%	70%	93,33%
Time	0,93s	0,64s	0,65s

Tabela 11. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	Euklides	Czebyszew	Uliczna
Recall PremierLeague	66,67%	100%	100%
Recall Bundesliga	100%	100%	100%
Recall SerieA	50%	50%	50%
Recall PrimeraDivision	100%	100%	100%
Recall Ligue1	100%	100%	100%
Recall France	88,89%	11,11%	88,89%
Precision PremierLeague	66,67%	27,27%	75%
Precision Bundesliga	100%	100%	100%
Precision SerieA	100%	100%	100%
Precision PrimeraDivision	100%	100%	100%
Precision Ligue1	87,5%	100%	100%
Precision PrimeiraLiga	88,89%	50%	88,89%

Tabela 12. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

5.3. Dokładność a wybór parametru k

	1	5	10	15	30	50
Overall Accuracy	54,71%	55,74%	70,44%	89,31%	88,13%	82,33%
Time	43,73s	41,34s	38,83s	43,69s	39,32s	43,69s

Tabela 13. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii PLACES.

	1	5	10	15	30	50
Recall USA	51,27%	52,27%	71,25%	96,21%	96,77%	99,42%
Recall UK	66,08%	72,38%	71,68%	70,63%	68,18%	0%
Recall Japan	59,38%	66,67%	46,88%	38,54%	0,52%	0%
Recall Canada	74,75%	67,34%	66,33%	45,79%	43,77%	0%
Recall West-Germany	77,19%	77,19%	78,07%	77,19%	77,19%	0%
Recall France	85,51%	86,96%	85,51%	85,51%	85,51%	78,26%
Precision USA	94,11%	94,33%	93,04%	92,69%	90,73%	82,86%
Precision UK	59,62%	79,01%	81,35%	82,45%	82,63%	0%
Precision Japan	21,27%	12,11%	58,82%	66,67%	14,29%	0%
Precision Canada	12,51%	15,52%	15,17%	75,14%	75,58%	0%
Precision West-Germany	54,32%	60,27%	60,96%	60,27%	60,27%	0%
Precision France	46,83%	49,59%	51,3 %	52,68%	53,15%	57,45%

Tabela 14. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	1	5	10	15	30	50
Overall Accuracy	88,11%	91,25%	92,80%	90,91%	90,58%	88,36%
Time	10,60s	10,94s	10,92s	10,80s	10,73s	10,75s

Tabela 15. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	1	5	10	15	30	50
Recall earn	88,93%	89,07%	92,33%	88,42%	87,99%	87,7%
Recall acq	87,93%	95,72%	94,84%	95,72%	95,39%	98,57%
Recall money-fx	77,89%	80%	80%	81,05%	82,11%	0%
Recall grain	0%	0%	0%	0%	0%	0%
Precision earn	93,46%	98,8%	98,38%	98,95%	99,1%	98,86%
Precision acq	83,26%	85,24%	88,89%	84,5%	83,96%	77,35%
Precision money-fx	71,15%	69,72%	71,7%	71,96%	70,27%	0%
Precision grain	0%	0%	0%	0%	0%	0%

Tabela 16. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	1	2	3	4
Overall Accuracy	93,33%	93,33%	93,33%	93,33%
Time	0,66s	0,93s	0,94s	0,90s

Tabela 17. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	1	2	3	4
Recall PremierLeague	100%	100%	100%	100%
Recall Bundesliga	100%	100%	100%	100%
Recall SerieA	50%	50%	50%	50%
Recall PrimeraDivision	100%	100%	100%	100%
Recall Ligue1	100%	100%	100%	100%
Recall PrimeiraLiga	88,89%	88,89%	88,89%	88,89%
Precision PremierLeague	75%	75%	75%	75%
Precision Bundesliga	100%	100%	100%	100%
Precision SerieA	100%	100%	100%	100%
Precision PrimeraDivision	100%	100%	100%	100%
Precision Ligue1	100%	100%	100%	100%
Precision PrimeiraLiga	88,89%	88,89%	88,89%	88,89%

Tabela 18. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

5.4. Dokładność a wybór ilości danych do cold startu

	5	10	15	30	50	100
Overall Accuracy	80,68%	67,61%	69,96%	89,22%	89,19%	71,62%
Time	40,61s	37,94s	41,20s	39,54s	37,79s	39,56s

Tabela 19. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii PLACES.

	5	10	15	30	50	100
Recall USA	80,68%	69,02%	71,77%	95,62%	94,94%	72,36%
Recall UK	0%	70,27%	70,45%	71,74%	69,92%	69,42%
Recall Japan	0%	37,13%	39,59%	42,86%	49,38%	57,14%
Recall Canada	0%	57,33%	56,62%	45,3%	45,32%	65,44%
Recall West-Germany	0%	76,61%	78,15%	76,92%	75%	79,41%
Recall France	0%	84,81%	85,14%	84,75%	82,05%	0
Precision USA	100%	93,67%	93,89%	93,14%	93,95%	95,92%
Precision UK	0%	82,54%	82%	82,16%	81,36%	79,89%
Precision Japan	0%	68,18%	65%	56,93%	50,31%	15,27%
Precision Canada	0%	11,63%	12,56%	73,03%	69,94%	16,25%
Precision West-Germany	0%	62,09%	61,59%	57,97%	50%	31,4%
Precision France	0%	58,77%	55,26%	50%	37,65%	0%

Tabela 20. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	5	10	15	30	50	100
Overall Accuracy	90,92%	90,64%	90,52%	91,01%	92,86%	93,44%
Time	9,44s	10,92s	10,86s	10,87s	10,76s	10,68s

Tabela 21. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	5	10	15	30	50	100
Recall earn	88,98%	87,79%	87,89%	88,34%	92,38%	93,16%
Recall acq	96,33%	95,98%	95,52%	96%	94,55%	93,98%
Recall money-fx	80,91%	80,95%	81%	81,18%	80%	86,67%
Recall grain	0%	100%	100%	0%	0%	0%
Precision earn	98,73%	99,03%	99,03%	98,94%	98,27%	98,38%
Precision acq	83,6%	83,79%	83,89%	84,47%	89,28%	91,13%
Precision money-fx	74,17%	73,91%	70,43%	70,41%	62,65%	26%
Precision grain	0%	36,36%	17,65%	0%	0%	0%

Tabela 22. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	1	2	3	4
Overall Accuracy	50%	93,33%	95,83%	94,74%
Time	0,89s	0,93s	0,91s	0,96s

Tabela 23. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	1	2	3	4	
Recall PremierLeague	0%	100%	100%	100%	
Recall Bundesliga	0%	100%	0%	0%	
Recall SerieA	0%	50%	100%	0%	
Recall PrimeraDivision	100%	100%	100%	100%	
Recall Ligue1	0%	100%	100%	100%	
Recall PrimeiraLiga	90%	88,89%	87,5%	85,71%	
Precision PremierLeague	0%	75%	66,67%	50%	
Precision Bundesliga	0%	100%	0%	0%	
Precision SerieA	0%	100%	100%	0%	
Precision PrimeraDivision	37,5%	100%	100%	100%	
Precision Ligue1	0%	100%	100%	100%	
Precision PrimeiraLiga	75%	88,89%	100%	100%	

Tabela 24. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

5.5. Porównanie dokładności dla cech wykorzystujących miary podobieństwa słów

	N-Gram	Levenshtein	Binarna
Overall Accuracy	73,5%	47,96%	86,34%
Time	32,35s	42,36s	9,15s

Tabela 25. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii PLACES.

	N-Gram	Levenshtein	Binarna
Recall USA	78,65%	55,27%	91,72%
Recall UK	61,19%	3,5%	69,93%
Recall Japan	45,31%	0%	56,77%
Recall Canada	30,3%	15,15%	49,16%
Recall West-Germany	71,93%	71,93%	75,44%
Recall France	71,01%	11,59%	79,71%
Precision USA	91,21%	82,07%	94,02%
Precision UK	44,76%	23,26%	74,63%
Precision Japan	21,01%	0%	45,23%
Precision Canada	53,89%	6,02%	63,76%
Precision West-Germany	28,18%	5,55%	44,56%
Precision France	17,07%	10,81%	41,67%

Tabela 26. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	N-Gram	Levenshtein	Binarna
Overall Accuracy	80,36%	57,45%	87,23%
Time	6,83s	9,70s	2,05s

Tabela 27. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	N-Gram	Levenshtein	Binarna
Recall earn	85,67%	74,67%	95,44%
Recall acq	71,79%	29,53%	74,64%
Recall money-fx	85,26%	74,74%	88,42%
Recall grain	0%	0%	0%
Precision earn	98,75%	99,81%	87,06%
Precision acq	86,28%	69,33%	93,15%
Precision money-fx	30,45%	8,99%	68,29%
Precision grain	0%	0%	0%

Tabela 28. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	N-Gram	Levenshtein	Binarna
Overall Accuracy	70%	33,33%	90%
Time	0,46s	0,91s	0,30s

Tabela 29. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	N-Gram	Levenshtein	Binarna
Recall PremierLeague	100%	0%	100%
Recall Bundesliga	0%	0%	100%
Recall SerieA	50%	50%	100%
Recall PrimeraDivision	75%	12,5%	87,5%
Recall Ligue1	42,86%	0%	100%
Recall PrimeiraLiga	88,89%	88,89%	77,78%
Precision PremierLeague	60%	0%	60%
Precision Bundesliga	0%	0%	100%
Precision SerieA	50%	6,25%	100%
Precision PrimeraDivision	75%	50%	100%
Precision Ligue1	100%	0%	100%
Precision PrimeiraLiga	66,67%	72,73%	87,5%

Tabela 30. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

5.6. Porównanie tekstowych i liczbowych metod obliczania cech

	N-Gram	Levenshtein	ilość słów kluczowych	Binarna
Overall Accuracy	73,5%	47,96%	86,84%	86,34%
Time	32,35s	42,36s	9,89s	9,15s

Tabela 31. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii PLACES.

	N-Gram	Levenshtein	ilość słów kluczowych	Binarna
Recall USA	78,65%	55,27%	93,6%	91,72%
Recall UK	61,19%	3,5%	67,48%	69,93%
Recall Japan	45,31%	0%	49,48%	56,77%
Recall Canada	30,3%	15,15%	42,42%	49,16%
Recall West-Germany	71,93%	71,93%	63,16%	75,44%
Recall France	71,01%	11,59%	79,71%	79,71%
Precision USA	91,21%	82,07%	92,65%	94,02%
Precision UK	44,76%	23,26%	74,52%	74,63%
Precision Japan	21,01%	0%	51,63%	45,23%
Precision Canada	53,89%	6,02%	64,29%	63,76%
Precision West-Germany	28,18%	5,55%	44,72%	44,56%
Precision France	17,07%	10,81%	48,25%	41,67%

Tabela 32. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	N-Gram	Levenshtein	ilość słów kluczowych	Binarna
Overall Accuracy	80,36%	57,45%	86,89%	87,23%
Time	6,83s	9,70s	2,23s	2,05s

Tabela 33. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	N-Gram	Levenshtein	ilość słów kluczowych	Binarna
Recall earn	85,67%	74,67%	96,82%	95,44%
Recall acq	71,79%	29,53%	71,35%	74,64%
Recall money-fx	85,26%	74,74%	91,58%	88,42%
Recall grain	0%	0%	0%	0%
Precision earn	98,75%	99,81%	86,05%	87,06%
Precision acq	86,28%	69,33%	95,73%	93,15%
Precision money-fx	30,45%	8,99%	62,14%	68,29%
Precision grain	0%	0%	0%	0%

Tabela 34. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	N-Gram	Levenshtein	ilość słów kluczowych	Binarna
Overall Accuracy	70%	33,33%	64,52%	90%
Time	0,46s	0,91s	0,19s	0,30s

Tabela 35. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	N-Gram	Levenshtein	ilość słów kluczowych	Binarna
Recall PremierLeague	100%	0%	100%	100%
Recall Bundesliga	0%	0%	0%	100%
Recall SerieA	50%	50%	100%	100%
Recall PrimeraDivision	75%	12,5%	87,5%	87,5%
Recall Ligue1	42,86%	0%	100%	100%
Recall PrimeiraLiga	88,89%	88,89%	0%	77,78%
Precision PremierLeague	60%	0%	42,86%	60%
Precision Bundesliga	0%	0%	0%	100%
Precision SerieA	50%	6,25%	22,22%	100%
Precision PrimeraDivision	75%	50%	100%	100%
Precision Ligue1	100%	0%	100%	100%
Precision PrimeiraLiga	66,67%	72,73%	0%	87,5%

Tabela 36. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

5.7. Porównanie różnych połączeń cech

	Wektor 1	Wektor 2	Wektor 3	Wektor 4
Overall Accuracy	75,73%	88,19%	89,4%	88,99%
Time	92,87s	16,46s	42,68s	49,22s

Tabela 37. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii PLACES.

	Wektor 1	Wektor 2	Wektor 3	Wektor 4
Recall USA	77,55%	94,21%	96,81%	95,51%
Recall UK	70,98%	73,43%	67,83%	70,98%
Recall Japan	53,12%	55,73%	36,98%	40,1%
Recall Canada	64,98%	42,42%	42,09%	48,15%
Recall West-Germany	78,07%	75,44%	78,95%	78,07%
Recall France	86,96%	82,61%	84,06%	86,96%
Precision USA	93,7%	93,36%	92,18%	92,92%
Precision UK	82,19%	74,47%	83,98%	82,19%
Precision Japan	52,85%	49,54%	73,2%	57,46%
Precision Canada	19,4%	75,45%	75,76%	72,96%
Precision West-Germany	59,73%	58,9%	62,5%	59,73%
Precision France	52,17%	52,78%	55,24%	53,57%

Tabela 38. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	Wektor 1	Wektor 2	Wektor 3	Wektor 4
Overall Accuracy	93,01%	93,13%	91,12%	92,76%
Time	21,44s	3,74s	9,73s	12,14s

Tabela 39. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	Wektor 1	Wektor 2	Wektor 3	Wektor 4
Recall earn	92,4%	92,98%	88,28%	92,11%
Recall acq	95,39%	94,84%	96,49%	95,17%
Recall money-fx	78,95%	78,95%	81,05%	78,95%
Recall grain	0%	0%	0%	0%
Precision earn	98,16%	98,02%	99,03%	98,15%
Precision acq	88,95%	89,63%	84,44%	88,56%
Precision money-fx	76,53%	78,12%	74,04%	75%
Precision grain	0%	0%	0%	0%

Tabela 40. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	Wektor 1	Wektor 2	Wektor 3	Wektor 4
Overall Accuracy	90%	90%	90%	88,83%
Time	1,71s	0,26s	0,65s	0,72s

Tabela 41. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	Wektor 1	Wektor 2	Wektor 3	Wektor 4
Recall PremierLeague	66,67%	66,67%	66,67%	66,67%
Recall Bundesliga	100%	100%	100%	100%
Recall SerieA	50%	50%	50%	50%
Recall PrimeraDivision	100%	100%	100%	100%
Recall Ligue1	100%	100%	100%	100%
Recall PrimeiraLiga	88,89%	88,89%	88,89%	88,24%
Precision PremierLeague	66,67%	66,67%	66,67%	66,67%
Precision Bundesliga	50%	50%	50%	50%
Precision SerieA	100%	100%	100%	100%
Precision PrimeraDivision	100%	100%	100%	100%
Precision Ligue1	100%	100%	100%	100%
Precision PrimeiraLiga	88,89%	88,89%	88,89%	88,24%

Tabela 42. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

5.8. Porównanie różnego stosunku danych treningowych do testowych

	30%/70%	50%/50%	70%/30%
Overall Accuracy	86,44%	66,4%	88,99%
Time	103,57s	59,19s	28,76s

Tabela 43. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii PLACES.

	30%/70%	50%/50%	70%/30%
Recall USA	92,97%	64,54%	95,32%
Recall UK	70,76%	75,45%	65,17%
Recall Japan	54,35%	71,37%	42,86%
Recall Canada	39,26%	68,98%	48,79%
Recall West-Germany	69,7%	80%	76,32%
Recall France	83,78%	88,42%	77,14%
Precision USA	92,73%	94,7%	93,38%
Precision UK	68,71%	80,66%	76,82%
Precision Japan	56,21%	70,3%	48,46%
Precision Canada	60,6%	12,46%	74,26%
Precision West-Germany	58,76%	68,89%	61,7%
Precision France	47,69%	54,19%	42,19%

Tabela 44. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii PLACES.

	30%/70%	50%/50%	70%/30%
Overall Accuracy	93,65%	94,4%	94,49%
Time	19,49s	11,70s	5,26s

Tabela 45. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii TOPICS.

	30%/70%	50%/50%	70%/30%
Recall earn	93,21%	94,43%	95,28%
Recall acq	94,64%	95,6%	94,38%
Recall money-fx	91,44%	84,83%	82,54%
Recall grain	100%	0%	0%
Precision earn	98,42%	98,5%	97,87%
Precision acq	91,21%	91,57%	92,83%
Precision money-fx	68,13%	82,55%	74,29%
Precision grain	37,5%	0%	0%

Tabela 46. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii TOPICS.

	30%/70%	50%/50%	70%/30%
Overall Accuracy	85,25%	87,5%	84,21%
Time	1,62s	0,79s	0,5s

Tabela 47. Tabela przedstawiająca procent poprawnie sklasyfikowanych obiektów oraz czas wykonywania klasyfikacji dla kategorii FOOTBALL.

	30%/70%	50%/50%	70%/30%
Recall PremierLeague	100%	66,67%	100%
Recall Bundesliga	100%	66,67%	0%
Recall SerieA	90%	87,5%	50%
Recall PrimeraDivision	80%	90%	100%
Recall Ligue1	90,91%	100%	100%
Recall France	54,55%	88,89%	87,5%
Precision PremierLeague	83,33%	50%	75%
Precision Bundesliga	75%	100%	0%
Precision SerieA	81,82%	87,5%	100%
Precision PrimeraDivision	100%	100%	100%
Precision Ligue1	90,91%	87,5%	100%
Precision PrimeiraLiga	85,71%	88,89%	77,78%

Tabela 48. Tabela przedstawiająca czułość i precyzję dla wybranych etykiet z kategorii FOOTBALL.

6. Dyskusja

6.1. Porównanie metod wyboru słów kluczowych oraz ich ilości

W eksperymencie 5.1 w tabelach 1, 2, 3, 4, 5 oraz 6 zostały przedstawione wyniki pokazujące różnice pomiędzy poszczególnymi metodami wyboru słów kluczowych. Opracowane przez nas metody wykorzystywały odpowiednio metodę Term Frequency oraz Document Frequency aby wydobyć słowa, które

najczęściej pojawiają się w dokumentach. Wybranie tylko najczęstszych słów byłoby jednak bardzo nieefektywne, ponieważ te same słowa pojawiały by się również w innych etykietach. Uniknęliśmy tego dodając metodę, która usuwa z wektorów słów dla danej etykiety słowa, które występowały równie często w innych etykietach. Zapewniamy przez to, że wybrane słowa będą unikalne dla danej etykiety. Jak możemy zauważyć w tabelach 1, 3 oraz 5 wybór ilości słów kluczowych ma duże znaczenie. Mała ilość słów kluczowych dostarcza nam słabe wyniki klasyfikacji. Jest to spowodowane tym, że artykuły, które są klasyfikowane nie posiadają w swoich tekstach słów kluczowych danej klasy. Zbyt duża ilość słów kluczowych również pogarsza wyniki klasyfikacji. Prawdopodobnie powodem jest zaszumienie danych testowych. W wielu tekstach pojawiają się słowa, które zostały wyznaczone dla innych niż przewidywana etykieta. W naszym przypadku najlepszym rozwiązaniem okazała się metoda Term Frequency z ilością słów kluczowych na poziomie 15 słów. Ilość słów kluczowych równa 30 dawała podobne wyniki klasyfikacji lecz większą złożoność obliczeniową.

6.2. Porównanie metryk odległości

Eksperyment 5.2 miał na celu pokazanie różnic pomiędzy metrykami odległości. W tabelach 7, 9 oraz 11 możemy zauważyć, że metoda Uliczna (2) osiąga lepsze wyniki od pozostowałych zarówno jeśli chodzi o złożoność obliczeniową oraz procent poprawnie zaklasyfikowanych obiektów. Prawdopodobnie jest to spowodowane tym, że w metryce euklidesowej (1) małe różnice pomiędzy wartościami stają się jeszcze mniejsze przez co są wręcz prawie negowane, natomiast duże różnice pomiędzy wartościami są wyolbrzymiane. Takie obliczanie odległości powoduje, że niektóre potencjalnie ważne wartości w wektorze są pomijane co może skutkować w gorszych wynikach. W przypadku metody Ulicznej wszystkie wartości obliczane są tak samo, nie ma tutaj zmniejszania ważności mniejszych różnic pomiędzy wektorami.

6.3. Dokładność a wybór parametru k

W eksperymencie 5.3 przedstawione zostały wyniki dotyczące wyników klasyfikacji dla różnych wartości parametru k. Dobór tego parametru ma duże znaczenie dla jakości klasyfikacji. W przedstawionych przez nas wynikach widać, że zbyt mała wartość tego parametru negatywnie wpływa na wyniki klasyfikacji dla k równego 1 oraz 5 procent dla klasyfikacji wyniósł 54-55% dla zestawu PLACES. Spowodowane jest to zaszumieniem danych testowych. Wiele z tekstów posiada słowa kluczowe, które zostały wyekstrahowane dla innych tagów niż przewidywano. Z drugiej strony zbyt duży parametr k powodował zbyt duże uzależnienie od najliczniejszej klasy jak można zauważyć w tabelach 14 oraz 16. Dla zestawu PLACES większość tekstów jest klasyfikowana jako USA. Teksty należące do Francji prawdopobonie znacząco różniły się od tekstów z USA dlatego część z nich została zaklasyfikowana poprawnie. W naszym przypadku na podstawie tabel 13,14 15, 16 najlepszy okazał się parametr k o wartości 15. W przypadku zestawu FOOTBALL wszystkie wartości osiągały podobne wyniki.

6.4. Dokładność a wybór ilości danych do cold startu

Na podstawie eksperymentu 5.4 określiliśmy ile danych powinno być branych do cold startu aby wyniki były możliwie jak najlepsze. Na podstawie tabel 20,22 oraz 24 widzimy, że zbyt mała ilość danych nie wystarcza do klasyfikacji danych. Dla ilości danych równych 5 wynik wyszedł 80,68% tabela 19, ale było to spowodowane tym, że wszystkie artykuły zostały zaklasyfikowane jako USA, których procent we wszystkich artykułach wynosi dokładnie 80,68%. Powodem, może być wybranie tekstów, których wektory cech były zerowe lub posiadały one wartości, które były bardzo zbliżone do artykułów USA. Zauważyliśmy, że zbyt duża ilość również negatywnie wpływa na wyniki - jest to zapewne spowodowane tym, że w danych do cold startu zostały wybrane artykuły, które były bardzo podobne do artykułów, których tagi były różne. Zatem znowu widzimy wpływ zaszumienia danych tym razem w aspekcie wyboru danych do cold startu. W naszym przypadku najlepsze wyniki uzyskaliśmy dla ilości danych 30 dla każdej etykiety.

6.5. Porównywanie dokładności dla cech wykorzystujących miary podobieństwa słów

Eksperyment ten wykonywaliśmy aby pokazać skuteczność i wydajność poszczególnych miar podobieństwa słów. Na podstawie tabel 25, 26,27, 28,29 oraz 30 widzimy, że najgorsze wyniki osiągane są dla miary Levenshteina (6). Zaskoczeniem były dla nas wyniki dla miar binarnej (4) oraz N-Gram (5). Spodziewaliśmy się, że miara n-gram osiągnie lepsze wyniki jeśli chodzi o procent poprawnie sklasyfikowanych danych, gdyż potrafi ona wyznaczyć słowa podobne do szukanych, w przeciwieństwie do miary binarnej, która zwraca 1 w momencie gdy słowa są równe. Wyniki takie mogą być spowodowane przez zbyt wysokie wartości miary podobieństwa dla słów, które wyglądają podobnie ale mają inne znaczenia przez co są podobne do słów kluczowych tagów innych niż oczekiwany. Najszybsze wyniki zostały osiągnięte przez miarę binarną co nie jest tutaj wielkim zaskoczeniem.

6.6. Porównanie tekstowych i liczbowych metod obliczania cech

W tym eskperymencie pokazaliśmy różnice pomiędzy metodami porównywania słów a liczbowymi metodami obliczania cech. Na podstawie tabel 31, 32,33, 34, 35 oraz 36 widać, że wyniki jeśli chodzi o czas są lepsze dla liczbowych metod obliczania cech, nie jest to zaskoczeniem, gdyż metody porównujące słowa są bardziej złożone obliczeniowe niż liczbowe metody obliczania cech. Podobnie jak w poprzednim eksperymencie widzimy, że lepsze wyniki klasyfikacji są osiągane przez binarną miarę podobieństwa oraz przez wyznaczenie ilości słów kluczowych, które wystąpiły w tekście.

6.7. Porównanie różnych połączeń cech

W tym eksperymencie skupiliśmy się na różnych połączeniach cech dla artykułów. W pierwszym wektorze wykorzystaliśmy wszystkie cechy, które umieściliśmy w naszym programie. Wyniki osiągnięte przez ten wektor były najgorsze ze wszystkich. Spowodowane jest to przez miarę Levenshteina,

która dawała nam negatywne wyniki w poprzednich eksperymentach. Wyniki osiągane przez pozostałe wektory były bardzo podobne. Najlepszy pod względem dokładności klasyfikacji okazał się wektor złożony z cech bazowanych na n-gramach i ilości słów kluczowych. Zostało to jednak osiągnięte dużym kosztem wydajności, różnica pomiędzy wektorem drugim a trzecim była ponad dwa razy większa. Jest to duża różnica w szczególności, że dokładność obu wektorów różni się jednym punktem procentowym.

6.8. Porównanie różnego stosunku danych treningowych do testowych

W tym eksperymencie pokazaliśmy różnice w wyborze różnego stosunku danych treningowych i testowych. Wyniki w zależności od różnego rodzaju zbioru danych znacznie się różniły. Spowodowane jest to rozłożeniem tekstów, które nadają się do ekstrakcji słów kluczowych. W tabeli 43 możemy zauważyć, że dla stosunku 30/70 wyniki klasyfikacji dają bardzo pozytywne wyniki natomiast dla stosunku 50/50 są one zdecydowanie gorsze. Powodem może być liczba artykułów, które posiadają charakterystyczne dla danych klas słowa kluczowe. Prawdopodobnie w pierwszych 30% danych treningowych znajdowały się teksty, w których znajdowały się słowa bardzo związane bezpośrednio z daną klasą. W kolejnych 20% danych musiały się zatem znaleźć teksty, które były bardzo zaszumione i dawały negatywne wyniki. Ważne zatem jest dobranie odpowiedniej ilości danych tak, aby znalazły się w nich słowa ważne dla danej klasy. W przypadku innych zestawów danych wyniki były lepsze im wyższy był stosunek danych treningowych i testowych.

7. Wnioski

1. Dobór odpowiednich metod wyboru słów kluczowych jest bardzo ważnym czynnikiem klasyfikacji tekstów. Ich niepoprawny wybór prowadzi do negatywnych wyników klasyfikacji.
2. Zbyt mała lub zbyt duża ilość słów kluczowych prowadzi do pogorszenia klasyfikacji.
3. W przypadku wektorów cech, gdzie każda cecha ma istotne znaczenie lepszym wyborem jest metryka Uliczna, w przeciwieństwie do Euklidesowej, która preferuje większe odległości pomiędzy pojedynczymi cechami
4. Zbyt mała wartość parametru k powoduje znaczące pogorszenie wyników klasyfikacji dla danych, które są bardzo zaszumione
5. Zbyt duża wartość parametru k powoduje pogorszenie wyników klasyfikacji dla danych, gdzie dominuje jedna z klas (jest liczniejsza)
6. Zaszumienie danych wpływa na wybór ilości danych do cold startu. Zbyt mała, bądź zbyt duża ilość danych negatywnie wpływa na wyniki klasyfikacji
7. Miara N-Gram osiąga lepsze wyniki dla miar podobieństwa słów niż miara Levenshtein
8. Poprawnie zaprogramowany klasyfikator jest w stanie klasyfikować teksty z różnych zestawów o różnej tematyce

9. Wybranie odpowiedniej ilości danych treningowych jest bardzo ważne dla poprawnej klasyfikacji. Należy zauważyć, że nie zawsze więcej danych oznacza lepiej. W przypadku, gdy więcej jest danych, które dają negatywne wyniki klasyfikacja nie będzie dawała oczekiwanych rezultatów.

Literatura

[1] https://www.datarobot.com/wiki/classification/
[2] https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python
[3] https://www.merriam-webster.com/dictionary/metric
[4] http://ics.p.lodz.pl/ aniewiadomski/ksr/ksr-projekt1.pdf
[5] http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection