Informatyka, studia STACJONARNE, I st. semestr VI

Komputerowe systemy rozpoznawania **2020/2021**

Prowadzący: dr inż. Marcin Kacprowicz poniedziałek, 12:00

Data oddania: _____ Ocena: _____

Maciej Lewandowski 224357
Kamil Dike 224282

Projekt 1. Klasyfikacja dokumentów tekstowych

Opis projektu ma formę artykułu naukowego lub raportu z zadania badawczego/doświadczalnego/obliczeniowego (wg indywidualnych potrzeb związanych np. z pracą inżynierską/naukową/zawodową).

Wybrane sekcje (rozdziały sprawozdania) są uzupełniane wg wymagań w opisie Projektu 1. i Harmonogramie ZAJĘĆ na WI-KAMP KSR jako efekty zadań w poszczególnych tygodniach.

1. Cel projektu

Celem zadania jest stworzenie systemu klasyfikującego teksty w zależności od kraju który jest opisywany przez tekst. System został oparty o metodę k-NN. Ponad to została przeanalizowana skuteczność działania programu w odniesieniu do nietraktowanego wektora cech.

2. Klasyfikacja nadzorowana metodą k-NN

Metoda k-NN służy do klasyfikacji obiektów. Opiera się na założeniu podobieństwa obiektów blisko położonych w przestrzeni cech. Jak podaje założenia dla algorytmu:
$numclass$ - liczba rozpoznawanych klas
dim - wymiar przestrzeni cech
num - liczba obiektów ciągu uczącego
$sampl[1...num][1...dim+1]$ - ciąg uczący
rec - identyfikator rozpoznanego obiektu

$obj[1...dim]$ - rozpoznawany obiekt
$dist(sampl[k], obj)$ - funkcja podająca odległość między i-tym elementem ciągu uczącego a rozpoznawanym obiektem
k - zmienna określająca ilość uwzględnianych sąsiadów
$tab[1...num][1...2]$ - tablica odległości
$sort(tab)$ - funkcja sortująca tablicę
$fun[1..numclass]$ - tabela wartości funkcji przynależności
$pointmax(fun)$ - funkcja wskazująca numer klasy, dla której wartość przynależności jest maksymalna. Algorytm według składa się z następujących kroków:

```
1.  wyzeruj tablicę fun
2.  wykonaj pętlę s od i=1 do num
3.  w pętli s przyporządkuj elementowi tablicy tab[i][1]
    wynik wywołania funkcji dist(sampl[i],obj)
4.  w pętli s przyporządkuj elementowi tablicy tab[i][2]
    element tablicy sampl[i][dim+1]
5.  zakończ pętlę s
6.  wykonaj sortowanie tablicy sort(tab)
7.  rozpocznij pętlę q od i=1 do i=k
8.  w pętli q przyporządkuj elementowi tablicy fun[tab[i][2]]
    element tablicy fun[tab[i][2]]+1
9.  zakończ pętlę q
10. przyporządkuj zmiennej rec wynik funkcji pointmax(fun)
```

Algorytm jako dane wejściowe pobiera obiekt do klasyfikacji obj oraz zmienną przechowującą informacje o klasie do którego zostanie przyporządkowany rec.Odległość dwóch obiektów określana jest poprzez określoną metrykę. Porównywane będą wektory cech reprezentujące obiekty.

2.1. Ekstrakcja cech, wektory cech

Na potrzeby reprezentacji obiektów poprzez wektory cech wybrano cechy:

1. Liczba słów w dokumencie
$$v_1 = \hat{A} \tag{1}$$
,gdzie
A oznacza artykuł taki, że $A = [s_1, s_2, s_3, ..., s_T]$
s_i oznacza i- te słowo w artykule
$\hat{A}$ oznacza moc zbioru A
2. Wartość logiczna z logiki trój-wartościowej określająca dominujący rodzaj jednostek występujących w tekście. Wartość cechy 1 oznacza że dominują w artykule jednostki układu SI. Wartość cechy 0 oznacza że w artykule dominują jednostki układu Imperialnego. Wartość cechy 1/2 oznacza że w artykule nie dominują jednostki układu SI anie jednostki układu imperialnego.
$$v_2 = l(A) \tag{2}$$
,gdzie
$l : \mathcal{A} \to \{0, \frac{1}{2}, 1\}$, l funkcja przyporządkowywuje artykułowi wartość lo-

giczną 0, 1/2 albo 1 w zależności od ilości wystąpień jednostek danego typu(si/imperialne).
$\mathcal{A}$ oznacza zbiór wszystkich możliwych wektorów reprezentujących artykuły.

3. Najczęściej występujący miesiąc

$$v_3 = m(A) \tag{3}$$

,gdzie
$m : \mathcal{A} \to \{0, 1, 2, ..., 12\}$, m funkcja przyporządkowywująca artykułowi wartość całkowitą od 0 do 12, w zależności od ilości wystąpień danego miesiąca w zbiorze A.

4. Najczęściej występujący typ spółki/firmy

$$v_4 = f(max(k(A, G_S))) \tag{4}$$

,gdzie
$\mathcal{G}$ zbiór wszystkich możliwych wektorów słów kluczowych
$G_S = [x_1, x_2, x_3, ..., x_j]$ wektor słów kluczowych rodzajów spółek
x_i oznacza i-te słowo kluczowe
$\mathcal{H}$ zbiór wszystkich możliwych wektorów częstości występowania słów kluczowych
H wektor częstości występowania słów kluczowych
$f : \mathcal{H} \to \mathcal{G}$, f jest funkcją przyporządkowującą zbiór częstości do zbioru słów kluczowych
$k : \mathcal{A}, \mathcal{G} \to \mathcal{H}$, k jest funkcją zwracającą wektor częstości dla zapewnionego artykułu oraz wektora słów kluczowych

5. Najczęściej występująca w tekście nazwa giełdy

$$v_5 = f(max(k(A, G_G))) \tag{5}$$

,gdzie
$G_g = [x_1, x_2, x_3, ..., x_j]$ wektor słów kluczowych nazw giełd

6. Najczęściej występująca nazwa morza lub oceanu

$$v_6 = f(max(k(A, G_M))) \tag{6}$$

,gdzie
$G_M = [x_1, x_2, x_3, ..., x_j]$ wektor słów kluczowych nazw mórz i oceanów

7. Względna ilość słów o długości do 4 znaków

$$v_7 = \frac{c(A, 0, 4)}{v_1} \tag{7}$$

,gdzie
$c : \mathcal{A}, N, M \to P$ c jest funkcją zliczającą ilość słów o długości od n do m znaków
$N = \{n : n \in \mathrm{N} \wedge n > 0\}$
$M = \{m : m \in \mathrm{N} \wedge m > n\}$
$P = \{p : p \in \mathrm{N}\}$

8. Względna ilość słów o długości od 4 do 8 znaków

$$v_8 = \frac{c(A, 4, 8)}{v_1} \tag{8}$$

9. Względna ilość słów o długości od 8 znaków

$$v_9 = \frac{c(A, 8, \infty)}{v_1} \tag{9}$$

10. Najczęściej występujący rok w artykule

$$v_{10} = yr(A) \tag{10}$$

,gdzie
$yr : \mathcal{A} \to \mathcal{P}$, yr to funkcja zwracająca najczęściej występującą datę w tekście

11. Ilość cen w tekście

$$v_{11} = dl(A) \tag{11}$$

,gdzie
$dl : \mathcal{A} \to \mathcal{P}$, yr to funkcja zwracająca najczęściej występujący rok w tekście

12. Liczba unikalnych słów

$$v_{12} = us(A) \tag{12}$$

,gdzie
$us : \mathcal{A} \to \mathcal{P}$, yr to funkcja zwracająca ilość różnych słów w tekście

2.2. Miary jakości klasyfikacji

Celem miar jakości klasyfikacji jest zbadanie dokonanej klasyfikacji. Ze względu na brak miary idealnej posłużymy się paroma następującymi miarami:

1. accuracy
2. precision
3. recall
4. F1

Do wyznaczenia miar jakości klasyfikacji korzystamy z tablicy pomyłek. Spis oznaczeń:
TP - prawdziwie pozytywna klasyfikacja
FP - fałszywie pozytywna klasyfikacja
FN - fałszywie negatywna klasyfikacja
TN - prawdziwie negatywna klasyfikacja

TP *Przykład dla TP* Jeśli obiekt A w rzeczywistości należy do klasy $\mathcal{A}$, i zostanie sklasyfikowany do klasy $\mathcal{A}$ wówczas klasyfikacja jest uznawana za prawdziwie pozytywną TP.

FP *Przykład dla FP* Jeśli obiekt A w rzeczywistości nie należy do klasy $\mathcal{A}$, i zostanie sklasyfikowany do klasy $\mathcal{A}$ wówczas klasyfikacja jest uznawana za fałszywie pozytywną FP.

FN *Przykład dla FN* Jeśli obiekt A w rzeczywistości należy do klasy $\mathcal{A}$, i nie zostanie sklasyfikowany do klasy $\mathcal{A}$ wówczas klasyfikacja jest uznawana za fałszywie negatywną FN.

TN *Przykład dla TN* Jeśli obiekt A w rzeczywistości nie należy do klasy $\mathcal{A}$, i nie zostanie sklasyfikowany do klasy $\mathcal{A}$ wówczas klasyfikacja jest uznawana za prawdziwie negatywną TN.

2.2.1. Accuracy

Dokładność określa sprawność klasyfikatora. Miara ta jest wspólna dla wszystkich klas. Dokładność wyraża się wzorem:

$$ACC = \frac{\Sigma TP}{\Sigma populacja} \tag{13}$$

2.2.2. Precision

Precyzja jest miarą liczoną dla danej klasy. Miara ta określa precyzje rozpoznawania w obrębie konkretnej klasy. Precyzja wyraża się wzorem:

$$PPV = \frac{\Sigma TP}{\Sigma TP + \Sigma FP} \tag{14}$$

2.2.3. Recall

Recall jest miarą liczoną dla danej klasy. Miara ta określa ilość rozpoznanych elementów danej klasy. Czułość wyraża się wzorem:

$$TPR = \frac{\Sigma TP}{\Sigma TP + \Sigma FN} \tag{15}$$

2.2.4. F1

F1 jest miarą liczoną dla danej klasy. Liczona jest na podstawie miar precision oraz recall jako ich średnia harmoniczna. Miarę F1 wyraża się wzorem:

$$TPR = 2 * \frac{PPV * TPR}{PPV + TPR} \tag{16}$$

3. Klasyfikacja z użyciem metryk i miar podobieństwa tekstów

Program umożliwia wybór trzech metryk do porównywania wektorów cech: metryki miejskiej, amplitudy kosinusowej oraz odległości euklidesowej. Metryka miejska

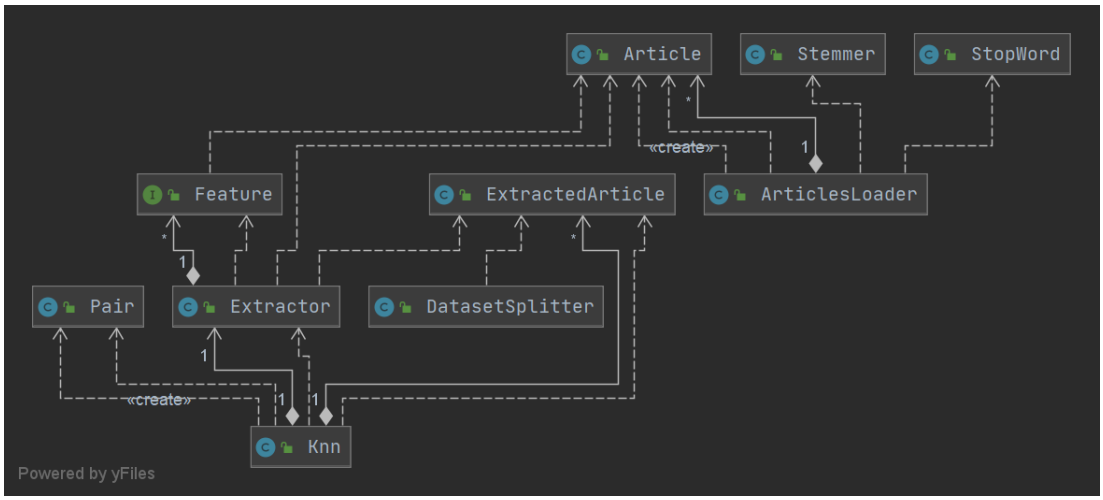
$$\rho_C(X, Y) = \Sigma^n{}_{i=1}|x_i - y_i| \tag{17}$$

Amplituda kosinusowa

$$r_{ca}(V_1, V_2) = \frac{|\Sigma^n{}_{i=1} v_{1i} \cdot v_{2i}|}{\sqrt{\Sigma^n{}_{i=1} v_{1i}^2 \cdot \Sigma^n{}_{i=1} v_{2i}^2}} \tag{18}$$

Odległość euklidesowa

$$\rho_E(X, Y) = \sqrt{\Sigma^n{}_{i=1}(x_i - y_i)^2} \tag{19}$$



Rysunek 1. Struktura aplikacji

Jako metrykę tekstową do ekstrakcji cech tekstowych zastosowano metodę *trigramów*, opisaną poniższym równaniem.

$$sim_3(s_1, s_2) = \frac{1}{N-2}\Sigma^{N-2}{}_{i=1}h(i) \tag{20}$$

, gdzie
$N-2$ - ilość możliwych trój-elementowych podciągów.
$h(i) = 1$ - jeśli trój-elementowy podciąg zaczynający się od i-tej pozycji w s_1 występuje przynajmniej raz w s_2, w innym przypadki $h(i) = 0$.

3.0.1. Wstępne wyniki miary accuracy

Uruchomiono program w czterech konfiguracjach:
Dla zbioru uczącego stanowiącego $60proc$ oraz dla 17 sąsiadów $ACC = 0.7666$.
Dla zbioru uczącego stanowiącego $60proc$ oraz dla 3 sąsiadów $ACC = 0.7693$.
Dla zbioru uczącego stanowiącego $30proc$ oraz dla 17 sąsiadów $ACC = 0.7913$.
Dla zbioru uczącego stanowiącego $30proc$ oraz dla 3 sąsiadów $ACC = 0.7379$.

4. Budowa aplikacji

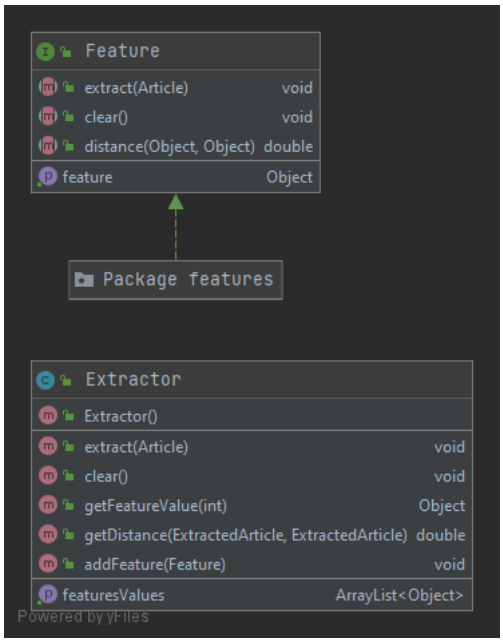
4.1. Diagramy UML

4.1.1. Struktura aplikacji

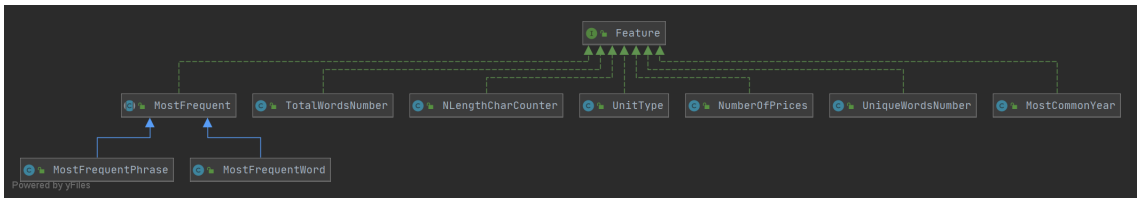
Aplikacja złożona jest z komponentów: extractor, features, knn, main, model, parser, utils. Struktura aplikacji została przedstawiona na rysunku 4.1.1 na stronie 6.

4.1.2. extractor

Pakiet ekstraktor udostępnia narzędzia umożliwiające ekstrakcję cech z tekstów. Pakiet ekstraktor zaprezentowano na rysunku 4.1.2 na stronie 7.



Rysunek 2. Pakiet extrator



Rysunek 3. Struktura pakietu features

4.1.3. features

Pakiet features udostępnia abstrakcję cechy oraz modeluje logikę konkretnych cech. Pakiet features zaprezentowano na rysunku 4.1.3 oraz 4.1.3 na stronie 7.

4.1.4. knn

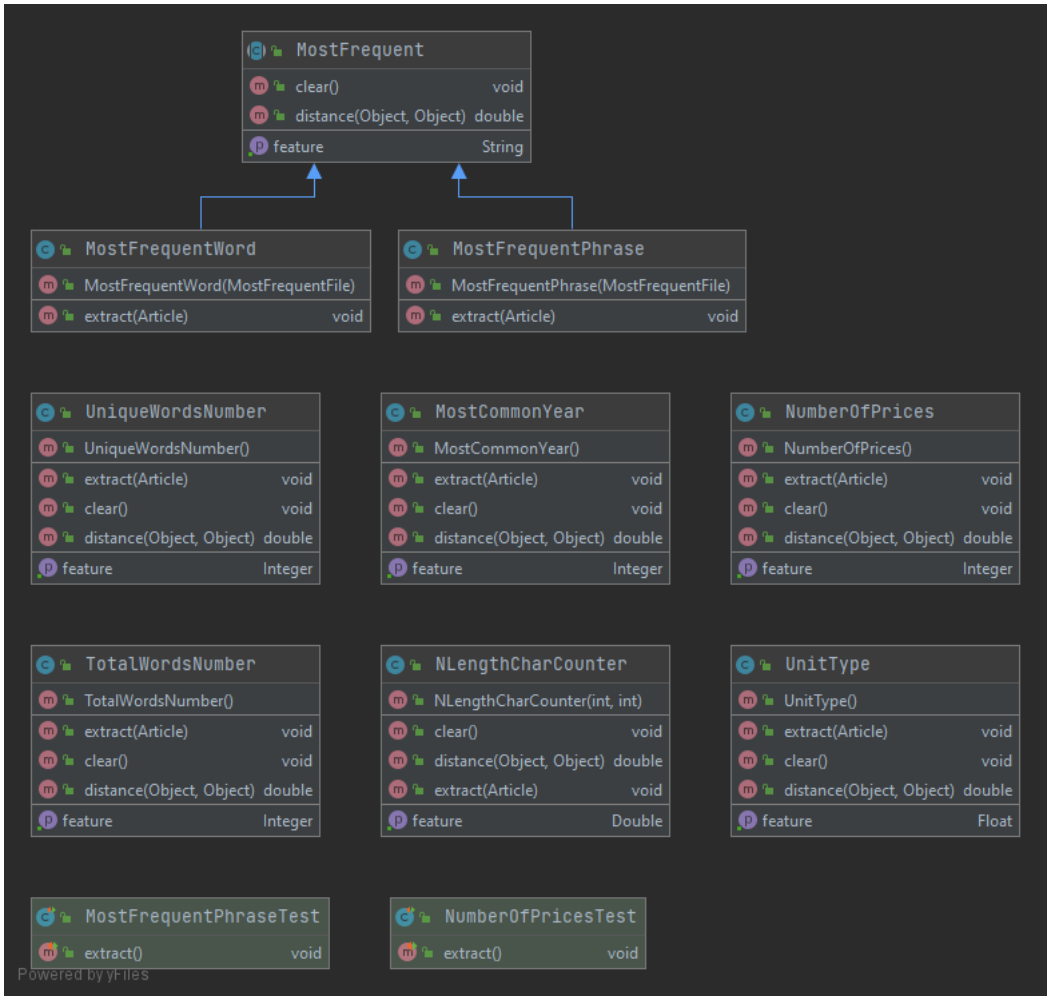
Pakiet knn udostępnia algorytm $k - NN$. Pakiet knn zaprezentowano na rysunku 4.1.4 na stronie 8.

4.1.5. main

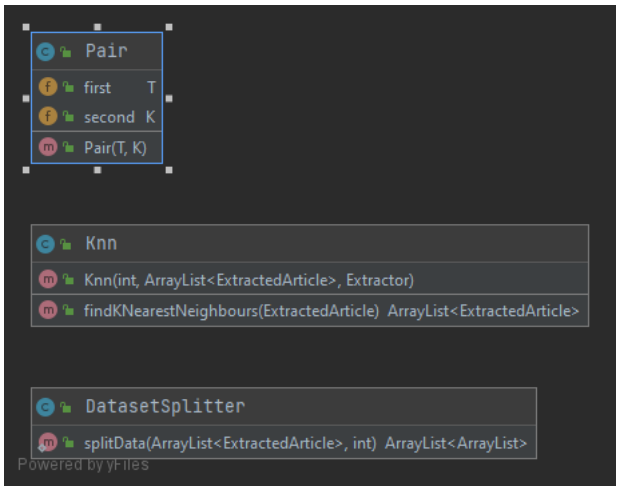
Pakiet main stanowi wejście, oraz implementuje logikę CLI. Pakiet main zaprezentowano na rysunku 4.1.5 na stronie 9.

4.1.6. model

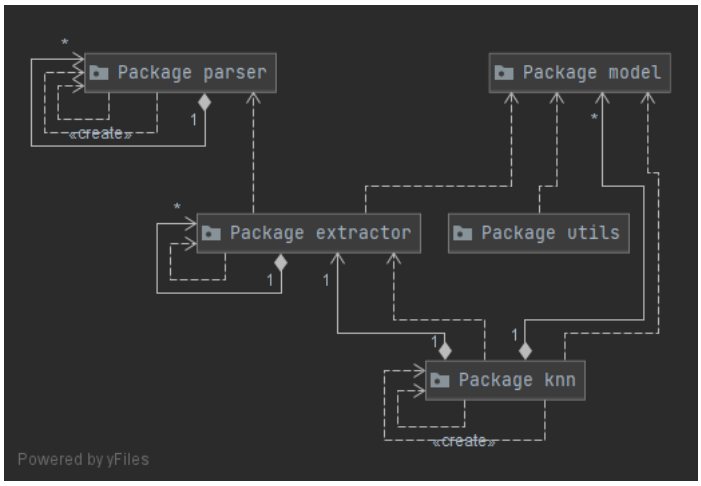
Pakiet model dostarcza model danych dla artykułu reprezentowanego jako wektor cech oraz pozostałe wymagane modele danych. Pakiet model zaprezentowano na rysunku 4.1.6 na stronie 9.



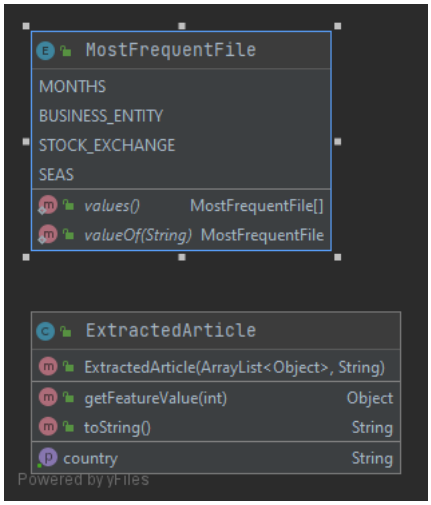
Rysunek 4. Pakiet features



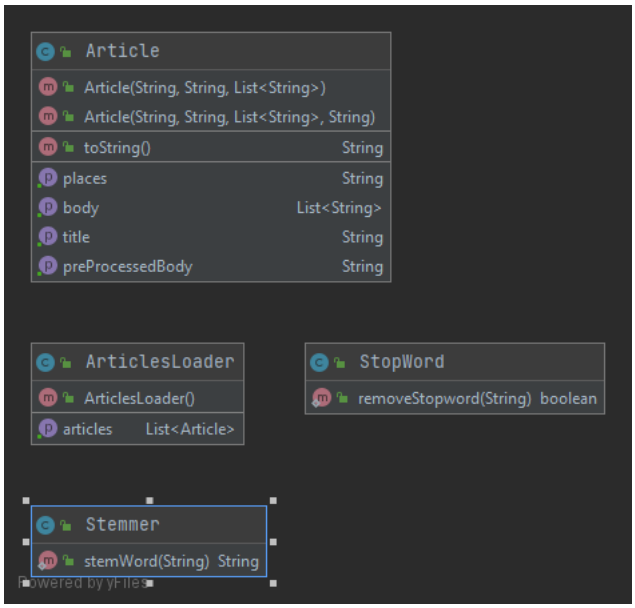
Rysunek 5. Pakiet knn



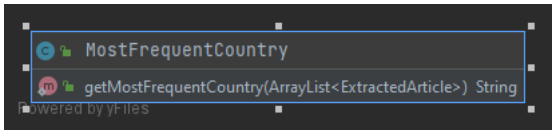
Rysunek 6. Pakiet main



Rysunek 7. Pakiet model



Rysunek 8. Pakiet parser



Rysunek 9. Pakiet utils

4.1.7. parser

Pakiet parser udostępnia metody odpowiedzialne za przygotowanie tekstu przed ekstrakcją cech. Pakiet parser zaprezentowano na rysunku 4.1.7 na stronie 10.

4.1.8. utils

Pakiet utils udostępnia narzędziowe metody wykorzystywane przez pozostałe pakiety. Pakiet utils zaprezentowano na rysunku 4.1.8 na stronie 10.

4.2. Prezentacja wyników, interfejs użytkownika

4.2.1. Interfejs użytkownika

Interfejs użytkownika jest w formie tekstowej. Poniżej pokazano przykładowe uruchomienie programu przy udziale zbioru treningowego 60*proc*, siedmiu sąsiadów, do klasyfikacji wykorzystano wszystkie kraje, oraz użyto wszystkich cech opisanych wyżej.

```
Okresl procent zbioru treningowego:
60
Okresl ilosc sasiadow KNN:
17
Wybierz interesujące kraje:
```

```
0. Wszystkie
1. USA
2. UK
3. Japan
4. Canada
5. West-Germany
6. France
Y. Przejdz dalej
0
Wybierz interesujące metryki:
0. Wszystkie
1. TotalWordsNumber
2. unitType
3. shortWords
4. middleWords
5. longWords
6. mostCommonYear
7. mostFrequentWordMonth
8. mostFrequentWordBusinessEntity
9. mostFrequentStockExchange
10. mostFrequentSea
11. numberOfPrices
12. uniqueWordsNumber
Y. Przejdz dalej
0
loading...
```

Wynik wygenerowany przez program będzie następujący:

```
Dla usa _________
Dokladnosc: 0.766563241467421
Precyzja: 0.789354044326577
Czulosc: 0.9643601018282805
F1: 0.8681249999999999


Dla uk _________
Dokladnosc: 0.766563241467421
Precyzja: 0.165
Czulosc: 0.0889487870619946
F1: 0.11558669001751312


Dla japan _________
Dokladnosc: 0.766563241467421
Precyzja: 0.5
Czulosc: 0.0
F1: 0.0
```

```
Dla canada _________
Dokladnosc: 0.766563241467421
Precyzja: 0.5
Czulosc: 0.0
F1: 0.0


Dla west-germany _________
Dokladnosc: 0.766563241467421
Precyzja: 0.5
Czulosc: 0.0
F1: 0.0


Dla france _________
Dokladnosc: 0.766563241467421
Precyzja: 0.5
Czulosc: 0.0
F1: 0.0
```

5. Wyniki klasyfikacji dla różnych parametrów wejściowych

Wyniki kolejnych eksperymentów wg punktów 2.-8. opisu projektu 1. Wykresy i tabele obowiązkowe, dokładnie opisane w „captions" (tytułach), konieczny opis osi i jednostek wykresów oraz kolumn i wierszy tabel.

Ewentualne wyniki realizacji punktu 9. opisu Projektu 1., czyli „na ocenę 5.0" i ich porównanie do wyników z części obowiązkowej.

Sekcja uzupełniona jako efekt zadania Tydzień 05 wg Harmonogramu Zajęć na WIKAMP KSR.

6. Dyskusja, wnioski

Dokładne interpretacje uzyskanych wyników w zależności od parametrów klasyfikacji opisanych w punktach 3.-8 opisu Projektu 1. Szczególnie istotne są wnioski o charakterze uniwersalnym, istotne dla podobnych zadań. Omówić i wyjaśnić napotkane problemy (jeśli były). Każdy wniosek/problem powinien mieć poparcie w przeprowadzonych eksperymentach (odwołania do konkretnych wyników: wykresów, tabel).
Dla końcowej oceny jest to najważniejsza sekcja sprawozdania, gdyż prezentuje poziom zrozumienia rozwiązywanego problemu.

** Możliwości kontynuacji prac w obszarze systemów rozpoznawania, zwłaszcza w kontekście pracy inżynierskiej, magisterskiej, naukowej, itp. **

Sekcja uzupełniona jako efekt zadania Tydzień 06 wg Harmonogramu Zajęć na WIKAMP KSR.

7. Braki w realizacji projektu 1.

Wymienić wg opisu Projektu 1. wszystkie niezrealizowane obowiązkowe elementy projektu, ewentualnie podać merytoryczne (ale nie czasowe) przyczyny tych braków.

Literatura

[1] R. Tadeusiewicz: Rozpoznawanie obrazów, PWN, Warszawa, 1991.

[2] A. Niewiadomski, Methods for the Linguistic Summarization of Data: Applications of Fuzzy Sets and Their Extensions, Akademicka Oficyna Wydawnicza EXIT, Warszawa, 2008.

Literatura zawiera wyłącznie źródła recenzowane i/lub o potwierdzonej wiarygodności, możliwe do weryfikacji i cytowane w sprawozdaniu.