Informatyka, studia STACJONARNE, I st. semestr VI

Komputerowe systemy rozpoznawania **2020/2021**

Prowadzący: prof. dr hab. inż. Adam Niewiadomski poniedziałek, 12:00

Data oddania: _____ Ocena: _____

Hubert Gawłowski 224298
Kamil Kiszko-Zgierski 224328

Projekt 1. Klasyfikacja dokumentów tekstowych

1. Cel projektu

Celem zadania jest zaimplementowanie algorytmu k-NN w technologii Java na potrzeby klasyfikacji tekstów oraz zbadanie wpływu wybranych cech liczbowych i tekstowych na skuteczność powyższej metody. W wyniku działania algorytmu teksty zostaną przyporządkowane do krajów, z jakich pochodzą. Badanie zostanie przeprowadzone na podstawie artykułów prasowych z agencji prasowej Reuters, które to pochodzą z 1987 roku, wszystkie teksty napisane są w języku angielskim, a przy klasyfikacji pod uwagę będą brane artykuły, które pochodzą z następujących krajów: Republika Federalna Niemiec, USA, Francja, Wielka Brytania, Kanada, Japonia.

2. Klasyfikacja nadzorowana metodą k-NN

Algorytm k-NN (od angielskich słów nearest neighbour - najbliższy sąsiad) to algorytm, którego działanie polega na przyporządkowaniu obiektu poddanego rozpoznawaniu do jednej z klas. Do wykorzystania tego algorytmu niezbędny jest zestaw klas, do których może należeć obiekt, zbiór danych uczących oraz rozpoznawany obiekt. Metoda k-NN należy do grupy metod minimalnoodległościowych, ponieważ o zaklasyfikowaniu obiektu do danej klasy decyduje najmniejsza odległość (zgodna z przyjętą metryką), pomiędzy rozpoznawanym obiektem oraz k-obiektami z ciągu uczącego. Wyszukiwanie

najmniejszej odległości pomiędzy obiektami można przedstawić za pomocą wzoru ogólnego:

$$\rho(x, x^{i,k}) = \min_{x^\mu \in U^i}(x, x^\mu) \tag{1}$$

gdzie ρ to wybrana metryka, U^i oznacza ciąg uczący, $x^{i,k}$ jest elementem zbioru U^i, a rozpoznawany obiekt to x [1].

Skuteczność algorytmu k-NN mierzona jest na podstawie odsetka poprawnych przyporządkowań obiektów do odpowiadających im klas.

2.1. Ekstrakcja cech, wektory cech

Pierwszym etapem, który należy wykonać w procesie rozpoznawania tekstów jest wyodrębnienie takich cech, aby jak najlepiej określały ich charakterystykę.Wszystkie artykuły są napisane w tym samym języku oraz w tej samej formie stylistycznej, dlatego w trakcie analizy skupiliśmy się na cechach liczbowych oraz tekstowych. Mając to na uwadzę, dokonaliśmy ekstrakcji poniższych cech:

1. Zapis cyfr - za pomocą analizy ciągu cyfr otrzymamy informacje o np. numerach telefonicznych, które to są charakterystyczne dla omawianego kraju. Z pomocą symboli matematycznych ceche tą możemy zapisać następująco:
$$c_1 = \frac{|s : s \in N \wedge s \in P_k|}{w} \tag{2}$$
gdzie N- zbiór ciągów cyfr znalezionych w dokumencie, P - zbiór charakterystycznych ciągów cyfr dla danego kraju k (Słowniki charakterystycznych ciągów cyfr znajdują się w Załączniku 1), w - waga przez jaką należy podzielić otrzymaną cechę.
Wzór ten zastosujemy dla każdego z rozpatrywanych krajów i dzięki porównaniu otrzymanych wyników uzyskamy informacje do której grupy, na podstawie tej cechy, sklasyfikować dany dokument.

   Przykład 2.1. *Fragment artykułu pt. "Offers USA direct service in Denmark" [3]*
   "[...]The service allows callers in Denmark to reach an ATT operator in the United States by dialing a single telephone number, 0430-0010, ATT said.[...]".

   Z powyższego fragmentu możemy wyodrębnić ciąg cyfr 0430-0010 i na jego podstawie sklasyfikować, do jakiej etykiety kraju możemy zaklasyfikować dany artykuł. Numery telefonów z różnych krajów mogą być rozpoznane na podstawie np. numeru kierunkowego, ich długości czy też formatu zapisu.

2. Waluty - wyciągnięcie z tekstów nazw najczęsciej używanych walut. Każdy z krajów posługuje się inną walutą [1], dlatego jest to cecha, która jasno

[1] Omawiane artykuły pochodzą z lat 80, kiedy we Francji i w Niemczech obowiązywała inna waluta (odpowiednio frank francuski i marka niemiecka). Kraje te przyjęły wspólną walutę, tj. euro dopiero w 2002 roku

charakteryzuje nam wybrane kraje. Wzory do tej cechy będą prezentowały się nasepująco:

$$c_2' = d_t(M) \tag{3}$$

gdzie d_t jest funkcją wyznaczającą t najczęściej występujących słów w danym zbiorze, M to zbiór znalezionych w artykule słów oznaczających waluty. Zbiór walut charakterystycznych dla krajów znajduje się w Załączniku 1.

$$c_2 = g_k(c_2') \tag{4}$$

gdzie g_k jest funkcją przyporządkowującą wyznaczone słowa do poszczególnego kraju k.
W wyniku porównania otrzymanych dla każdego kraju wartości będziemy mogli stwierdzić do którego kraju, na podstawie tej cechy, można przyporządkować podany tekst.

Przykład 2.2. *Fragment artykułu pt. "Maxtor agrees to acquire U.S. design" [3]*
"[...]They said the arrangement, which is subject to a number of conditions including U.S. Design shareholder approval, calls for Maxtor to issue 12 mln dlrs worth of its own common stock in exchange for all of U.S. Design.[...]".

W powyższym fragmencie została wymieniona waluta o nazwie dolar ("dlrs"). Mimo, że najbardziej popularnym dolarem jest dolar amerykański, natomiast na świecie jest jeszcze wiele innych walut, których pierwszym członem jest słowo "dolar", np. dolar kanadyjski, dolar australijski. Ten fakt należy również wziąć pod uwagę w momencie wyznaczania zbiorów rozmytych. Z podanego fragmentu wynika także, że aby w pełni skorzystać z tej cechy, należy uwzględnić nie tylko pełne nazwy walut, ale również ich skróty, które również się pojawiają w artykułach.

3. Częstość występowania dat - zliczenie, jak często w podanych tekstach występują elementy określające czas. Wydaje się, że ich częstość będzie się różnić w zależności od pochodzenia tekstu. Powyższą cechę można przedstawić następująco:

$$c_3 = \frac{|s : s \in D \wedge s \in A|}{|A|} \tag{5}$$

gdzie D - zbiór słów oznaczających daty (zbiór formatów dat, według których słowo będzie klasyfikowane jako data znajduje się w Załączniku 1), A - zbiór wszystkich słów z artykułu.

Przykład 2.3. *Fragment artykułu pt. "USDA comments on export sales" [3]*

"[...] In comments on its Export Sales Report, the department said sales of 1.0 mln tonnes to the USSR – previously reported under the daily reporting system – were the first sales for delivery to the USSR under the fourth year of the U.S.-USSR Grains Supply Agreement, which began October 1. [...] Egypt, Japan and Iraq were the major wheat buyers for delivery in the current year, while sales to China decreased by 30,000 tonnes for the current season, but increased by 90,000 tonnes for the 1987/88 season, which begins June 1. [...]".

W przytoczonym fragmencie zapis daty został wykorzystany 3 razy ("October 1", "1987/88", "June 1"). Wobec tego, uważamy, że opisywana cecha będzie korzystnie wpływać na proces klasyfikacji tekstów.

4. Format zapisu dat - w zależności od kraju format zapisu dat różni się. Cechę tą zapisać można za pomocą wzorów:

$$c_4' = f(B) \tag{6}$$

gdzie f - funkcja wyznaczająca zapis datowy z podanego zbioru, B - zbiór wszystkich wyrażeń znajdujących się w artykule (wyrażenie traktujemy jako połączenie minimum dwóch słów). Formaty zapisu dat, jakie będą brane pod uwagę przy wyznaczaniu znajdują się w Załączniku 1.

$$c_4 = t_k(c_4') \tag{7}$$

gdzie t_k jest funkcją przyporządkowującą wyznaczone wyrażenia do poszczególnego kraju k.
Jako, że w kilku krajach stosowany jest ten sam zapis datowy, funkcja ta może (a wręcz jest to bardzo prawdopodobne) zwrócić taką samą wartość dla kilku krajów. W wyniku porównania otrzymanych dla każdego kraju wartości będziemy mogli stwierdzić do którego kraju (bądź kilku krajów z równym prawdopodobińśtwem), na podstawie tej cechy, można przyporządkować podany tekst.

Przykład 2.4. *Fragment artykułu pt. "Software services extends warrants" [3]*
"[...]Software Services of America Inc said its board has extended the expiration date of its warrants until August 31 from April 30.[...]"

Daty występujące w tym fragmencie ("August 31" i "April 30") są zapisane w formacie: miesiąc dzień. Uważamy, że w zależności od tego, z jakiego kraju pochodzi artykuł format zapisu dat może się różnić.

5. Ogólna liczba słów - zliczenie wszystkich słów występujących w tekście. Uważamy, że w zależności od tego, jakiego kraju tekst dotyczy, ich długość może być różna. Liczbę wyrazów znajdujących się w tekście można przedstawić następująco:

$$c_5 = |A| \tag{8}$$

gdzie A - zbiór wszystkich słów z artykułu.

6. Częstość słów rozpoczynających się wielką literą - słowa takie będą oznaczały najczęściej nazwy własne np. imiona, nazwiska, nazwy budynków lub będą to rozwinięcia skrótów. Pisząc o jednym kraju może być używane więcej takich słów, a o innych mniej. Z tej grupy wykluczymy jednak wyrazy składające się wyłącznie z wielkich liter (o których mowa będzie w punkcie następnym) oraz słowa pisane z wielkiej litery z uwagi na początek zdania. Aby odróżnić wielkie litery od małych, trzeba na początku dokonać odwzorowania liter w słowie na kody ASCII i następnie sprawdzić, czy odpowiedni kod ASCII znajduje się w przedziale od 65 do 90. W postaci wzoru wygląda to następująco:

$$f(l) = \begin{cases} 1 & \text{jeśli } l \in < 65, 90 > \\ 0 & \text{jeśli } l \notin < 65, 90 > \end{cases} \tag{9}$$

gdzie l oznacza pojedyńczy znak zapisany za pomocą kodu ASCII, a funkcja f(l) zwraca 1 dla liter zapisanych wielką literą, a 0 dla pozostałych znaków. Natomiast w celu obliczenia częstości słów rozpoczynających się wielką literą należy skorzystać z poniższego wzoru:

$$c_6 = \frac{|s : s \in Z \wedge s \notin W \wedge s \notin M|}{|A|} \tag{10}$$

gdzie Z - zbiór słów, rozpoczynających się w arrtykule wielką literą, A - zbiór wszystkich słów z artykułu, W - zbiór słów, pisanych w artykule wielkimi literami, M - zbiór słów, które rozpoczynają w artykule zdania.

Przykład 2.5. *Fragment artykułu pt. "U.S. Auto Union will fight to stop job/wage cuts" [3]*
"[...]The United Auto Workers union (UAW) vowed to fight wage and job cuts in a round of labour talks starting in July that cover nearly 500,000 workers at General Motors Corp and Ford Motor Co[...]".

W tym krótkim fragmencie występuje aż 9 słów rozpoczynających się wielką literą, jednocześnie nie będących pierwszym słowem w zdaniu oraz nie będących słowem składających się tylko z wielkich liter. Słowa te są w tym fragmencie związane z nazwami własnymi oraz nazwą miesiąca. Uważamy, że przede wszystkim stosowanie nazw własnych może być związane z tym, z jakiego kraju pochodzi podany dokument.

7. Częstość słów pisanych wielkymi literami - najczęściej będą to skróty. Uważamy, że w zależności od opisywanego kraju, ilość wykorzystywanych skrótów może się różnić. Do policzenia wystąpień słów zapisanych wielkimi literami należy wykorzystać poniższy wzór:

$$c_7 = \frac{|s : s \in W|}{|A|} \tag{11}$$

gdzie W - zbiór słów, pisanych w artykule wielkimi literami, A - zbiór wszystkich słów z artykułu.

Przykład 2.6. *Fragment artykułu pt."France approves large defence spending increase" [3]*
"[...]The budget represents a six pct annual increase, starting next year, well above the 3.5 pct NATO recommends for members of its military command. France is a member of NATO but does not belong to its integrated military command.[...]"

W powyższym fragmencie skrót NATO(Organizacja Traktatu Północnoatlantyckiego) występuje 2 razy. Według nas, częstość występowania skrótów, w danym artykule może mieć związek z tym, jakiego kraju dotyczy tekst.

8. Układ SI/imperialny - zdecydowanie częściej w artykułach z krajów anglojęzycznych będzie stosowany układ imperialny, natomiast w pozostałych - układ SI. Wyliczenie liczby wystąpień jednostek w układzie SI można wyrazić następująco:

$$c'_8 = \frac{|s : s \in S \wedge s \in A|}{|A|} \tag{12}$$

gdzie S - zbiór słów oznaczających jednostki układu SI, A - zbiór wszystkich słów z artykułu.
Z kolei wzór do wyliczenia liczby wystąpień jednostek w układzie imperialnym przedstawia się w poniższy sposób:

$$c''_8 = \frac{|s : s \in I \wedge s \in A|}{|A|} \tag{13}$$

gdzie I - zbiór słów oznaczających jednostki układu imperialnego, A - zbiór wszystkich słów z artykułu.
Zbiór słów oznaczających jednostki układu SI oraz układu imperialnego znajduje się w Załączniku 1.
Jako opisywaną cechę zapisujemy różnicę wystąpień jednostek w układzie SI oraz imperialnym, czyli:

$$c_8 = c'_8 - c''_8 \tag{14}$$

Przykład 2.7. *Fragment artykułu pt. "Sun in North Dakota oil find" [3]*
"[...]flowed 660 barrels of oil and 581,000 cubic feet of natural gas per day through a 13/64 inch choke from depths of 13,188 to 13,204 feet.[...]".

W powyższym tekście można zauważyć występowanie jednostkek z układu imperialnego, tj. cale(inch) i stopy(feet). Wobec tego, można przypuszczać, że tekst ten pochodzi z jednego z krajów anglojęzycznych.

9. Częstość występowania cytatów - kolejna cecha, która wydaje się różnić w zależności od kraju, o którym mowa w artykule. Liczba cytatów zostanie uzyskana w wyniku obliczenia liczby występowania słów, gdzie przedostatni znak to ',' lub '.', a ostatni '"'. Wyznaczenie tej cechy można zaprezentować w postaci wzoru:

$$c_9 = \frac{|s : s \in Y|}{|A|} \tag{15}$$

gdzie Y - zbiór cytatów występujących w artykule, A - zbiór wszystkich słów z artykułu.

Przykład 2.8. *Fragment artykułu pt. "Hughes changes stance on merger after suit" [3]*
"[...]"I think the merger is not going through," said Phil Pace, analyst at Kidder, Peabody and Co. He said the merger "lost a lot of its appeal" when the U.S. Department of Justice required that Baker sell off its Reed Tool Co operation.[...]".

W podanym fragmencie cytat wystąpił 2 razy. Uważamy, że artykuły dotyczące różnych krajów będą też zawierały różną liczbę cytatów.

10. Słowa kluczowe - sporządzone zostaną listy elementów identyfikujących każdy z krajów. Określenia te będą związane z elementami charakterystycznymi dla danego kraju. Możemy do nich zaliczyć nazwy geograficzne, znane osoby, nazwy firm itp.. Do wyznaczenia słów kluczowych należy skorzystać ze wzoru:

$$c_{10} = \frac{|s : s \in K \wedge s \in A|}{|A|} \tag{16}$$

gdzie K - zbiór słów kluczowych (zbiory słów kluczowych dla każdego kraju znajdują się w Załączniku 1), A - zbiór wszystkich słów z artykułu.

Przykład 2.9. *Fragment artykułu pt. "Currency futures to key off G-5, G-7 meetings" [3]*
"[...]News of an agreement among G-5 and G-7 finance ministers meeting in Washington this week will be key to the direction of currency futures at the International Monetary Market, but any such agreement will need to go beyond the Paris accord to stem the recent rise in futures, financial analysts said.[...]".

Powyższy tekst zawiera 2 słowa kluczowe - Washington i Paris. Washington związane jest z USA, natomiast Paris z Francją. Chcąc przyporządkować ten fragment biorąc pod uwagę tylko i wyłącznie cechę związaną ze słowami kluczowymi zostałby dopasowany z równym prawdopodobieństwem do Francji lub USA.

11. Najczęściej występujące słowa - wyodrębnienie z artykułów najczęściej występujących słów, z pominięciem słów znajdujących się na tzw. stopli-

ście, tj. liście najczęściej używanych słów w języku angielskim [2]. Zabieg ten ma na celu podniesienie jakości klasyfikacji poprzez wyszukanie słów, które charakteryzują treść artykułu. Pominięcie tej operacji skutkowałoby niejednoznacznym zaklasyfikowaniem tekstów, co w konsekwencji obniżyłoby skuteczność algorytmu. Wyznaczenie opisywanej cechy można przedstawić w postaci operacji na zbiorach:

$$c_{11} = d_t(A - S) \tag{17}$$

gdzie d_t jest funkcją wyznaczającą t najczęściej występujących słów w danym zbiorze, A to zbiór słów w artykule, a S jest zbiorem słów znajdujących się na stopliście.

Przykład 2.10. *Fragment artykułu pt. "Houston oil trust" [3]*
"[...] The most significant factor for the lack of a distribution this month is the establishment of additional special cost escrow accounts, the company said, adding, that there may be no cash distribution in other months or during the remainder of the year [...]"

Dla powyższego przykładu załóżmy, że stoplista obejmuje 100 najczęściej używanych słów w języku angielskim. Stosując wzór (2) okazuje się, że najczęściej występującym charakterystycznym słowem jest distribution, które pojawiło się w tekście 3 razy.

Ostatecznie, wektor wyekstrahowanych cech będzie się prezentował następująco:

$$v = [c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}] \tag{18}$$

Zbiór numerów cech tekstowych możemy przedstawić w następujący sposób:

$$T = \{1, 2, 4, 10, 11\} \tag{19}$$

Natomiast zbiór numerów cech liczbowych wygląda następująco:

$$L = \{3, 5, 6, 7, 8, 9\} \tag{20}$$

2.2. Miary jakości klasyfikacji

Podczas procesu klasyfikacji niezbędne jest określenie jak skuteczna i jakościowa jest prowadzona klasyfikacja. W tym celu posłużyliśmy się tablicą(macierzą) pomyłek oraz miarami jakości klasyfikacji.[5] Macierz pomyłek zawiera ilość próbek przypisanej do poszczególnej grupy (prawdziwie pozytywna, fałszywie pozytywna, fałszywie negatywna, prawdziwie negatywna). Grupy te powstają poprzez zestawienie ze sobą klasy rzeczywistej oraz klasy predykowanej. Tablica pomyłek prezentuje się następująco:

Zostały wykorzystane następujące miary jakości, które zostaną przedstawione wzorami, w których oznaczenia odnosić się będą do Tabeli 1:

		Klasa rzeczywista	
		pozytywna	negatywna
Klasa predykowana	pozytywna	prawdziwie pozytywna (TP)	fałszywie pozytywna (FP)
	negatywna	fałszywie negatywna (FN)	prawdziwie negatywna (TN)

Tabela 1. Tablica pomyłek

— Accuracy (dokładność) - miara, która oznacza dokładność całej klasyfikacji:

$$ACC = \frac{TP + TN}{L} \tag{21}$$

gdzie L - liczba wszytskich sklasyfikowanych próbek

— Precision (precyzja) - miara oznaczająca jak często określoną klasę udało się zakwalifikować poprawnie:

$$PPV = \frac{TP}{TP + FP} \tag{22}$$

— Recall (czułość) - określa jak dużo wystąpień obiektów danej klasy zakwalifikowaliśmy do tejże klasy:

$$TPR = \frac{TP}{TP + FN} \tag{23}$$

— F1 - miara stanowiąca średnią harmoniczną z miar Precision i Recall [6]:

$$F_1 = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR} \tag{24}$$

Dla przykładu obliczmy miary jakości dla poniższej tabeli:

		Klasa rzeczywista	
		$Japonia$	$\sim Japonia$
Klasa predykowana	$Japonia$	56	12
	$\sim Japonia$	17	36

Tabela 2. Tablica pomyłek z przykładowymi danymi

$$ACC = \frac{56 + 36}{121} = 0,76 \tag{25}$$

$$PPV = \frac{56}{56 + 12} = 0,82 \tag{26}$$

$$TPR = \frac{56}{56 + 17} = 0,77 \tag{27}$$

$$F_1 = 2 \cdot \frac{0,82 \cdot 0,77}{0,82 + 0,77} = 0,79 \tag{28}$$

Jako, że Precission oraz Recall są miarami jakości dla określonej klasy, aby uzyskać wynik dla całego zbioru artykułów wprowadziliśmy średnią ważoną:

— Dla Precission wzór wygląda wygląda następująco:

$$TPR_{waz} = \frac{\sum_{i=1}^{n} w_i TPR_i}{\sum_{i=1}^{n} w_i} \tag{29}$$

gdzie n - liczba klas w procesie klasyfikacji, TPR_i - miara Precission obliczona dla i-tej klasy, w_i - liczba dokumentów z i-tej klasy, zatem $\sum_{i=1}^{n} w_i$ oznacza liczbę wszystkich klasyfikowanych dokumentów.

— Dla Recall wzór wygląda analogicznie, jak powyżej:

$$PPV_{waz} = \frac{\sum_{i=1}^{n} w_i PPV_i}{\sum_{i=1}^{n} w_i} \tag{30}$$

gdzie n -liczba klas w procesie klasyfikacji, PPV_i - miara Recall obliczona dla i-tej klasy, w_i - liczba dokumentów z i-tej klasy, zatem $\sum_{i=1}^{n} w_i$ oznacza liczbę wszystkich klasyfikowanych dokumentów.

3. Klasyfikacja z użyciem metryk i miar podobieństwa tekstów

Nie wszystkie wartości w wektorze cech są od razu zapisane w formie liczbowej, niektóre z nich są w postaci tekstowej. Jak wspomniano w sekcji 1 numery cech dla cech tekstowych znajdują się w zbiorze oznaczonym w sprawozdaniu numerem (19).Należy zatem zaimplemetować miarę podobieństwa tekstów, która porówna dwie cechy tekstowe ze sobą, a następnie zamienić elementy wektora cech z postaci tekstowej na postać liczbową. W naszej aplikacji celu porównania cech tekstowych wykorzystaliśmy metodę n-gramów.[7] Metoda ta określa podobieństwo łańcuchów tekstowych s_1, s_2 w oparciu o ilość wspólnych podciągów n-elementowych. W zastosowanym przez nas algorytmie jako n przyjęliśmy liczbę 3, zatem posługiwaliśmy się trigramami. Wzór ogólny opisujący metodę n-gramów prezentuje się następująco:

$$sim_n(s_1, s_2) = \frac{1}{N-n+1} \sum_{i=1}^{N-n+1} h(i) \tag{31}$$

gdzie
$h(i) = 1$ jeśli n-elementowy podciąg zaczynający się od i-tej pozycji w słowie s_1 pojawia się przynajmniej raz w słowie s_2 (inaczej $h(i) = 0$);
$N(s_1), N(s_2)$ - ilość liter w słowach s_1 i s_2;
$N = max\{N(s_1), N(s_2)\}$;
$N - n + 1$ - ilość możliwych n-elementowych podciągów w s_1.

W naszym przypadku, dla trigramów wzór będzie wyglądał następująco:

$$sim_3(s_1, s_2) = \frac{1}{N-2} \sum_{i=1}^{N-2} h(i) \tag{32}$$

Weźmy na przykład porównanie dwóch słów: s_1=$MISSISIPI$, s_2=$MISSOURI$.
$N(s_1) = 9$, $N(s_2) = 8$, $N = max\{N(s_1), N(s_2)\} = 9$.

$$sim_3(s_1, s_2) = \frac{1}{7}\sum_{i=1}^{7} h(i) = \frac{2}{7} \approx 0.29 \tag{33}$$

ponieważ 2 trigramy z $MISSISIPI$ występują w $MISSOURI$.

Po wykonaniu metody n-gramów podobieństwo dwóch wzorców tekstowych zostaje wyrażone za pomocą liczby. Liczbę tą należy następnie zastosować w odpowiedniej metryce (wzory 35, 41, 46). W tym celu, do jednego z wektora cech, które porównujemy, w miejsce cechy tekstowej wpisujemy wynik poniższego wzoru:

$$w_\nu = 1 - sim_3(c^\mu_\nu, c^\eta_\nu)\ dla\ \nu \in T \tag{34}$$

gdzie c^μ oraz c^η oznaczają wektory cech, a ν to numer cechy, $sim_3(c^\mu_\nu, c^\eta_\nu)$ oznacza podobieństwo między dwoma cechami tekstowymi, T jest to zbiór numerów dla cech tekstowych (wzór nr 19)
W drugim wektorze natomiast w to miejsce wstawiamy 0. Tym sposobem obliczona wartość w metodzie n-gramów będzie mogła być wykorzystana w metryce. Dlatego dla dwóch porównywanych tekstów najpierw należy zastosować miarę podobieństwa, potem podmienić wartości tekstowe w wektorze cech na wartości liczbowe, a następnie obliczyć metrykę.

Jak już zostało wspomniane w rozdziale 2., algorytm k-NN należy do grupy algorytmów minimalnoodległościowych. Aby zaimplementować jego działanie niezbędna jest zatem metryka, która pozwoli w transparentny sposób wyznaczyć odległości pomiędzy wektorami cech. Z tego względu, w ramach badań zostaną wykorzystane trzy poniższe rodzaje metryk:

— Euklidesowa - aby obliczyć odległość pomiędzy dwoma wektorami cech należy obliczyć pierwiastek z sumy kwadratów różnic wszystkich kolejnych cech obu wektorów. Wzór opisujący metrykę Euklidesową wygląda następująco [1]:

$$\rho_1(c^\mu, c^\eta) = \begin{cases} \sqrt{\sum_{\nu=1}^{n_L}(c^\mu_\nu - c^\eta_\nu)^2} & dla\ \nu \in L \\ \sqrt{\sum_{\nu=1}^{n_T}(w_\nu)^2} & dla\ \nu \in T \end{cases} \tag{35}$$

gdzie n_L to liczba cech liczbowych, n_T - liczba cech tekstowych, ν - numer cechy, w_ν to obliczona zgodnie ze wzorem nr 34 wartość dla ν - tej cechy tekstowej. L jest to zbiór cech liczbowych (wzór nr 20), a T jest to zbiór cech tekstowych (wzór nr 19)

Przykład 3.1. *Obliczanie metryki Euklidesowej*
Dajmy dwa wektory cech:

$$c^\mu = [1, 2, 3] \tag{36}$$

$$c^\eta = [-2, -3, -4] \tag{37}$$

Wynik obliczenia metryki euklidesowej przedstawia się następująco:

$$\rho_1(c^\mu, c^\eta) = \sqrt{(1-(-2))^2 + (2-(-3))^2 + (3-(-4))^2} \tag{38}$$

$$\rho_1(c^\mu, c^\eta) = \sqrt{3^2 + 5^2 + 7^2} \tag{39}$$

$$\rho_1(c^\mu, c^\eta) \approx 9.11 \tag{40}$$

— uliczna - w celu obliczenia metryki ulicznej należy obliczyć sumę wartości bezwzględnych z różnic pomiędzy kolejnymi cechami z obu wektorów. Wzór opisujący metrykę uliczną wygląda następująco [1]:

$$\rho_2(c^\mu, c^\eta) = \begin{cases} \sqrt{\sum_{\nu=1}^{n_L} |c_\nu^\mu - c_\nu^\eta|} & dla\ \nu \in L \\ \sqrt{\sum_{\nu=1}^{n_T} |w_\nu|} & dla\ \nu \in T \end{cases} \tag{41}$$

Przykład 3.2. *Obliczanie metryki ulicznej*
Dajmy dwa wektory cech:

$$c^\mu = [3, 5, 8] \tag{42}$$

$$c^\eta = [-2, 0, 10] \tag{43}$$

Wynik obliczenia metryki ulicznej przedstawia się następująco:

$$\rho_2(c^\mu, c^\eta) = |3 - (-2)| + |5 - 0| + |8 - 10| \tag{44}$$

$$\rho_2(c^\mu, c^\eta) = 12 \tag{45}$$

— Czebyszewa - aby obliczyć metrykę Czebyszewa należy wyznaczyć maksymalną wartość bezwzględną z różnic pomiędzy kolejnymi cechami z obu wektorów. Wzór opisujący tę metrykę wygląda następująco [1]:

$$\rho_3(c^\mu, c^\eta) = \max_{1 \le \nu \le n} \begin{cases} |c_\nu^\mu - c_\nu^\eta| & dla\ \nu \in L \\ |w_\nu| & dla\ \nu \in T \end{cases} \tag{46}$$

gdzie: n - liczba wszystkich cech
Przykład 3.3. *Wykorzystanie metryki Czebyszewa*
Dajmy dwa wektory cech:

$$c^\mu = [6, 3, 5, -1] \tag{47}$$

$$c^\eta = [3, 4, 1, 9] \tag{48}$$

Wynik obliczenia metryki ulicznej przedstawia się następująco:

$$\rho_3(c^\mu, c^\eta) = max|6 - 3|, |3 - 4|, |5 - 1|, | - 1 - 9| \tag{49}$$

$$\rho_3(c^\mu, c^\eta) = 10 \tag{50}$$

Poniżej przedstawiamy wstępne wyniki miary Accuracy dla próbnych klasyfikacji na ograniczonym zbiorze tekstów (klasyfikacje przeprowadziliśmy na zbiorze 1291 tekstów z 13531 możliwych):
— Pierwszą klasyfikację przeprowadziliśmy dla następujących parametrów:

  - K najblższych sąsiadów: 5
  - Proporcja podziału zbioru: 80 - treningowy, 20 - testowy
  - Metryka: Uliczna
  - Wybrane cechy: Wszystkie

  Miara Accuracy w tym przypadku wyniosła: 0,84
- Przy drugiej próbnej klasyfikacji użyliśmy poniższych parametrów:
  - K najblższych sąsiadów: 6
  - Proporcja podziału zbioru: 75 - trenigowy/25 - testowy
  - Metryka: Czebyszewa
  - Wybrane cechy: 1. Częstość słów pisanych wielkimi literami, 2. Ogólna liczba słów, 3. Częstość występowania cytatów, 4. Częstość słów rozpoczynających się wielką literą, 5. Najczęściej występujące słowo, 6. Zapis cyfr, 7. Jednostki w układzie SI/imperialnym

  Miara Accuracy w tym przypadku wyniosła: 0,83
- Dla trzeciej klasyfikacji parametry były następujące:
  - K najblższych sąsiadów: 4
  - Proporcja podziału zbioru: 70 - treningowy/30 - testowy
  - Metryka: Euklidesowa
  - Wybrane cechy: 1. Ogólna liczba słów, 2. Częstość występowania dat, 3. Częstość słów rozpoczynających się wielką literą, 4. Częstość słów pisanych wielkimi literami, 5. Jednostki w układzie SI/ imperialnym

  Miara Accuracy w tym przypadku wyniosła: 0,75
- Natomiast dla czwartej, ostatniej klasyfikacji użyliśmy poniższych parametrów:
  - K najblższych sąsiadów: 3
  - Proporcja podziału zbioru: 60 - treningowy/40 - testowy
  - Metryka: Euklidesowa
  - Wybrane cechy: 1. Format zapisu dat, 2. Słowa kluczowe, 3. Najczęściej występująca waluta, 4. Najczęściej występujące słowo.

  Miara Accuracy w tym przypadku wyniosła: 0,82

Najwyższy wynik miary Accuracy spośród wstępnych klasyfikacji otrzymaliśmy dla pierwszej klasyfikacji, a zatem dla 5 najblższych sąsiadów, proporcji podziału zbioru 75/25, metryce Czebyszewa i dla wszystkich cech. Wyniki były jednak bardzo podobne, a zbiór artykułów na którym działaliśmy był mocno ograniczony. Poza tym 4 klasyfikacje to zdecydowanie za mało, aby wyciągnąć wnioski dotyczące wpływu określonych parametrów na wynik miary Accuracy.

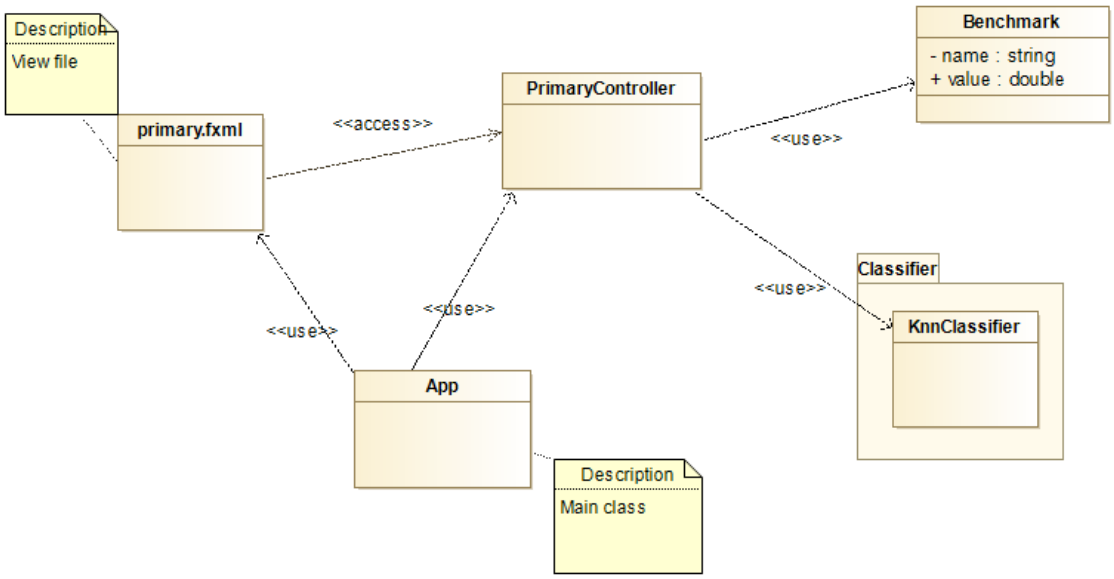
4. Budowa aplikacji

4.1. Diagramy UML

Nasza aplikacja będzie się składać z 4 modułów: moduł ekstrakcji, moduł klasyfikatora, moduł DAO (zarządzanie wczytywaniem danych z plików) oraz moduł interfejsu graficznego.

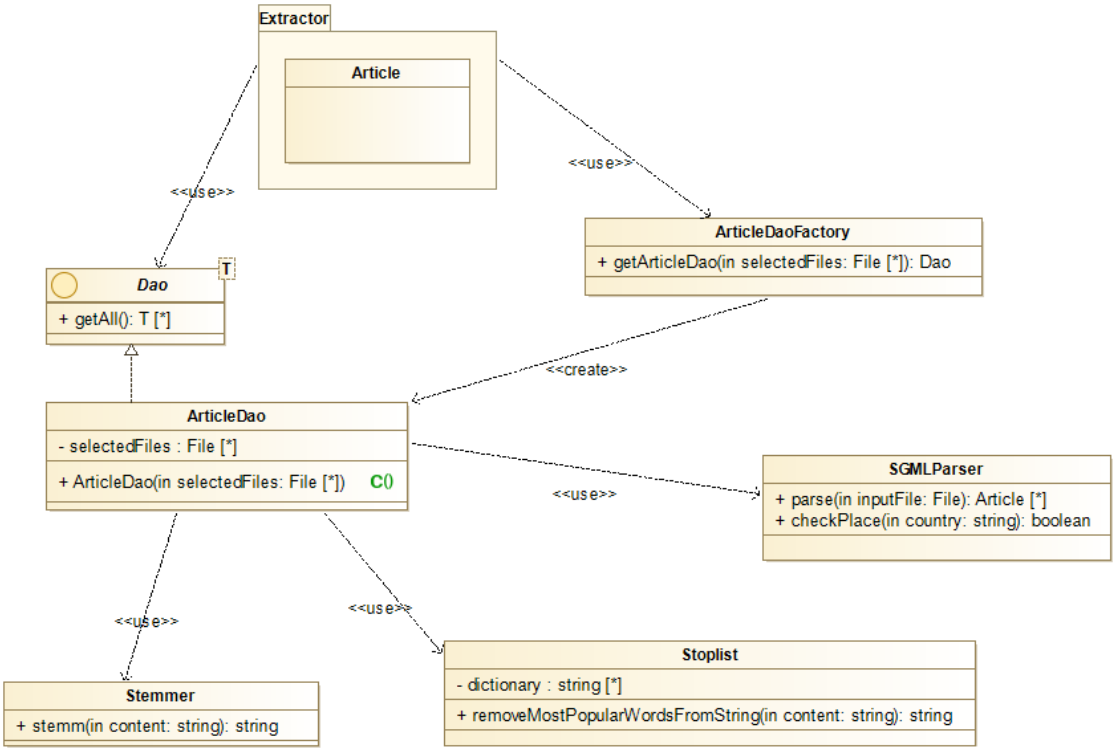
Moduł GUI został zaimplementowany zgodnie ze wzorcem projektowym

MVC (Model-View-Controller) i odpowiada za prezentację danych w ramach interfejsu graficznego. Model stanowi najważniejszą warstwę aplikacji, ponieważ w tej części zaimplementowany został algorytm kNN. Z kolei zadaniem części View (plik z rozszerzeniem fxml) jest wyświetlenie informacji w oknie aplikacji, natomiast Controller odpowiada za połączenie modelu aplikacji z widokiem.



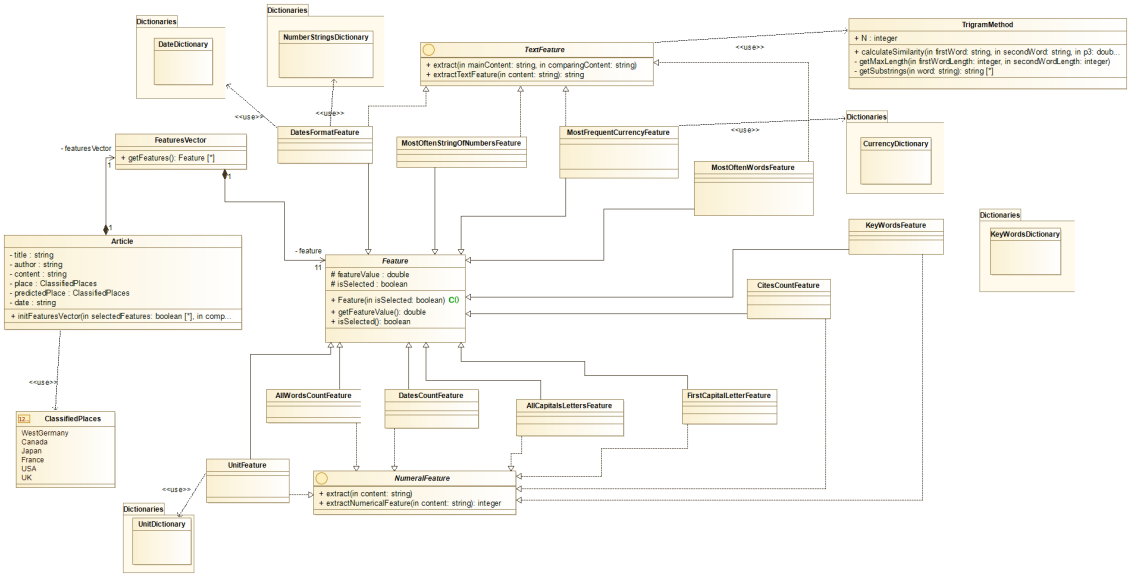
Rysunek 1. Diagram UML dla modułu GUI

Zadaniem modułu DAO jest umożliwienie wczytania plików z artykułami, które zostaną poddane klasyfikacji. Implementacja tego modułu została przeprowadzona zgodnie ze wzorcem projektowym DAO, dzięki czemu dostarczony zostaje jednolity interfejs do komunikacji między aplikacją, a źródłem danych. Poza wczytywaniem plików, w module DAO następuje także przygotowanie pliku do przetwarzania, tj. tekst zostaje poddany procesowi stematyzacji (klasa Stemm) oraz zostają usunięte najbardziej popularne słowa z języka angielskiego (klasa Stoplist), co ma na celu podniesienie jakości klasyfikacji.



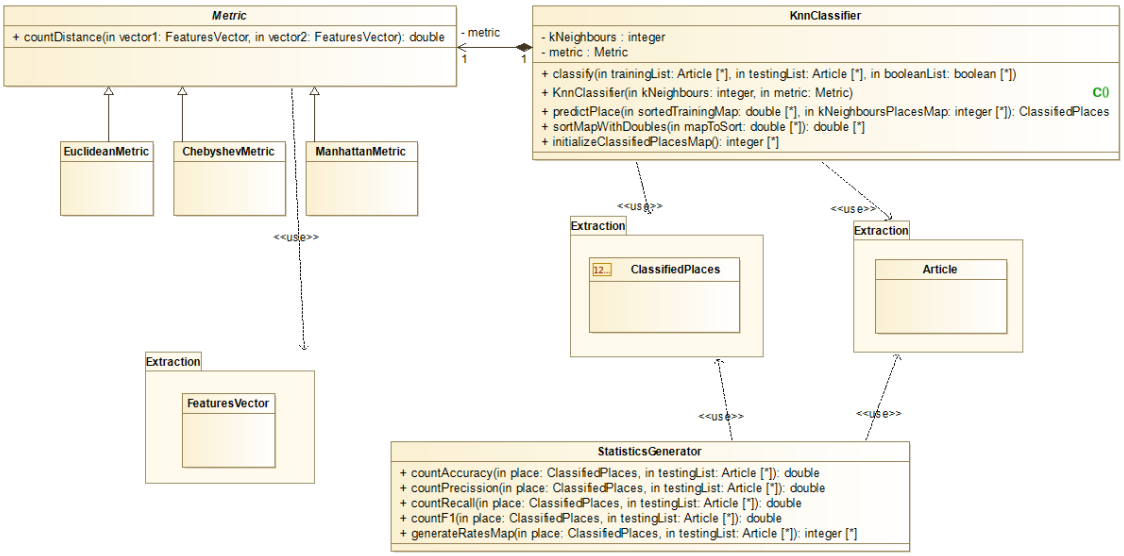
Rysunek 2. Diagram UML dla modułu DAO

Moduł ekstrakcji będzie odpowiedzialny za odwzorowanie tekstu na wektor cech. Wektor cech zaimplementowany zostanie z wykorzystaniem klasy FeaturesVector. W klasie FeaturesVector znajduje się 11 pól, każde oznaczające jedną z cech, które zostały przez nas wybrane i przedstawione w sekcji 2. Dla każdej cechy stworzona została odpowiadająca klasa. Każda z klas, która reprezentuje cechę, implementuje interfejs: dla cech liczbowych jest to interfejs NumericFeature, natomiast dla cech tekstowych jest n to interfejs TextFeature. Oba interfejsy mają metodę extract(Article article), różnica polega na zwracanej wartości. W pierwszym przypadku jest to liczba typu integer (oznaczająca obliczoną liczbę np. słów w tekście), zaś w drugim przypadku jest to lista ciągów znaków (oznaczająca wyznaczoną, znalezioną listę wyrazów, które spełniają warunki danej klasy).



Rysunek 3. Diagram UML dla modułu ekstrakcji

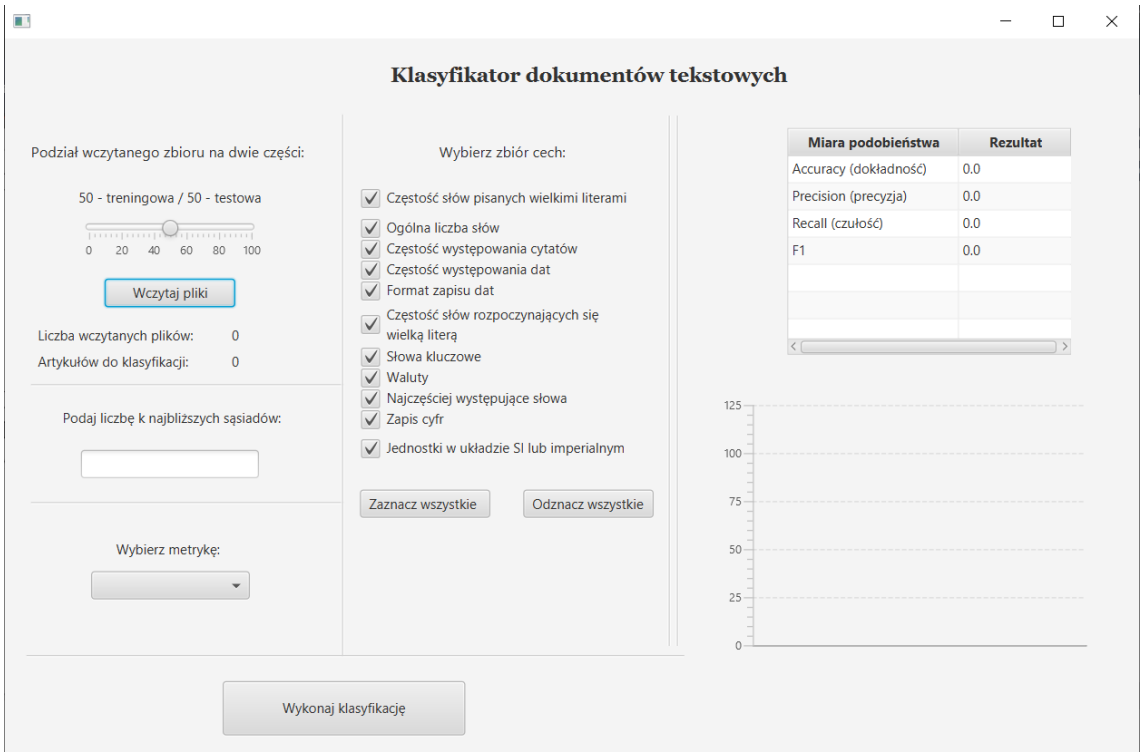
Moduł klasyfikatora będzie odpowiedzialny za klasyfikację artykułów do odpowiednich etykiet places przy pomocy algorytmu kNN. Z tego względu powstanie klasa KnnClassifier, której zadaniem będzie sklasyfikowanie artykułu z wykorzystaniem jednej z trzech metryk, tj. metryki Czebyszewa, metryki ulicznej (Manhattan) lub metryki euklidesowej. W tym celu powstały cztery klasy: klasa Metric jest to klasa abstrakcyjna, natomiast klasy EuclideanMetric, ChebyshevMetric oraz ManhattanMetric są klasami dziedziczącymi. Poza metryką, niezbędne do klasyfikacji są paramtery: liczba najbliższych sąsiadów (kNeighbours) oraz stosunek liczby artykułów w części treningowej do części testowej. Do wykorzystania klasy KnnClassifier niezbędny jest moduł ekstrakcji, ponieważ klasa KnnClassifier wykorzystuje klasę Artykuł oraz kategorie ClassifiedPlaces znajdujące się w typie enumerate.



Rysunek 4. Diagram UML dla modułu klasyfikatora

4.2. Prezentacja wyników, interfejs użytkownika

Aplikacja została wykonana w technologii Java w wersji 11[8] (najnowsza wersja LTS) przy wykorzystaniu Apache Maven w wersji 3.6.3[9]. Do stworzenia interfejsu graficznego posłużyliśmy biblioteką JavaFX w wersji 13[10]. W celu uruchomienia aplikacji należy po zainstalowaniu Maven na własnym kpmputerze, z poziomu wiersza poleceń znajdując się w folderze głównym projektu wykonać polecenie: mvn install, a następnie z wiersza poleceń z poziomu modułu GUI wykonać polecenie: mvn clean javafx:run. Aplikacje można także uruchomić z poziomu IDE. Po uruchomieniu aplikacji ukaże nam się interfejs użytkownika, w którym możemy wybrać jak ma zostać podzielony zbiór (w jakich proporcjach na część treningową i testową), wczytać pliki, w których znajdują się teksty do analizy, podać licbę k najbliższych sąsiadów dla klasyfiaktora k-NN, wybrać metrykę oraz zbiór cech wykorzystywanych w procesie klasyfikacji oraz wykonać klasyfikację. W efekcie ukażą nam się następujące informacje: liczba wczytanych plików, liczba artykułów podlegających klasyfikacji oraz rezultaty 4 miar podobieństwa - Accuracy, Precision, Recall i F1, a także wykres słupkowy pokazujący liczbę artykułów w każdej klasie oraz liczbę artykułów przyporządkowanej do danej klasy.



Rysunek 5. Interfejs użytkownika dla aplikacji.

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.

Załączniki

1. Załącznik nr 1 - Zbiór słowników wykorzystywanych przy ekstrakcji cech.

Literatura

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

[2] Corpus of Contemporary American English: Most frequent english words [przeglądany 20.03.2021], Dostępny w: https://www.english-corpora.org/

[3] Repozytorium Uniwersytety Kalifornijskiego w Irvine do nauki uczenia maszynowego: Artykuły agencji Reuters[przeglądany 20.03.2021], Dostępny w: http://archive.ics.uci.edu/ml/machine-learning-databases/reuters21578-mld/

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

[5] Tablica pomyłek [przeglądany 28.03.2021], Dostępny w: https://pl.wikipedia.org/wiki/Tablica_pomyłek

[6] F-score [przeglądany 28.03.2021] Dostępny w: https://en.wikipedia.org/wiki/F-score

[7] A. Niewiadomski, Materiały, przykłady i ćwiczenia do przedmiotu Komputerowe Systemy Rozpoznawania, 2020.

[8] Dokumetacja Java 11 [przeglądany 11.04.2021] Dostępny w : https://docs.oracle.com/en/java/javase/11/

[9] Dokumentacja Maven 3.6.3 [przeglądany 11.04.2021] Dostępny w: https://maven.apache.org/docs/3.6.3/release-notes.html

[10] Dokumentacja JavaFx 13 [przeglądany 11.04.2021] Dostępny w: https://openjfx.io/javadoc/13/

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

Załącznik 1.

Zbiór słowników wykorzystywanych przy ekstrakcji cech

- **Zbiór wszystkich słów oznaczających waluty (wraz z skrótami):** [dolar, dlr, cent, german mark, penny, deutschmark, pfennig, french franc, canadian dollar, yen, sen, rin, pound sterling, str]
    - **Dla USA:** [dollar, dlr, cent]
    - **Dla RFN:** [german mark, deutschmark, pfennig]
    - **Dla Japonii:** [yen, sen, rin]
    - **Dla Francji:** [cent, french franc]
    - **Dla Kanady:** [canadian dolar, canadian cent]
    - **Dla Wielkiej Brytanii:** [pound sterling, penny, str]
- **Zbiór jednostek imperialnych (wraz ze skrótami) – używane w USA i Wielkiej Brytanii:** [inch, foot , ft, yard, yd, mile, mi, ounce, oz, galon, gal, quart, qt, stone, st, pound, lb, ton, Fahrenheit ,°F]
- **Zbiór jednostek układu SI (wraz ze skrótami) – używane w RFN, Japonii, Francji i Kanadzie:** [gram, g , kilogram, kg, meter, m, centimeter, cm, Celsius,°C]
- **Zbiór numerów kierunkowych:** [+1, +001, +44, +0044, +49, +0049, +33, +0033, +81, +0081]
- **Zbiór formatów dat:** [RRRR-MM-DD, DD.MM.RRRR, DD-MM-RRRR, DD/MM/RRRR, MM.DD.RRRR, MM/DD/RRRR, MM-DD-RRRR, , "Miesiąc Dzień", "Dzień Miesiąc", "Dzień Miesiąc Rok", "Miesiąc Dzień Rok", „Dzień{th} Miesiąc”, „Dzień{th} of Miesiąc Rok”, „Dzień{th} of Miesiąc, Rok”]
  Daty możemy zapisać w formie numerycznej (np. 1980-02-13) lub słownej (np. "ósmy sierpnia" lub "August 8"). Przy zapisie słownym należy pamiętać, że dzień może być zapisany liczbą lub słownie. W powyższym zbiorze R oznacza cyfry w zapisie lat, M - cyfry w zapisie miesięcy, natomiast D - cyfry w zapisie dni. Zapis umieszczony w cudzysłowie oznacza zapis słowny daty (czyli dla "Miesiąc Dzień Rok" jest to np. "July 5 1992"). Z tego wniosek, że zapis słowny odnosi się zarówno do dat zapisanych w całości słownie, jak i takich, w których tylko część daty jest zapisana słownie, a pozostała część za pomocą cyfr).
- **Zbiory słów kluczowych dla poszczególnych krajów:**
    - **Dla USA:**
      [Alabama
      Alaska
      Arizona
      Arkansas
      California
      Colorado
      Connecticut
      Delaware
      Florida
      Georgia
      Hawaii
      Idaho
      Illinois
      Indiana

Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Missisipi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virginia
Washington
West Virginia
Wisconsin
Wyoming

New York
Los Angeles
Chicago
Houston
Phoenix
Philadelphia
San Antonio

San Diego
Dallas
San Francisco

Washington
Jefferson
Lincoln
Roosevelt
Wilson
Truman
Eisenhower
Kennedy
Nixon
Carter
Reagan

Smith
Johnson
Williams
Jones
Brown
Davis
Miller
Wilson
Moore
Taylor]

- **Dla Kanady:**

[Alberta
British Columbia
Manitoba
New Brunswick
Newfoundland and Labrador
Northwest Territories
Nova Scotia
Nunavut
Ontario
Prince Edward Island
Quebec
Saskatchewan
Yukon

Toronto
Montreal
Vancouver

Calgary
Edmonton
Ottawa
Winnipeg
Quebec
Hamilton

Douglas
Trudeau
Fox
Banting
Macdonald

Smith
Brown
Tremblay
Martin
Roy
Gagnon
Lee
Wilson
Johnson
MacDonald]

- **Dla Japonii:**

[Hokkaido
Honshu
Kyushu
Shikoku
Okinawa

Tokyo
Yokohama
Osaka
Nagoya
Sapporo
Kobe
Kyoto
Fukuoka
Kawasaki
Hiroshima

Mutsuhito
Yoshihito

Hirohito
Miki
Fukuda
Ōhira
Nakasone
Takeshita

Sato
Suzuki
Takahashi
Tanaka
Watanabe
Ito
Nakamura
Kobayashi
Yamamoto
Kato]

- **Dla Wielkiej Brytanii:**

[England
Scotland
Northern Ireland
Wales
Yorkshire

London
Birmingham
Leeds
Glasgow
Sheffield
Manchester
Edinburgh
Liverpool
Bristol
Cardiff

Queen Elisabeth II
Princess Diana
Churchill
Thatcher
Callaghan
King Phillip

Smith
Jones

Williams
Brown
Taylor
Davies
Wilson
Evans
Thomas
Roberts]

- **Dla Francji**

[Auvergne
Rhône-Alpes
Burgundy
Franche-Comté
Brittany
Centre-Val de Loire
Corsica
French Guiana
Alsace
Champagne-Ardenne
Lorraine
Guadeloupe
Nord-Pas-de-Calais
Picardy
Île-de-France
Martinique
Mayotte
Lower Normandy
Upper Normandy
Aquitaine
Limousin
Poitou-Charentes
Languedoc-Roussillon
Midi-Pyrénées
Pays de la Loire
Provence-Alpes-Côte d'Azur
Réunion

Paris
Lyon
Marseille
Toulouse
Lille
Bordeaux
Nice

Nantes
Strasbourg
Cannes

de Gaulle
Poher
Pompidou
Mitterrand
Chirac

Martin
Bernard
Dubois
Thomas
Robert
Richard
Petit
Durand
Leroy
Moreau]

- **Dla RFN:**
  [\taBayern
  Hessen
  Schleswig-Holstein
  Niedersachsen
  Nordrhein-Westfallen
  Rhein-Pfalz
  Saarland
  Baden-Wurttemberg

  Berlin
  Bremen
  Dortmund
  Bonn
  Frankfurt
  Hamburg
  Stuttgart
  Koln
  Dusseldorf
  Munchen

  Adenauer
  Brandt
  Kohl

Muller
Schmidt
Schneider
Fischer
Weber
Meyer
Wagner
Becker
Schulz
Hoffmann]