Informatyka, studia dzienne, II st. semestr II

Komputerowe systemy rozpoznawania 2013/2014

Prowadzący: dr hab. inż. Adam Niewiadomski, prof. PŁ poniedziałek, 10:30

Data oddania: _____ Ocena: _____

Mateusz Grotek 186816
Paweł Tarasiuk 186875

Zadanie 1.: Ekstrakcja cech, miary podobieństwa, klasyfikacja

1. Cel

Celem niniejszego zadania jest zbadanie różnych metod ekstrakcji cech oraz miar podobieństwa dla tekstu i zastosowanie ich w procesie klasyfikacji słabym klasyfikatorem kNN. Omówione zostaną metody znane w literaturze oraz nasze własne pomysły, a wszystkie te elementy znajdą swoje odzwierciedlenie w przygotowanej przez nas implementacji, co umożliwi nam wygenerowanie i ocenę wyników działania różnych metod.

2. Wprowadzenie

2.1. Podstawowe algorytmy

Klasyfikator kNN (k nearest neighbours, k najbliższych sąsiadów) to słaby klasyfikator, posiadający jednak istotną zaletę jaką jest prostota implementacji. Słabość klasyfikatora rozumiemy tak jak w teorii boostingu – słabym nazywamy taki klasyfikator, który nie zawsze jest w stanie osiągnąć dowolnie wysoką dokładność klasyfikacji. Dlatego jest on szeroko stosowany, gdyż pomimo swojej prostoty daje relatywnie dobre wyniki. Dodatkowo można go użyć w technikach typu AdaBoost, które pozwalają z kilku słabych klasyfikatorów stworzyć jeden silny. Jego idea polega na wybraniu takiej kategorii która jest najczęstsza wśród sąsiadów wektora podlegającego klasyfikacji. Potencjalni sąsiedzi są dostarczani do klasyfikatora na etapie jego uczenia.

Aby wskazać najbliższych sąsiadów należy oczywiście użyć jakiejś miary odległości, czyli metryki. Istnieje też odmiana klasyfikatora, którą moglibyśmy określić jako k najbardziej podobnych sąsiadów. W tej odmianie zamiast odległości w zadanej metryce używana jest funkcja podobieństwa. Pokazuje to jak szerokie zastosowanie może mieć ten klasyfikator.

W naszym projekcie użyliśmy następujących funkcji podobieństwa:

— miara Jaccarda na słowach po filtracji stoplistą i stemizacji jako podobieństwo tekstów

$$J(A,B) = \frac{\mu(A \cap B)}{\mu(A \cup B)} \tag{1}$$

— metoda n-gramów (użyliśmy metody 3-gramów, gdyż tekst jest tekstem angielskim, przy czym użyliśmy 3-gramów w mierze podobieństwa zdań)

$$\mathrm{sim}_3(s_1,s_2) = \frac{1}{N-2} \sum_{i=1}^{N-2} h(i) \tag{2}$$

$$\mu_{N_z}(z_1, z_2) = \frac{1}{N} \sum_{i=1}^{N(z_1)} \max_{j=1,\ldots,N(z_1)} \mathrm{sim}_3(s_{1j}, s_{2i}) \tag{3}$$

— autorska metoda bazująca na słowach kluczowych i prostym współczynniku dopasowania (patrz sekcja 2.5)

Użyte zostały następujące metryki:

— metryka euklidesowa

$$d_e(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

— metryka uliczna

$$d_1(X,Y) = \sum_{i=1}^{n} |x_i - y_i| \tag{5}$$

— metryka Czebyszewa

$$d_\infty(X,Y) = \max_{i=1,\ldots,n} |x_i - y_i| \tag{6}$$

Klasyfikacji można dokonywać bezpośrednio na podanych tekstach, jednak takie rozwiązanie charakteryzuje się wysokim kosztem obliczeniowym. Dlatego często lepiej jest wyekstrahować ze zbiorów pewne wektory cech, które pozwolą szybko sklasyfikować dane teksty. Wybór wektorów cech zależy od postawionego zadania. W naszym projekcie użyliśmy trzech sposobów ekstrakcji cech:

1. naiwny sposób bazujący na użyciu wszystkich słów (po filtracji przez stoplistę i stemizacji), ilość wymiarów wektora jest zależna od ilości wszystkich słów w zbiorze uczącym (patrz sekcja 2.2)
2. bazujący na słowach kluczowych (jak powyżej, ale wybierane są tylko słowa kluczowe, co zmniejsza ilość wymiarów) (patrz sekcja 2.3)
3. autorska metoda bazująca na zbiorach rozmytych (opis poniżej) (patrz sekcja 2.4)

Sama klasyfikacja została wykonana na dwóch zbiorach danych, każdy podzielony na podzbiór uczący i podzbiór testowy, przy czym sposób podziału jest wybierany w aplikacji. Można dokonać podziału w sposób losowy (przy zadanych proporcjach), lub w ustalony sposób (60% początkowych tekstów jako zbiór uczący, reszta to zbiór testowy). Użyliśmy następujących zbiorów danych:

1. zestaw krótkich wiadomości prasowych firmy Reuters,
2. przygotowany przez nas zestaw fragmentów tekstów znanych pisarzy.

Teksty prasowe zostały sklasyfikowane na podstawie dwóch kategorii: miejsca którego dotyczą i tematu. Jeśli chodzi o teksty znanych pisarzy, to zostały one sklasyfikowane kategorią, którą są nazwiska pisarzy.

2.2. Naiwna metoda ekstrakcji cech

Do celów porównawczych zaimplementowaliśmy naiwną metodę ekstrakcji. Metoda ta działa w następujący sposób. Najpierw teksty są dzielone na słowa. Słowa te podlegają prostej stemizacji, a następnie filtrowane są przez stop-listę. Otrzymujemy zestaw skończonych ciągów słów. Każdemu słowu z zestawu (w sensie typu, a nie egzemplarza słowa) przypisywana jest pozycja w wektorze (w kolejności występowania słów). Następnie dla każdego tekstu budowany jest wektor w którym na kolejnych pozycjach występują ilości znalezionych słów.

2.3. Metoda ekstrakcji cech bazująca na słowach kluczowych

Ta metoda działa podobnie do powyższej, jednakże zamiast użycia wszystkich słów, zostały użyte tylko słowa kluczowe. Dla notatek prasowych użyliśmy zestawu słów kluczowych znajdujących się w kolekcji. Dla autorów opracowaliśmy automatycznie własny zestaw słów kluczowych. Dzięki użyciu słów kluczowych zmniejszyła się ilość wymiarów w wektorze cech.

2.4. Autorska metoda ekstrakcji cech

Poniżej omówimy autorskie metody ekstrakcji cech i miarę podobieństwa, których użyliśmy. Główną ideą naszej metody ekstrakcji są słowa kluczowe. Zbiór takich słów należy przygotować dla konkretnego zadania. Może on być wyznaczony ręcznie, lub wygenerowany automatycznie na podstawie danych uczących. Zbiór takich słów dzielimy na podzbiory na podstawie ich znaczenia. Można oczywiście używać tylko jednego podzbioru w którym są umieszczone wszystkie słowa kluczowe, ale zwiększenie ilości podzbiorów może poprawić klasyfikację, ze względu na zwiększenie wymiarowości przestrzeni. W jednym podzbiorze powinny być słowa oznaczająca podobne obiekty, na przykład nazwy geograficzne. Następnie dla każdej kategorii względem której klasyfikujemy przygotowujemy osobny zestaw podzbiorów. Inaczej mówiąc jeżeli kategorią jest „kanada” a w jednym z podzbiorów znajdują się „organizacje”, to tworzymy podzbiór „kanadyjskie organizacje”. Widać, że jest to metoda analogiczna do tego jak ludzie kategoryzują obiekty. Wystarczy następnie zauważyć, że określenie „kanadyjskie organizacje” możemy modelować jako

pewien zbiór rozmyty. Niektóre organizacje są stricte kanadyjskie, inne tylko częściowo. Na podstawie danych uczących, lub wiedzy eksperckiej jesteśmy w stanie podać funkcję przynależności do danego zbioru.

Mając taki zestaw zbiorów możemy wyekstrahować cechy z tekstów zbioru uczącego i na tej podstawie nauczyć klasyfikator. Ekstrakcja przebiega następująco. Wstępnie każdy element wektora cech jest liczbą, która jest sumą wartości funkcji przynależności dla każdego słowa, kluczowego z podanego zbioru rozmytego, które wystąpiło w podanym tekście. Opisuje to wzór

$$V_T(n) = \sum_{s \in K(n) \cap T} \mu_{K(n)}(s) \tag{7}$$

gdzie:
— T to tekst (jako zbiór słów) z którego ekstrahujemy cechy,
— n to numer składowej wektora cech którą obliczamy,
— $K(n)$ to n-ty zbiór rozmyty słów kluczowych,
— $\mu_{K(n)}(s)$ to funkcja przynależności słowa s do zbioru $K(n)$.
Następnie wektor V_T jest normalizowany do długości równej 1 według wzoru

$$V'_T = \frac{V_T}{|V_T|} \tag{8}$$

Przykładowo, załóżmy, że mamy 2 teksty postaci:

T1 Showers continued throughout the week in the Bahia cocoa zone. Ms Smith said there is still some doubt as to how much old crop cocoa is still available as harvesting has practically come to an end.

T2 "I'd put it off as long as they conceivably could," said Lawrence Cohn, analyst with Merrill Lynch, Pierce, Fenner and Smith.

Załóżmy także, że tekst 1 jest przyporządkowany do kategorii „el-salvador", a tekst 2 do kateogrii „usa". Do dyspozycji mamy 3 słowa (frazy) kluczowe: „bahia", „smith", „merill lynch". Załóżmy także, że mamy dwa następujące zbiory rozmyte słów (fraz) kluczowych:

$$\begin{aligned}
\text{elSalvadorWords} &= \{ \\
&\quad < \text{bahia}; 1{,}0 >, \\
&\quad < \text{smith}; 0{,}2 >, \\
&\quad < \text{merill lynch}; 0{,}0 > \\
&\} \\
\text{usaWords} &= \{ \\
&\quad < \text{bahia}; 0{,}0 >, \\
&\quad < \text{smith}; 0{,}7 >, \\
&\quad < \text{merill lynch}; 0{,}9 > \\
&\}
\end{aligned}$$

Naszym zadaniem jest utworzyć dwa wektory cech opisujących powyższe teksty. Zgodnie z powyższymi wzorami Wektory te będą wyglądać następująco:

$$\begin{aligned}
V_{T1} &= < 1{,}2; 0{,}7 > \\
V_{T2} &= < 0{,}2; 1{,}6 >
\end{aligned}$$

a po normalizacji:

$$\begin{aligned} V'_{T1} &\approx < 0{,}86; 0{,}50 > \\ V'_{T2} &\approx < 0{,}12; 0{,}99 > \end{aligned}$$

2.5. Autorska funkcja podobieństwa

Natomiast jeśli chodzi o obliczanie funkcji podobieństwa, metoda ta również opiera się na słowach kluczowych. Dla słów kluczowych, które znajdują się w obu tekstach, liczony jest prosty współczynnik dopasowania (simple matching coefficient), wyznaczony wzorem

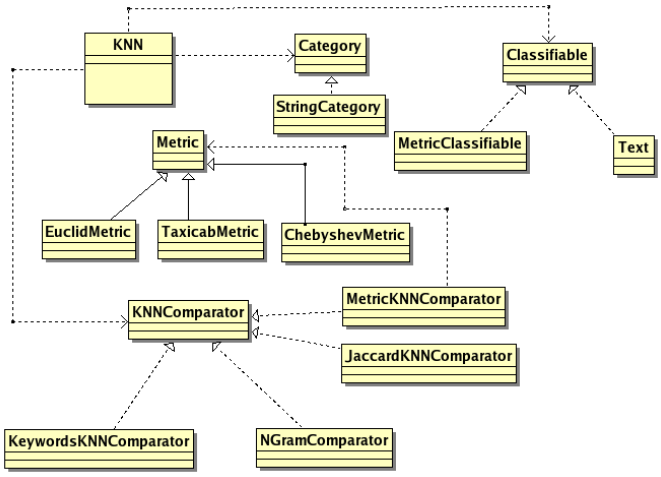
$$SMC(A,B) = \frac{\mu(A \cap B) + \mu(A^C \cap B^C)}{\mu(X)} \tag{9}$$

Dzięki użyciu słów kluczowych jesteśmy w stanie znaleźć przestrzeń X, gdyż jest to po prostu zestaw wszystkich słów kluczowych. Na przykład dla powyższych tekstów i zbiorów rozmytych mamy następującą wartość funkcji:

$$SMC(T1,T2) = \frac{1}{3}$$

3. Opis implementacji

Implementacja programu została wykonana w języku Java. Poniżej prezentujemy orientacyjny wycinek diagramu klas naszego programu.



Rysunek 1. Wycinek uproszczonego diagramu klas aplikacji

Przygotowane rozwiązanie zawiera w sobie także interfejs graficzny wykonany przy wykorzystaniu biblioteki Swing. Wybór ten został podyktowany tym, że biblioteka ta jest domyślnie obecna w środowisku uruchomieniowym Javy, dzięki czemu nasz projekt nie ma żadnych dodatkowych zewnętrznych zależności. Interfejs pozwala wybrać przetwarzany zestaw danych, dopasować parametry wskazywania zbioru uczącego i testowego oraz daje dostęp

do wszystkich dostępnych w projekcie metod i parametrów ekstrakcji cech, porównywania próbek i klasyfikacji. Po ustaleniu wszystkich właściwości zadanego problemu program wyświetla uogólnioną informację o tym, jaka część próbek została sklasyfikowana poprawnie oraz pozwala zapisać szczegółowe wyniki w plikach csv. Dostępne są trzy tryby generowania szczegółowych wyników:

— Raport z poszczególnych próbek ze zbioru testowego (do jakiej kategorii powinny należeć, a do jakiej zostały zaklasyfikowane)
— Statystyki TPR (jaka część próbek które powinny do niej należeć faktycznie do niej trafiła) i PPV (jaka część próbek zaklasyfikowanych do danej kategorii faktycznie do niej należy) opisujące każdą z kategorii
— Macierz, w której dla każdej pary właściwa kategoria/kategoria wskazana przez klasyfikator przechowywane są liczby przypadków w których tak się stało

4. Materiały i metody

Eksperymenty przeprowadzane były dla trzech problemów klasyfikacji:

1. **Kraje** – klasyfikacja próbek tekstowych ze zbioru *Reuters-21578 Text Categorization Collection Data Set*, gdzie jako etykiety wybrane zostały kraje, których dotyczy dana próbka. Wybrane zostały wyłącznie te próbki, które dotyczą dokładnie jednego kraju i w przypadku których krajem tym jest Kanada, Francja, Japonia, Wielka Brytania, USA lub RFN. W przypadku tego problemu mamy do dyspozycji **13441** próbek, w których (po uproszczonej stemizacji i usunięciu słów nieznaczących) występuje **28145** różnych słów.
2. **Tematy** – próbki wybierane z tego samego zbioru, co w przypadku problemu *Kraje*, jednakże tym razem klasami względem których chcemy rozróżniać próbki są tematy artykułów. Wzięte pod uwagę zostały te próbki, których lista tematów zawiera temat *interest* albo temat *grain* (próbki dotyczące obu tych tematów jednocześnie były pomijane). Problem klasyfikacji polega zatem na podziale zbioru wybranych próbek na dwie klasy, odpowiadające charakterystycznym tematom. W przypadku tego problemu mamy do dyspozycji **998** próbek, w których występuje **7533** rozróżnialnych, uwzględnianych przez nasz program słów.
3. **Autorzy** – tym razem wykorzystaliśmy własny zbiór tekstów, przygotowany w oparciu o dane przygotowane w ramach inicjatywy *Project Gutenberg*. Z danych tekstowych z plików *The Complete Works of William Shakespeare* [2] oraz *The Works of Lord Byron, Vol. 4* [3] usunięte zostały przypisy, a następnie wybrane zostały ciągłe próbki zawierające po kilka zdań lub wersów, tak aby każda z nich zawierała od 50 do 200 słów. Kryterium klasyfikacji tak przygotowanych próbek jest autor tekstu (po usunięciu przypisów, próbki wybierane są wprost z dzieł literackich, więc autorem każdej z próbek jest William Shakespeare lub Lord George Gordon Byron). W przypadku tego problemu mamy do dyspozycji **568** próbek, w których występuje **6028** rozróżnialnych, uwzględnianych przez nasz program słów.

Dla każdego ze przedstawionych powyżej problemów testowaliśmy różne podejścia do wyboru zbioru uczącego oraz zbioru sprawdzającego. Domyślnym ustawieniem było potraktowanie 60% początkowych próbek jako zbioru uczącego, zaś pozostałych 40% próbek – jako zbioru sprawdzającego. Oceniane było także zachowanie na mniejszych zbiorach – wtedy zbiór uczący wybierany był zawsze od początku, zaś zbiór sprawdzający – od końca (podanie odpowiedniego argumentu w przygotowanej przez nas implementacji powoduje, że dozwolone jest nachodzenie na siebie tych dwóch zbiorów, zatem suma ich miar może przekroczyć liczbę wszystkich próbek w danym problemie). Dodatkowa opcja pozwala spowodować, aby przy części testów wybierać próbki w sposób pseudolosowy (zamiast stosowania schematu opisanego powyżej).

Kluczowym elementem testów jest porównanie różnych miar podobieństwa próbek (bądź równoważnie – różnych funkcji odległości między próbkami). W zastosowanych przez nas metodach nacisk położony był albo na ekstrakcję cech (po której można było zastosować metody porównania oparte na różnych metrykach w $\mathbb{R}^n$) albo na zaawansowane miary podobieństwa. Możliwości obejmowały:

- W przypadku zastosowania metod opartych na metrykach w $\mathbb{R}^n$
  - Wybór metody ekstrakcji cech:
    - poprzez obliczanie częstości występowania poszczególnych słów: *„wszystko”* (patrz sekcja 2.2)
    - poprzez obliczanie częstości występowania wybranych wcześniej słów kluczowych: *„wybr. sł.”* (patrz sekcja 2.3)
    - poprzez wykorzystanie opisanej w poprzednich sekcjach metody wykorzystującej zbiory rozmyte: *„zb. rozm.”* (patrz sekcja 2.4)
  - Wybór odległości na $\mathbb{R}^n$:
    - euklidesowej: „d_e” (patrz równanie 4)
    - ulicznej: „d_1” (patrz równanie 5)
    - Czebyszewa: „d_∞” (patrz równanie 6)
- W przypadku zastosowania metod opartych na szczególnych miarach podobieństwa próbek
  - miary Jaccarda (patrz równanie 1)
  - metody n-gramów (patrz równanie 3)
  - miary podobieństwa opartej o słowa kluczowe (patrz sekcja 2.5)

Zdefiniowany zbiór słów kluczowych składa się ze skrótów giełdowych, nazw organizacji, nazwisk, nazw geograficznych oraz nazw tematów dołączonych do zbioru danych *Reuters-21578 Text Categorization Collection Data Set*. Po wyróżnieniu pojedynczych słów oznacza to zastosowanie zbioru **853** słów kluczowych.

Dodatkowym elementem testów była ocena znaczenia parametru k na wyniki klasyfikacji. Większość prób wykonywana była przy ustawieniu $k = 5$, lecz dla typowych przypadków wykonano dodatkowe próby przy $k = 3$ oraz $k = 11$, aby zaobserwować jaki wpływ będzie to miało na wyniki.

5. Wyniki

[1] przekształcenie na wektor cech w R^n
[2] patrz sekcja 2.2
[3] patrz równanie 4
[4] patrz równanie 5
[5] patrz równanie 6
[6] patrz sekcja 2.3
[7] patrz sekcja 2.4
[8] patrz równanie 1
[9] patrz sekcja 2.5
[10] patrz równanie 3

Tabela 1: Wyniki przeprowadzonych prób dla kategoryzacji względem krajów

L.p.	Zest.	Zb. tr. i ucz.	Podobieństwo	k	Kategoria	TPR	PPV
1	Kraje	4% i 6%	R^{n1}, wszystko[2], ${d_e}^3$	3	canada	15,00%	8,57%
					france	0,00%	0,00%
					japan	30,00%	60,00%
					uk	20,00%	33,33%
					usa	92,20%	93,74%
					west-germany	0,00%	0,00%
					poprawne:	**84**,**94**%	
2	Kraje	4% i 6%	R^{n1}, wszystko[2], ${d_e}^3$	5	canada	0,00%	0,00%
					france	0,00%	0,00%
					japan	30,00%	50,00%
					uk	20,00%	33,33%
					usa	95,48%	92,26%
					west-germany	0,00%	0,00%
					poprawne:	**87**,**36**%	
3	Kraje	4% i 6%	R^{n1}, wszystko[2], ${d_e}^3$	11	canada	0,00%	0,00%
					france	0,00%	0,00%
					japan	30,00%	60,00%
					uk	10,00%	33,33%
					usa	99,59%	91,68%
					west-germany	0,00%	0,00%
					poprawne:	**90**,**89**%	
4	Kraje	4% i 6%	R^{n1}, wszystko[2], ${d_1}^4$	5	canada	0,00%	0,00%
					france	0,00%	0,00%
					japan	0,00%	0,00%
					uk	40,00%	11,76%
					usa	94,87%	92,77%
					west-germany	0,00%	0,00%
					poprawne:	**86**,**62**%	
5	Kraje	4% i 6%	R^{n1}, wszystko[2], ${d_\infty}^5$	5	canada	5,00%	5,88%
					france	0,00%	0,00%
					japan	40,00%	50,00%
					uk	10,00%	11,11%
					usa	94,25%	91,62%
					west-germany	0,00%	0,00%
					poprawne:	**86**,**43**%	
6	Kraje	10% i 15%	R^{n1}, wszystko[2], ${d_e}^3$	5	canada	3,28%	4,88%
					france	12,50%	9,09%
					japan	11,76%	80,00%
					uk	43,75%	26,92%
					usa	95,69%	91,58%
					west-germany	0,00%	0,00%
					poprawne:	**85**,**65**%	

Tabela 1: Wyniki przeprowadzonych prób dla kategoryzacji względem krajów

L.p.	Zest.	Zb. tr. i ucz.	Podobieństwo	k	Kategoria	TPR	PPV
7	Kraje	20% i 30%	R^{n1}, wszystko[2], ${d_e}^3$	5	canada	5,26%	16,33%
					france	36,67%	40,74%
					japan	39,39%	58,21%
					uk	61,61%	51,49%
					usa	95,94%	89,44%
					west-germany	12,96%	100,00%
					poprawne:	**84,98%**	
8	Kraje	4% i 6%	R^{n1}, wybr. sł.[6], ${d_e}^3$	5	canada	15,00%	100,00%
					france	33,33%	50,00%
					japan	40,00%	66,67%
					uk	20,00%	33,33%
					usa	98,97%	92,51%
					west-germany	0,00%	0,00%
					poprawne:	**91,45%**	
9	Kraje	10% i 15%	R^{n1}, wybr. sł.[6], ${d_e}^3$	5	canada	31,15%	90,48%
					france	25,00%	25,00%
					japan	52,94%	94,74%
					uk	34,38%	68,75%
					usa	98,56%	92,09%
					west-germany	21,43%	37,50%
					poprawne:	**90,78%**	
10	Kraje	20% i 30%	R^{n1}, wybr. sł.[6], ${d_e}^3$	5	canada	32,89%	83,33%
					france	40,00%	50,00%
					japan	52,53%	75,36%
					uk	53,57%	69,77%
					usa	97,59%	90,38%
					west-germany	22,22%	41,38%
					poprawne:	**88,29%**	
11	Kraje	40% i 60%	R^{n1}, wybr. sł.[6], ${d_e}^3$	5	canada	28,39%	72,00%
					france	56,18%	75,76%
					japan	65,09%	82,63%
					uk	65,69%	84,81%
					usa	97,71%	89,71%
					west-germany	41,04%	70,51%
					poprawne:	**88,41%**	
12	Kraje	4% i 6%	R^{n1}, zb. rozm.[7], ${d_e}^3$	5	canada	25,00%	83,33%
					france	66,67%	66,67%
					japan	70,00%	100,00%
					uk	30,00%	60,00%
					usa	99,38%	93,80%
					west-germany	0,00%	0,00%
					poprawne:	**93,12%**	

Tabela 1: Wyniki przeprowadzonych prób dla kategoryzacji względem krajów

L.p.	Zest.	Zb. tr. i ucz.	Podobieństwo	k	Kategoria	TPR	PPV
13	Kraje	10% i 15%	R^{n1}, zb. rozm.[7], ${d_e}^3$	5	canada	29,51%	90,00%
					france	50,00%	57,14%
					japan	61,76%	84,00%
					uk	40,62%	72,22%
					usa	98,82%	92,33%
					west-germany	25,00%	70,00%
					poprawne:	**91,52%**	
14	Kraje	20% i 30%	R^{n1}, zb. rozm.[7], ${d_e}^3$	5	canada	35,53%	72,97%
					france	50,00%	48,39%
					japan	64,65%	80,00%
					uk	60,71%	78,16%
					usa	97,32%	91,14%
					west-germany	24,07%	56,52%
					poprawne:	**89,10%**	
15	Kraje	40% i 60%	R^{n1}, zb. rozm.[7], ${d_e}^3$	3	canada	31,55%	38,76%
					france	58,43%	57,78%
					japan	72,17%	80,95%
					uk	65,36%	80,32%
					usa	94,49%	90,55%
					west-germany	43,28%	69,05%
					poprawne:	**86,37%**	
16	Kraje	40% i 60%	R^{n1}, zb. rozm.[7], ${d_e}^3$	5	canada	28,08%	72,36%
					france	60,67%	68,35%
					japan	74,53%	82,29%
					uk	68,63%	80,15%
					usa	97,62%	90,59%
					west-germany	38,81%	77,61%
					poprawne:	**88,88%**	
17	Kraje	40% i 60%	R^{n1}, zb. rozm.[7], ${d_\infty}^5$	5	canada	29,34%	72,09%
					france	66,29%	68,60%
					japan	75,94%	83,85%
					uk	66,99%	79,77%
					usa	97,52%	90,82%
					west-germany	41,79%	74,67%
					poprawne:	**89,01%**	
18	Kraje	40% i 60%	R^{n1}, zb. rozm.[7], ${d_1}^4$	5	canada	28,08%	72,36%
					france	53,93%	72,73%
					japan	70,75%	82,87%
					uk	65,69%	83,06%
					usa	97,89%	90,00%
					west-germany	38,81%	77,61%
					poprawne:	**88,67%**	

Tabela 1: Wyniki przeprowadzonych prób dla kategoryzacji względem krajów

L.p.	Zest.	Zb. tr. i ucz.	Podobieństwo	k	Kategoria	TPR	PPV
19	Kraje	20% i 30%, losowanie	R^{n1}, zb. rozm.[7], d_e[3]	5	canada	34,87%	73,61%
					france	50,00%	48,39%
					japan	65,66%	80,25%
					uk	60,71%	78,16%
					usa	97,46%	91,04%
					west-germany	18,52%	55,56%
					poprawne:	**89**,**10**%	
20	Kraje	40% i 60%, losowanie	R^{n1}, zb. rozm.[7], d_e[3]	5	canada	29,57%	72,04%
					france	60,14%	70,34%
					japan	67,97%	83,77%
					uk	70,43%	83,26%
					usa	97,19%	89,39%
					west-germany	34,36%	67,00%
					poprawne:	**87**,**86**%	
21	Kraje	40% i 60%, losowanie, zachodzenie	R^{n1}, zb. rozm.[7], d_e[3]	5	canada	32,07%	69,70%
					france	66,01%	78,91%
					japan	75,09%	83,14%
					uk	74,32%	85,36%
					usa	97,32%	90,75%
					west-germany	52,26%	78,20%
					poprawne:	**89**,**19**%	
22	Kraje	60% i 90%, losowanie, zachodzenie	R^{n1}, zb. rozm.[7], d_e[3]	5	canada	33,99%	78,25%
					france	71,24%	82,56%
					japan	79,19%	88,50%
					uk	74,33%	87,91%
					usa	98,20%	91,22%
					west-germany	45,67%	82,53%
					poprawne:	**90**,**35**%	
23	Kraje	4% i 6%	Jaccard[8]	5	canada	5,00%	8,33%
					france	100,00%	100,00%
					japan	30,00%	50,00%
					uk	40,00%	44,44%
					usa	96,71%	92,72%
					west-germany	0,00%	0,00%
					poprawne:	**89**,**59**%	
24	Kraje	10% i 15%	Jaccard[8]	5	canada	16,39%	43,48%
					france	12,50%	33,33%
					japan	32,35%	91,67%
					uk	46,88%	55,56%
					usa	98,56%	91,23%
					west-germany	3,57%	33,33%
					poprawne:	**89**,**44**%	

Tabela 1: Wyniki przeprowadzonych prób dla kategoryzacji względem krajów

L.p.	Zest.	Zb. tr. i ucz.	Podobieństwo	k	Kategoria	TPR	PPV
25	Kraje	20% i 30%	Jaccard[8]	5	canada	17,11%	50,00%
					france	40,00%	60,00%
					japan	54,55%	71,05%
					uk	67,86%	78,35%
					usa	98,39%	90,74%
					west-germany	20,37%	84,62%
					poprawne:	**88,69%**	
26	Kraje	4% i 6%	Wg. sł. kluczowych[9]	5	canada	5,00%	50,00%
					france	0,00%	0,00%
					japan	10,00%	100,00%
					uk	20,00%	100,00%
					usa	99,79%	91,18%
					west-germany	0,00%	0,00%
					poprawne:	**91,08%**	
27	Kraje	10% i 15%	Wg. sł. kluczowych[9]	5	canada	19,67%	92,31%
					france	0,00%	0,00%
					japan	41,18%	93,33%
					uk	31,25%	55,56%
					usa	99,24%	90,93%
					west-germany	10,71%	42,86%
					poprawne:	**90,11%**	
28	Kraje	4% i 6%	N-gramy[10]	5	canada	0,00%	0,00%
					france	0,00%	0,00%
					japan	0,00%	0,00%
					uk	10,00%	16,67%
					usa	98,36%	91,24%
					west-germany	0,00%	0,00%
					poprawne:	**89,22%**	

Tabela 2: Wyniki przeprowadzonych prób dla kategoryzacji względem tematów

L.p.	Zest.	Zb. tr. i ucz.	Podobieństwo	k	Kategoria	TPR	PPV
1	Tematy	40% i 60%	R^{n1}, wszystko[2], ${d_e}^3$	3	grain	93,78%	99,06%
					interest	98,86%	92,51%
					poprawne:	**96,00%**	
2	Tematy	40% i 60%	R^{n1}, wszystko[2], ${d_e}^3$	5	grain	93,78%	98,60%
					interest	98,29%	92,47%
					poprawne:	**95,75%**	
3	Tematy	40% i 60%	R^{n1}, wszystko[2], ${d_e}^3$	11	grain	94,67%	100,00%
					interest	100,00%	93,58%
					poprawne:	**97,00%**	
4	Tematy	40% i 60%	R^{n1}, wszystko[2], ${d_\infty}^5$	5	grain	84,89%	90,95%
					interest	89,14%	82,11%
					poprawne:	**86,75%**	
5	Tematy	40% i 60%	R^{n1}, wszystko[2], ${d_1}^4$	5	grain	83,56%	98,95%
					interest	98,86%	82,38%
					poprawne:	**90,25%**	
6	Tematy	20% i 30%, losowanie	R^{n1}, wszystko[2], ${d_e}^3$	5	grain	75,24%	100,00%
					interest	100,00%	78,51%
					poprawne:	**87,00%**	
7	Tematy	40% i 60%, losowanie	R^{n1}, wszystko[2], ${d_e}^3$	5	grain	86,03%	100,00%
					interest	100,00%	82,76%
					poprawne:	**91,64%**	
8	Tematy	40% i 60%, losowanie, zachodzenie	R^{n1}, wszystko[2], ${d_e}^3$	5	grain	91,81%	100,00%
					interest	100,00%	90,18%
					poprawne:	**95,33%**	
9	Tematy	60% i 90%, losowanie, zachodzenie	R^{n1}, wszystko[2], ${d_e}^3$	5	grain	92,04%	99,79%
					interest	99,74%	90,33%
					poprawne:	**95,33%**	
10	Tematy	40% i 60%	R^{n1}, wybr. sł.[6], ${d_e}^3$	5	grain	85,78%	98,47%
					interest	98,29%	84,31%
					poprawne:	**91,25%**	
11	Tematy	40% i 60%	R^{n1}, zb. rozm.[7], ${d_e}^3$	5	grain	98,67%	96,52%
					interest	95,43%	98,24%
					poprawne:	**97,25%**	
12	Tematy	40% i 60%	Jaccard[8]	5	grain	98,22%	99,10%
					interest	98,86%	97,74%
					poprawne:	**98,50%**	
13	Tematy	40% i 60%	Wg. sł. kluczowych[9]	5	grain	86,22%	100,00%
					interest	100,00%	84,95%
					poprawne:	**92,25%**	
14	Tematy	40% i 60%	N-gramy[10]	5	grain	93,33%	99,53%
					interest	99,43%	92,06%
					poprawne:	**96,00%**	

Tabela 3: Wyniki przeprowadzonych prób dla kategoryzacji względem autorów

L.p.	Zest.	Zb. tr. i ucz.	Podobieństwo	k	Kategoria	TPR	PPV
1	Autorzy	40% i 60%	R^{n1}, wszystko[2], ${d_e}^3$	3	Byron	76,42%	71,05%
					Shakespeare	72,95%	78,07%
					poprawne:	**74,56%**	
2	Autorzy	40% i 60%	R^{n1}, wszystko[2], ${d_e}^3$	5	Byron	76,42%	70,43%
					Shakespeare	72,13%	77,88%
					poprawne:	**74,12%**	
3	Autorzy	40% i 60%	R^{n1}, wszystko[2], ${d_e}^3$	11	Byron	73,58%	71,56%
					Shakespeare	74,59%	76,47%
					poprawne:	**74,12%**	
4	Autorzy	40% i 60%	R^{n1}, wszystko[2], ${d_\infty}^5$	5	Byron	63,21%	62,04%
					Shakespeare	66,39%	67,50%
					poprawne:	**64,91%**	
5	Autorzy	40% i 60%	R^{n1}, wszystko[2], ${d_1}^4$	5	Byron	93,40%	54,10%
					Shakespeare	31,15%	84,44%
					poprawne:	**60,09%**	
6	Autorzy	40% i 60%	R^{n1}, wybr. sł.[6], ${d_e}^3$	5	Byron	32,08%	58,62%
					Shakespeare	80,33%	57,65%
					poprawne:	**57,89%**	
7	Autorzy	40% i 60%	R^{n1}, zb. rozm.[7], ${d_e}^3$	5	Byron	23,58%	53,19%
					Shakespeare	81,97%	55,25%
					poprawne:	**54,82%**	
8	Autorzy	40% i 60%	Jaccard[8]	5	Byron	53,77%	79,17%
					Shakespeare	87,70%	68,59%
					poprawne:	**71,93%**	
9	Autorzy	40% i 60%	Wg. sł. kluczowych[9]	5	Byron	30,19%	50,79%
					Shakespeare	74,59%	55,15%
					poprawne:	**53,95%**	
10	Autorzy	40% i 60%	N-gramy[10]	5	Byron	1,89%	100,00%
					Shakespeare	100,00%	53,98%
					poprawne:	**54,39%**	

Wykonana seria pomiarów, poza zawartymi w tabeli informacjami o skuteczności klasyfikacji dała nam także pogląd na szybkość poszczególnych metod. Czasy wykonania zależą oczywiście od parametrów komputera, na którym uruchamiany jest program. W przypadku wybranej przez nas konfiguracji referencyjnej (znaczenie mają tu tylko proporcje pomiędzy przedstawionymi czasami) dla zestawu danych **Kraje** czasy klasyfikacji różnymi metodami prezentują się następująco:

	10% danych	**25% danych**	**50% danych**	**pełny zestaw danych**
Zb. rozm.	< 1 min	< 1 min	około 1 min	około 5 min
$\mathbb{R}^n$, wybr. sł.	< 1 min	około 5 min	około 15 min	około 1,5 h
Jaccard	około 1 min	około 5 min	około 40 min	*zbyt długo*
$\mathbb{R}^n$, wszystkie	około 5 min	około 30 min	około 3 h	*zbyt długo*
Wg. sł. kluczowych	około 10 min	około 1,5 h	*zbyt długo*	*zbyt długo*
n-gramy	około 30 min	*zbyt długo*	*zbyt długo*	*zbyt długo*

Tabela 4. Przybliżone czasy dla wybranego zestawu

Zapis *zbyt długo* oznacza, że szacowany czas wykonania testu istotnie przekraczał 3 h i wykonywanie go nie było w naszym odczuciu warunkiem koniecznym aby móc wyciągnąć wnioski z niniejszego zadania laboratoryjnego.

6. Dyskusja

Pierwszym wnioskiem, który możemy wysnuć z powyższych wyników jest fakt, że zwiększając parametr k (w granicach od 3 do 11) jakość klasyfikacji rośnie. Widać także, że wybór metryki ma jedynie niewielki wpływ na jakość klasyfikacji tekstów notatek prasowych. Ma ona natomiast dużo większy wpływ dla klasyfikacji tekstów literackich. Jeśli chodzi o jakość konkretnych metod klasyfikacji, to najlepsza dla zbioru notatek prasowych (klasyfikacja według kraju i według tematu) okazała się nasza autorska metoda bazująca na zbiorach rozmytych. Pozwoliła ona w rozsądnym czasie sklasyfikować wszystkie teksty ze skutecznością sięgającą 89%. Sensownym, aczkolwiek dużo wolniej działającym wyborem jest też metoda oparta na mierze Jaccarda. Dla zbioru tekstów znanych pisarzy bardzo dobre wyniki uzyskała metoda naiwnej ekstrakcji z metryką euklidesową, a także metoda bazująca na mierze Jaccarda. Niektóre metody nie są przydatne, gdyż ich czas obliczeń jest zbyt długi dla dużego zbioru tekstów. Do tych metod należą: metoda n-gramów, naiwna ekstrakcja bazująca na wszystkich słowach a także podobieństwo bazujące na słowach kluczowych.

Uogólniając powyższe widać, że metoda bazująca na mierze Jaccarda, choć nie jest to metoda najszybsza, to sprawdza się w różnych zastosowaniach. Co więcej metoda ta jest prosta w implementacji i nie wymaga żadnej wstępnej preparacji danych, na przykład znajdywania słów kluczowych. Jeżeli zależy nam na szybkości, a określony problem pozwala na zastosowanie metody ekstrakcji opartej o zbiory rozmyte, to warto jej użyć. Jednakże w

problemach, w których trudno jest odnaleźć właściwe słowa kluczowe, jak na przykład w problemie rozpoznawania autorstwa, lepiej użyć metody bardziej ogólnej. Metoda ekstrahująca wszystkie słowa ma dobrą skuteczność, ale jej czas działania niestety ją dyskwalifikuje. Zupełnie natomiast nie radzi sobie miara bazująca na n-gramach, ze względu na powolność działania, co wynika z dużej ilości obliczeń niezbędnych do zastosowania metody.

7. Wnioski

— Nasza autorska metoda bazująca na zbiorach rozmytych jest bardzo skuteczna w problemach w których można wyróżnić słowa kluczowe.
— Dla problemów, w których wyróżnienie słów kluczowych nie jest skuteczne dobrą metodą jest metoda oparta na mierze Jaccarda.
— Metoda n-gramów działa zbyt wolno. Wynika to jednak z faktu, że liczymy n-gramy dla liter. Możliwe, że gdybyśmy użyli modyfikacji tej metody, w której zamiast liter użylibyśmy całych słów, to metoda byłaby szybsza i dawała lepsze wyniki.

Literatura

[1] Niewiadomski A., 2009, *Materiały, przykłady i ćwiczenia do przedmiotu Komputerowe Systemy Rozpoznawania*, skrypt
[2] Shakespeare W., 1564-1616, *The Complete Works of William Shakespeare*, Project Gutenberg, `http://www.gutenberg.org/ebooks/100`
[3] Lord Byron, G. G. 1788-1824, *The Works of Lord Byron, Vol. 4*, Project Gutenberg, `http://www.gutenberg.org/ebooks/20158`