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.

   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 charakteryzuje nam wybrane kraje.

   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.[...]".

[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

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:

$$c_3 = \frac{|s : s \in D \wedge s \in A|}{|A|} \tag{2}$$

gdzie D - zbiór słów oznaczających daty, 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.

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 dłu-

gość może być różna. Liczbę wyrazów znajdujących się w tekście można przedstawić następująco:

$$c_5 = |A| \tag{3}$$

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:

$$f(l) = \begin{cases} 1 & \text{jeśli } l \in <65, 90> \\ 0 & \text{jeśli } l \notin <65, 90> \end{cases} \tag{4}$$

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:

$$c_6 = \frac{|s : s \in Z \wedge s \notin W \wedge s \notin M|}{|A|} \tag{5}$$

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:

$$c_7 = \frac{|s : s \in W|}{|A|} \tag{6}$$

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:

$$c'_8 = \frac{|s : s \in S \wedge s \in A|}{|A|} \tag{7}$$

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:

$$c''_8 = \frac{|s : s \in I \wedge s \in A|}{|A|} \tag{8}$$

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

$$c_8 = c'_8 - c''_8 \tag{9}$$

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:

$$c_9 = \frac{|s : s \in Y|}{|A|} \tag{10}$$

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:

$$c_{10} = \frac{|s : s \in K \wedge s \in A|}{|A|} \tag{11}$$

gdzie K - zbiór słów kluczowych, 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. stopliście, 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 ob-

niżyłoby skuteczność algorytmu. Wyznaczenie opisywanej cechy można przedstawić w postaci operacji na zbiorach:

$$c_{11} = d_t(A - S) \tag{12}$$

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:

$$v = [c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}] \tag{13}$$

gdzie s $\in \langle$ 1, 11 $\rangle$, a c_s oznacza cechę o odpowiednim numerze

2.2. Miary jakości klasyfikacji

Miary jakości klasyfikacji (Accuracy, Precision, Recall, F1). We wprowadzeniu zaprezentować minimum teorii potrzebnej do realizacji zadania, tak by inżynier innej specjalności zrozumiał dalszy opis.

Stosowane wzory, oznaczenia z objaśnieniami znaczenia symboli użytych w doświadczeniu. Oznaczenia jednolite w obrębie całego sprawozdania. Opis zawiera przypisy do bibliografii zgodnie z Polską Normą, (zob. materiały BG PŁ).

Sekcja uzupełniona jako efekt zadania Tydzień 03 wg Harmonogramu Zajęć na WIKAMP KSR.

3. Klasyfikacja z użyciem metryk i miar podobieństwa tekstów

Wzory, znaczenia i opisy symboli zastosowanych metryk z przykładami. Wzory, opisy i znaczenia miar podobieństwa tekstów zastosowanych w obliczaniu metryk dla wektorów cech z przykładami dla każdej miary [4]. Oznaczenia jednolite w obrębie całego sprawozdania. Wstępne wyniki miary Accuracy dla próbnych klasyfikacji na ograniczonym zbiorze tekstów (podać parametry i kryteria wyboru wg punktów 3.-8. z opisu Projektu 1.).

Sekcja uzupełniona jako efekt zadania Tydzień 04 wg Harmonogramu Zajęć na WIKAMP KSR.

4. Budowa aplikacji

4.1. Diagramy UML

Diagramy UML i zwięzłe opisy: idei aplikacji, modułu ekstrakcji i modułu klasyfikatora.
Sekcja uzupełniona jako efekt zadania Tydzień 03 wg Harmonogramu Zajęć na WIKAMP KSR.

4.2. Prezentacja wyników, interfejs użytkownika

Krótki ilustrowany opis jak użytkownik może korzystać z aplikacji, w szczególności wprowadzać parametry klasyfikacji i odczytywać wyniki. Wersja JRE i inne wymogi niezbędne do uruchomienia aplikacji przez użytkownika na własnym komputerze.
Sekcja uzupełniona jako efekt zadania Tydzień 04 wg Harmonogramu Zajęć na WIKAMP KSR.

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] 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.

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