Komputerowe systemy rozpoznawania
2024/2025
Projekt 1. Klasyfikacja dokumentów tekstowych

Dominik Gałkowski, 247659
Jan Śladowski, 247806
Prowadzący: dr inż. Marcin Kacprowicz

24 marca 2025

1 Cel projektu

Celem projektu jest przygotowanie aplikacji, która będzie dokonywała klasyfikacji zbioru dokumentów tekstowych metodą k-NN. Jej zadaniem będzie przydzielenie obiektu do odpowiedniej klasy. W trakcie działania programu konieczne będzie dokonanie ekstrakcji wektorów cech z artykułów dostępnych pod linkiem: `https://archive.ics.uci.edu/dataset/137/reuters+21578+text+categorization+collection`.

2 Klasyfikacja nadzorowana metodą k-NN. Ekstrakcja cech, wektory cech

Metoda k-NN (k-Nearest Neighbors) jest algorytmem leniwym, co oznacza, że nie tworzy wewnętrznej reprezentacji danych uczących, tylko przechowuje wszystkie wzorce uczące. Dopiero po pojawieniu się wzorca testowego, dla którego wyznaczana jest odległość względem wszystkich wzorców uczących, algorytm poszukuje rozwiązania. [1]. Algorytm k-NN wymaga dwóch kluczowych parametrów, metryki, za pomocą, której wyznacza odległości obiektu testującego od wszystkich wzorców uczących oraz liczby sąsiadów k, czyli elementów do których badany element ma najbliżej. Decyzja klasyfikacyjna opiera się na najczęstszej klasie wśród k najbliższych sąsiadów. W przypadku naszego projektu odległość pomiędzy obiektami oznacza skalę podobieństwa tekstów.

W projekcie ekstrakcja cech charakterystycznych tekstu jest dokonywana poprzez stworzenie wektora cech, opisanego na podstawie następujących cech:

1. Długość tekstu - cecha ta oznacza liczbę słów, z których składa się dany

artykuł, co pozwala na porównanie długości różnych tekstów.

$$len = \sum_{i=0}^{n} x_i \tag{1}$$

gdzie x = liczba liter ≥ 3, n = liczba słów w tekście.

2. Dominująca waluta - cecha ta reprezentowana jest poprzez nazwę waluty, ze zbioru walut kluczowych, która pojawia się najczęściej w badanym artykule. Na przykład w przypadku, gdy w badanym tekście pojawi się dwukrotnie słowo "U.S. Dollar"i tylko raz "Japanese Yen"to dla tej cechy zostanie zwrócona wartość tekstowa "U.S. Dollar".

$$w = \arg\max_{w \in W} f(w) \tag{2}$$

gdzie W - zbiór walut kluczowych, $f(w)$ - liczba wystąpień waluty w w tekście.

3. Nazwy miejsca - cecha ta jest reprezentacją tekstową wszystkich miejsc, np. nazw miast lub regionów pojawiających się ze zbioru miejsc kluczowych. Przykład: "AMR Corp will hold a press conference this morning in New York at 0900 EST, a company spokesman said."wynikiem dla tego cytatu będzie zbiór $M' = \{\text{New York}\}$.

$$M' = x \in M \wedge x \in T \tag{3}$$

gdzie M - zbiór miejsc kluczowych, T - zbiór słów znajdujących się w tekście, x = liczba liter ≥ 3.

4. Liczba unikalnych słów - cecha oznaczająca wystąpienia słów unikalnych, czyli takich, które nie pojawiają się więcej niż jeden raz w tekście. Przykład: "AMR Corp will hold a press conference this morning in New York at 0900 EST, a company spokesman said. And the next week also in New York", słowa "New York"nie zostaną zliczone.

$$uk = \mid x : x \in T \wedge f(x) = 1 \mid \tag{4}$$

gdzie T - zbiór słów znajdujących się w tekście, $f(x)$ - funkcja zwracająca liczbę wystąpień słowa x w tekście, x = liczba liter ≥ 3.

5. Średnia długość słowa - cecha opisująca średnią długość słów w badanym tekście.

$$al = \frac{\sum_{i=0}^{m} a_i}{\sum_{i=0}^{n} x_i} \tag{5}$$

gdzie a_i - litera, x - liczba liter ≥ 3, n = liczba słów w tekście, m = liczba liter w tekście.

6. Liczba słów kluczowych w pierwszych 3 zdaniach - cecha ta oznacza bezwzględną liczbę wystąpień słów ze zbioru słów kluczowych w pewnym fragmencie tekstu (pierwsze 3 zdania).

$$fw = \mid x : x \in K \wedge x \in T_{\mathrm{y}} \mid \tag{6}$$

gdzie K - zbiór słów kluczowych, T_y - zbiór słów znajdujący się w pierwszych trzech zdaniach tekstu, x = liczba liter ≥ 3.

7. Liczba słów zaczynających się wielką literą - cecha ta oznaczą liczbę wystąpień słów zaczynających się wielką literą, nie uzwlędniając przy tym słów rozpoczynających nowe zdanie.

$$bw = \sum_{i=0}^{n} x_i \tag{7}$$

gdzie x = słowo zaczynające się wielką literą, n = liczba słów zaczynających się wielka literą w tekście

8. Pierwsze kluczowe słowo w tekście - cecha opisująca pierwsze znalezione słowo znajdujące się w zbiorze słów kluczowych. Przykład: "AMR Corp will hold a press conference this morning in New York at 0900 EST, a company spokesman said."wynikiem dla tego cytatu będzie x_{first} = New York.

$$x_{first} = \min\{x : x \in K \wedge x \in T\} \tag{8}$$

gdzie K - zbiór słów kluczowych, T - zbiór słów znajdujących się w tekście, x = liczba liter ≥ 3.

9. Liczba słów kluczowych - cecha ta oznacza bezwzględną liczbę wystąpień słów ze zbioru słów kluczowych.

$$kw = \mid x : x \in K \wedge x \in T \mid \tag{9}$$

gdzie K - zbiór słów kluczowych, T - zbiór słów znajdujących się w tekście, x = liczba liter ≥ 3.

10. Względna liczba słów kluczowych - cecha która reprezentuje stosunek słów kluczowych do długości całego tekstu.

$$rw = \frac{\mid x : x \in K \wedge x \in T \mid}{\sum_{i=0}^{n} x_i} \tag{10}$$

gdzie K - zbiór słów kluczowych, x = liczba liter ≥ 3, T - zbiór słów znajdujących się w tekście, n = liczba słów w tekście

11. Nazwiska - cecha ta jest reprezentacją tekstową wszystkich nazwisk pojawiających się ze zbioru nazwisk kluczowych. Przykład: "Wallis was quoted as saying the Reagan Administration wants Japanese cooperation so the

White House can ensure any U.S."wynikiem dla tego cytatu będzie zbiór $N' = \{\text{Reagan}\}$.

$$N' = x \in N \wedge x \in T \tag{11}$$

gdzie N - zbiór nazwisk kluczowych, T - zbiór słów znajdujących się w tekście, $x =$ liczba liter ≥ 3.

Wektor cech będzie miał postać:

$$v = [c1, c2, c3, c4, c5, c6, c7, c8, c9, c10, c11] \tag{12}$$

3 Miary jakości klasyfikacji

W celu określenia jakości klasyfikacji wykorzystywane są cztery miary jakości. Accuracy (dokładność) określa jaka część obiektów, ze wszystkich zaklasyfikowanych, została zaklasyfikowana poprawnie. Precision (precyzja), dzięki niej dowiadujemy się, ile wśród obiektów sklasyfikowanych do danej klasy jest rzeczywiście tej klasy. Recall (czułość) opisuje jaki jest udział poprawnie sklasyfikowanych obiektów wśród wszystkich obiektów tej klasy. F1 to średnia harmoniczna pomiędzy precyzją a czułością. Każda z opisanych miar jakości przyjmuje wartości z zakresu $< 0, 1 >$, przy czym wyższa wartość każdej miary oznacza, że więcej tekstów zostało poprawnie sklasyfikowanych.
Dokładność jest obliczana dla wszystkich klas jednocześnie:

$$ACC = \frac{TC + TJ + TU + TF + TW + TUK}{T + N} \tag{13}$$

Precison, Recall oraz F1 jest obliczane dla każdej klasy oddzielnie:

$$PPV_X = \frac{TX}{TX + NX} \tag{14}$$

$$TPR_X = \frac{TX}{X} \tag{15}$$

$$F1_X = \frac{2 * PPV_X * TPR_X}{PPV_X + TPR_X} \tag{16}$$

gdzie:

ACC - accuracy, PPV - precision, TPR - recall, $X \in \{C, J, U, F, W, UK\}$,
T - liczba poprawnie zaakwalifikowanych tekstów,
N - liczba niepoprawnie zaakwalifikowanych tekstów,
C - teksty o Kanadzie,
J - teksty o Japonii,
U - teksty o USA,
F - teksty o Francji,
W - teksty o Niemczech Zachodnich,
UK - teksty o Wielkiej Brytanii.

4 Metryki i miary podobieństwa tekstów w klasyfikacji

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 [2]. Oznaczenia jednolite w obrębie całego sprawozdania. **Podaj metryki i miary podobieństwa nie z literatury (te wystarczy zacytować linkiem), ale konkretne ich postaci stosowane w zadaniu. Jakie zakresy wartości przyjmują te miary i metryki, co oznaczają ich wartości? Podaj przykładowe wartości dla przykładowych wektorów cech**.
Sekcja uzupełniona jako efekt zadania Tydzień 04 wg Harmonogramu Zajęć na WIKAMP KSR.

5 Wyniki klasyfikacji dla różnych parametrów wejściowych

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 05 wg Harmonogramu Zajęć na WIKAMP KSR.**

6 Dyskusja, wnioski, sprawozdanie końcowe

Wyniki kolejnych eksperymentów wg punktów 2.-8. opisu projektu 1. Każdorazowo podane parametry, dla których przeprowadzana eksperyment. Wykresy (np. słupowe) i tabele wyników 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.Dokładne interpretacje uzyskanych wyników w zależności od parametrów klasyfikacji opisanych w punktach 3.-8 opisu Projektu 1. 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 zadań Tydzień 05 i 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] https://home.agh.edu.pl/ horzyk/lectures/miw/KNN.pdf

[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.