

Data oddania: _____

Ocena: _____

Mateusz Walczak 216911

Konrad Kajszczyk 216790

Zadanie 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja*

1. Cel

Celem zadania było stworzenie aplikacji służącej do klasyfikacji artykułów prasowych metodą k-NN. Korzystając z różnych metod wyboru słów kluczowych i ekstrakcji wektorów cech oraz istniejących miar podobieństwa, należało porównać przypisane przez naszą aplikację kategorie artykułów do tych faktycznych. Należało również podjąć próbę opracowania własnej miary podobieństwa i/lub metryki.

2. Wprowadzenie

Algorytm k najbliższych sąsiadów jest bardzo prostym klasyfikatorem probabilistycznym. Niekiedy mówi się, że algorytm k-NN jest leniwy. Wynika to z faktu, że nie tworzy on wewnętrznej reprezentacji danych treningowych (uczących), ale rozpoczyna poszukiwanie rozwiązania dopiero podczas analizy konkretnego wzorca ze zbioru testowego.

Algorytm przechowuje zbiór wszystkich wzorców uczących, względem których obliczana jest odległość wzorca testowego, zdefiniowana poprzez odpowiednią metrykę. Następnie algorytm wybiera k wzorców treningowych, nazywanych sąsiadami, do których aktualnie badany wzorec testowy ma

* SVN: <https://github.com/Walducha1908/KSR1>

najmniejszą odległość. Ostateczny rezultat - kategoria, do której zostanie przypisany analizowany wzorzec - stanowi najczęściej występująca kategoria wśród k najbliższych sąsiadów.

2.1. Metryki

Do obliczenia odległości pomiędzy tekstami posłużyliśmy się następującymi metrykami:

- Metryka Euklidesowa - w celu obliczenia odległości $d_e(x, y)$ między dwoma punktami x, y należy obliczyć pierwiastek kwadratowy z sumy kwadratów różnic wartości współrzędnych o tych samych indeksach, zgodnie ze wzorem:

$$d_e(x, y) = \sqrt{(y_1 - x_1)^2 + \dots + (y_n - x_n)^2} \quad (1)$$

- Metryka uliczna (Manhattan, miejska) - w celu obliczenia odległości $d_m(x, y)$ między dwoma punktami x, y należy obliczyć sumę wartości bezwzględnych różnic współrzędnych punktów x oraz y , zgodnie ze wzorem:

$$d_m(x, y) = \sum_{k=1}^n |x_k - y_k| \quad (2)$$

- Metryka Czebyszewa - w celu obliczenia odległości $d_{ch}(x, y)$ między dwoma punktami x, y należy obliczyć maksymalną wartość bezwzględnych różnic współrzędnych punktów x oraz y , zgodnie ze wzorem:

$$d_{ch}(x, y) = \max_i |x_i - y_i| \quad (3)$$

- Metryka Hamminga - definiujemy jako ilość różnic pomiędzy dwoma wektorami o tej samej długości. Aby obliczyć odległość $d_h(x, y)$ między dwoma punktami x, y należy posłużyć się wzorem [1] :

$$d_h(x, y) = \sum_{i=1}^n |h(i)|, \quad (4)$$

gdzie

$$h(i) = \begin{cases} 0 & \text{jeśli } v_{1i} = v_{2i} \\ 1 & \text{w przeciwnym wypadku} \end{cases} \quad (5)$$

- Odległość Canberra - ważona wersja metryki ulicznej, aby obliczyć odległość $d_c(x, y)$ między dwoma punktami x, y należy posłużyć się wzorem:

$$d_c(x, y) = \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (6)$$

2.2. Wyznaczanie słów kluczowych

Aby wyznaczyć słowa kluczowe posługujemy się poniższą metodą:

- Term frequency - metoda polegająca na zliczeniu liczby wystąpień danego słowa we wszystkich dokumentach.

Przeprowadzamy obliczenia na zbiorze wszystkich posiadanych danych (w naszym przypadku na wszystkich artykułach) i otrzymujemy zestaw par - słowo i wartość. Taki zestaw par sortujemy malejąco po wartości i wybieramy n pierwszych słów. Wybrane n słów staje się słowami kluczowymi.

Taki schemat powtarzamy l razy, gdzie l jest liczbą kategorii na jakie klasyfikujemy. Ostatecznie otrzymujemy l zestawów słów kluczowych, przy czym każdy zestaw reprezentuje inną kategorię. Otrzymane zbiory słów kluczowych oznaczamy:

$$K_1, K_2, \dots, K_{l-1}, K_l. \quad (7)$$

Otrzymany zbiór słów kluczowych będziemy używać we wszystkich iteracjach programu. Słowa kluczowe będą niezmiennie, a wszystkie przeprowadzone przez nas eksperymenty będą bazowały na tym samym zbiorze słów kluczowych.

2.3. Wyznaczanie ważonych słów kluczowych

W celach poprawienia jakości klasyfikacji wprowadzono "ważone słowa kluczowe". Tak nazwaliśmy zestaw par - słowo kluczowe i waga (wartość zmiennoprzecinkowa), z wykorzystaniem których przeprowadziliśmy takie same eksperymenty jak z wykorzystaniem "zwykłych" słów kluczowych, opisanych w poprzednim podpunkcie.

Ważone słowa kluczowe to nic innego jak obliczony wcześniej, ten sam zestaw słów, jednak ubogacony o wagę, obliczaną zgodnie z opracowanym przez nas wzorem:

$$W_i = \left(1 - \frac{N_{W_i \in K_l}}{l - 1}\right)^2, \quad (8)$$

gdzie W_i - waga i -tego słowa kluczowego, l - liczba kategorii, $N_{W_i \in K_l}$ - liczba kategorii słów kluczowych (innych od swojej własnej), w których i -te słowo kluczowe występuje.

Dla jasności przeanalizujemy przykład. Niech $l = 3$, a obliczone słowa kluczowe mają postać:

$$K_1 = \{"jesien", "ogon", "krowa"\}, \quad (9)$$

$$K_2 = \{"wiosna", "ogon", "pies"\}, \quad (10)$$

$$K_3 = \{"lato", "ogon", "krowa"\}, \quad (11)$$

Obliczmy wartości wag dla wybranych słów kluczowych z powyższego zestawu. Dla słowa "jesien" otrzymamy następującą wartość:

$$W_{jesien} = \left(1 - \frac{0}{2}\right)^2 = 1, \quad (12)$$

słowo "jesien" wystąpiło tylko w jednej, "swojej" kategorii, ma zatem największą możliwą wagę.

Dla słowa "krowa":

$$W_{krowa} = \left(1 - \frac{1}{2}\right)^2 = 0.25, \quad (13)$$

słowo "krowa" wystąpiło w jednej dodatkowej kategorii (łącznie w dwóch).

Dla słowa "ogon":

$$W_{ogon} = \left(1 - \frac{2}{2}\right)^2 = 0, \quad (14)$$

słowo "ogon" wystąpiło we wszystkich kategoriach, dlatego też uznajemy, że nie ma dla nas żadnego znaczenia, jego waga jest równa 0.

Z powyższych rozważań bardzo jasno wynika, że wagi słów kluczowych mogą osiągać wartości z przedziału $\langle 0; 1 \rangle$.

2.4. Cechy poddawane ekstrakcji

Ekstrakcja cech charakterystycznych tekstu - w tym celu tworzymy wektor cech, który opisuje tekst (w naszym przypadku artykuł) na podstawie konkretnych, zdefiniowanych cech. Poniżej znajduje się opis wszystkich cech użytych w doświadczeniu.

Przyjęto następujące oznaczenia:

T_i - zbiór słów do badania,

K - stały zbiór słów kluczowych¹,

$N_{K \in T}$ - liczba wystąpień elementów zbioru K w zbiorze T ,²

$C_i(T, K)$ - wartość funkcji cechy.

2.4.1. Liczba wystąpień wszystkich słów kluczowych w całym artykule

Cecha opisująca liczbę słów kluczowych, które występują w całej sekcji głównej artykułu (body).

$$C_1(T_1, K) = N_{K \in T_1}, \quad (15)$$

gdzie T_1 - zbiór słów sekcji głównej artykułu.

2.4.2. Liczba wystąpień wszystkich słów kluczowych w tytule artykułu

Cecha opisująca liczbę słów kluczowych, które występują w tytule artykułu (title).

$$C_2(T_2, K) = N_{K \in T_2}, \quad (16)$$

gdzie T_2 - zbiór słów tytułu artykułu.

¹ Na który składają się zbiory $K_1, K_2, \dots, K_{l-1}, K_l$.

² W przypadku ważonych słów kluczowych będzie to suma iloczynów liczby wystąpień poszczególnych elementów zbioru K w zbiorze T i odpowiadających im wag.

2.4.3. Liczba wystąpień wszystkich słów kluczowych w sekcji daty artykułu

Cecha opisująca liczbę słów kluczowych, które występują w sekcji daty artykułu (dateline).

$$C_3(T_3, K) = N_{K \in T_3}, \quad (17)$$

gdzie T_3 - zbiór słów sekcji daty artykułu.

2.4.4. Stosunek liczby wystąpień wszystkich słów kluczowych do ogólnej liczby słów w artykule

Cecha opisująca stosunek liczby słów kluczowych, które występują w całej sekcji głównej artykułu (body), do całkowitej liczby słów występujących w części głównej.

$$C_4(T_4, K) = \frac{N_{K \in T_4}}{|T_4|}, \quad (18)$$

gdzie T_4 - zbiór słów sekcji głównej artykułu, $|T_4|$ - liczba elementów (słów) zbioru sekcji głównej artykułu.

2.4.5. Liczba wystąpień wszystkich słów kluczowych w pierwszych 50 słowach artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 50 słowach sekcji głównej artykułu. Jeśli artykuł jest krótszy niż 50 słów to bierzemy pod uwagę wszystkie występujące w nim słowa.

$$C_5(T_5, K) = N_{K \in T_5}, \quad (19)$$

gdzie T_5 - pierwsze 50 słów sekcji głównej artykułu.

2.4.6. Liczba wystąpień wszystkich słów kluczowych w pierwszych 10% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 10% sekcji głównej artykułu.

$$C_6(T_6, K) = N_{K \in T_6}, \quad (20)$$

gdzie T_6 - pierwsze 10% słów sekcji głównej artykułu.

2.4.7. Liczba wystąpień wszystkich słów kluczowych w pierwszych 20% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 20% sekcji głównej artykułu.

$$C_7(T_7, K) = N_{K \in T_7}, \quad (21)$$

gdzie T_7 - pierwsze 20% słów sekcji głównej artykułu.

2.4.8. Liczba wystąpień wszystkich słów kluczowych w pierwszych 50% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w pierwszych 50% sekcji głównej artykułu.

$$C_8(T_8, K) = N_{K \in T_8}, \quad (22)$$

gdzie T_8 - pierwsze 50% słów sekcji głównej artykułu.

2.4.9. Liczba wystąpień wszystkich słów kluczowych w pierwszym paragrafie

Cecha opisująca liczbę słów kluczowych, które występują w pierwszym paragrafie sekcji głównej artykułu.

$$C_9(T_9, K) = N_{K \in T_9}, \quad (23)$$

gdzie T_9 - pierwszy paragraf sekcji głównej artykułu.

2.4.10. Liczba wystąpień wszystkich słów kluczowych w ostatnich 50 słowach artykułu

Cecha opisująca liczbę słów kluczowych, które występują w ostatnich 50 słowach sekcji głównej artykułu. Jeśli artykuł jest krótszy niż 50 słów to bierzemy pod uwagę wszystkie występujące w nim słowa.

$$C_{10}(T_{10}, K) = N_{K \in T_{10}}, \quad (24)$$

gdzie T_{10} - ostatnie 50 słów sekcji głównej artykułu.

2.4.11. Liczba wystąpień wszystkich słów kluczowych w ostatnich 10% artykułu

Cecha opisująca liczbę słów kluczowych, które występują w ostatnich 10% sekcji głównej artykułu.

$$C_{11}(T_{11}, K) = N_{K \in T_{11}}, \quad (25)$$

gdzie T_{11} - ostatnie 10% słów sekcji głównej artykułu.

2.4.12. Liczba wystąpień wszystkich słów kluczowych w ostatnim paragrafie

Cecha opisująca liczbę słów kluczowych, które występują w ostatnim paragrafie sekcji głównej artykułu.

$$C_{12}(T_{12}, K) = N_{K \in T_{12}}, \quad (26)$$

gdzie T_{12} - ostatni paragraf sekcji głównej artykułu.

3. Opis implementacji

Należy tu zamieścić krótki i zwięzły opis zaprojektowanych klas oraz powiązań między nimi. Powinien się tu również znaleźć diagram UML (diagram klas) prezentujący najistotniejsze elementy stworzonej aplikacji. Należy także podać, w jakim języku programowania została stworzona aplikacja.

4. Materiały i metody

W tym rozdziale omówione zostaną poszczególne eksperymenty jakie wykonano z użyciem naszego programu.

Klasyfikacje artykułów przeprowadzano ze względu na dwa różne rodzaje etykiet. Pierwszym z nich była lokalizacja (place). Kategorie (etykiety) jakie wyróżniliśmy były następujące: west-germany, usa, france, uk, canada, japan. Klasyfikacja przeprowadzana była jedynie z wykorzystaniem artykułów, których pole "places" przyjmowało jedną z powyższych wartości.

Drugim rodzajem etykiet był temat (topic). Kategorie (etykiety) jakie wyróżniliśmy były następujące: earn, trade, money-supply, acq. Podobnie jak w pierwszym przypadku, klasyfikacja przeprowadzana była jedynie z wykorzystaniem artykułów, których pole "topics" przyjmowało jedną z powyższych wartości.

4.1. Wpływ liczby k sąsiadów oraz wyboru metryki na klasyfikację

Klasyfikacja tekstów została wykonana z wykorzystaniem zbioru (zwykłych) słów kluczowych. Eksperymenty wykonano z użyciem wszystkich pięciu metryk. Dla każdego przypadku testowego dokonano klasyfikacji tekstu dla następujących wartości współczynnika k :

$$k \in \{1, 3, 4, 6, 8, 10, 12, 14, 17, 20\}. \quad (27)$$

W każdym przypadku testowym zbiór treningowy stanowił 70% artykułów, zaś zbiór testowy 30% artykułów.

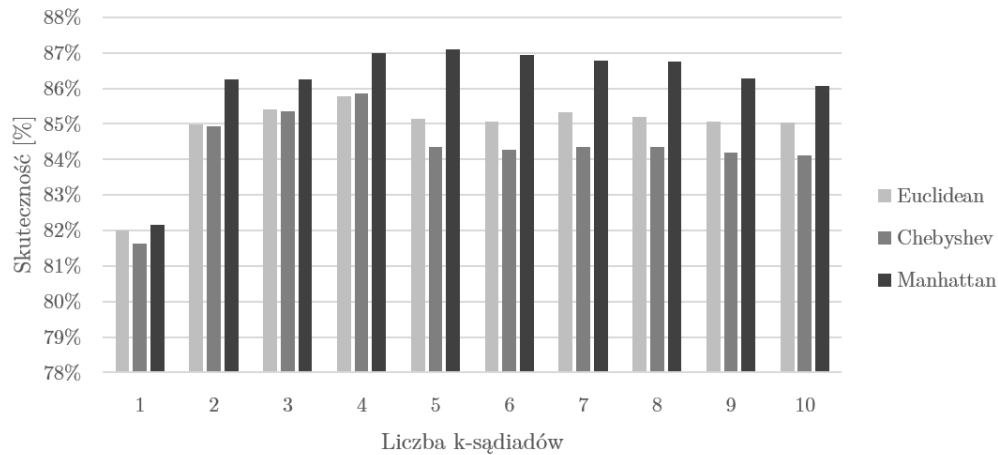
4.2. Wpływ podziału tekstów na zbiory treningowe i testowe na klasyfikację

Klasyfikacja tekstów została wykonana z wykorzystaniem zbioru (zwykłych) słów kluczowych. Eksperymenty przeprowadzono posługując się metryką Euklidesową. Wartość parametru k była stała i wynosiła $k = 6$. Przeprowadzono klasyfikacje dla pięciu różnych podziałów artykułów na zbiory testowe i treningowe:

- Zbiór treningowy: 40% artykułów, zbiór testowy 60%,
- Zbiór treningowy: 50% artykułów, zbiór testowy 50%,
- Zbiór treningowy: 60% artykułów, zbiór testowy 40%,
- Zbiór treningowy: 70% artykułów, zbiór testowy 30%,
- Zbiór treningowy: 80% artykułów, zbiór testowy 20%.

5. Wyniki

W tej sekcji należy zaprezentować, dla każdego przeprowadzonego eksperymentu, kompletny zestaw wyników w postaci tabel, wykresów itp. Powinny być one tak ponazywane, aby było wiadomo, do czego się odnoszą. Wszystkie tabele i wykresy należy oczywiście opisać (opisać co jest na osiach, w kolumnach itd.) stosując się do przyjętych wcześniej oznaczeń. Nie należy tu komentować i interpretować wyników, gdyż miejsce na to jest w kolejnej sekcji. Tu również dobrze jest wprowadzić oznaczenia (tabel, wykresów) aby móc się do nich odwoływać poniżej.



Rysunek 1.

6. Dyskusja

Sekcja ta powinna zawierać dokładną interpretację uzyskanych wyników eksperymentów wraz ze szczegółowymi wnioskami z nich płynącymi. Najcenniejsze są, rzecz jasna, wnioski o charakterze uniwersalnym, które mogą być istotne przy innych, podobnych zadaniach. Należy również omówić i wyjaśnić wszystkie napotkane problemy (jeśli takie były). Każdy wniosek powinien mieć poparcie we wcześniej przeprowadzonych eksperymentach (odwołania do konkretnych wyników). Jest to jedna z najważniejszych sekcji tego sprawozdania, gdyż prezentuje poziom zrozumienia badanego problemu.

7. Wnioski

W tej, przedostatniej, sekcji należy zamieścić podsumowanie najważniejszych wniosków z sekcji poprzedniej. Najlepiej jest je po prostu wypunktować. Znow, tak jak poprzednio, najistotniejsze są wnioski o charakterze uniwersalnym.

Literatura

- [1] A. Niewiadomski *Materiały, przykłady i ćwiczenia do przedmiotu Komputerowe Systemy Rozpoznawania*. 19 czerwca 2012.