Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2019/2020**
Prowadzący: prof. dr hab. inż. Adam Niewiadomski poniedziałek, 12:15

Data oddania: _____ Ocena: _____

Mateusz Walczak 216911
Konrad Kajszczak 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 aplikacje kategorie artykułów do tych faktycznych. Należało również podjąć próbe 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 naiwny lub leniwy. Wynika to z faktu, że nie tworzy on wewnętrznej reprezentacji danych treningowych (uczących), ale ropoczyna 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 wzorzec 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:
$$d_e(x, y) = \sqrt{(y_1 - x_1)^2 + \cdots + (y_n - x_n)^2} \tag{1}$$
- Metryka uliczna (Manhattan, miejska) - w celu obliczenia odległości $d_e(x, y)$ między dwoma punktami x, y należy obliczyć sumę wartości bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:
$$d_m(x, y) = \sum_{k=1}^{n} |x_k - y_k| \tag{2}$$
- Metryka Czebyszewa - w celu obliczenia odległości $d_e(x, y)$ między dwoma punktami x, y należy obliczyć maksymalną wartość bezwzględnych różnic współrzędnych punktów x oraz y, zgodnie ze wzorem:
$$d_{ch}(x, y) = \max_i |x_i - y_i| \tag{3}$$

2.2. Metody wyboru słów kluczowych

W ramach zadania zostały użyte następujące metody wyboru słów kluczowych:

- Term frequency - metoda polegająca na zliczeniu liczby wystąpień danego słowa we wszystkich dokumentach.
- Document frequency - metoda polegająca na zliczeniu liczby dokumentów w których dane słowo występuje przynajmniej raz.

W przypadku obu metod otrzymujemy zestaw par - słowo i wartość obliczoną za pomocą jednej z powyższych metod. Taki zestaw par sortujemy malejąco po wartości i wybieramy n pierwszych słów. Wybrane n słów staje się słowami kluczowymi.

2.3. 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 - zbiór słów do badania,

K - zbiór słów kluczowych,
$N_{K \in T}$ - liczba wystąpień elementów zbioru K w zbiorze T,
$C_i(T, K)$ - wartość funkcji cechy.

2.3.1. Liczba wystąpień wszystkich słów kluczowych w całym artykule

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

$$C_1(T, K) = N_{K \in T}, \tag{4}$$

gdzie T - zbiór słów sekcji głównej artykułu.

2.3.2. Liczba wystąpień wszystkich słów kluczowych w tytule artykułu

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

$$C_2(T, K) = N_{K \in T}, \tag{5}$$

gdzie T - zbiór słów tytułu artykułu.

2.3.3. Liczba wystąpień wszystkich słów kluczowych w sekcji daty artykułu

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

$$C_3(T, K) = N_{K \in T}, \tag{6}$$

gdzie T - zbiór słów sekcji daty artykułu.

2.3.4. 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.

$$C_4(T, K) = \frac{N_{K \in T}}{|T|}, \tag{7}$$

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

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

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

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

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

$$C_7(T, K) = \textit{ liczba wystapien elementow zbioru } K \textit{ w zbiorze } T \tag{8}$$

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

2.3.9. Liczba wystąpień wszystkich słów kluczowych w pierwszym akapicie

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

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

2.3.12. Liczba wystąpień wszystkich słów kluczowych w ostatnim akapicie

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 miejscu należy opisać, jak przeprowadzone zostały wszystkie badania, których wyniki i dyskusja zamieszczane są w dalszych sekcjach. Opis ten powinien być na tyle dokładny, aby osoba czytająca go potrafiła wszystkie przeprowadzone badania samodzielnie powtórzyć w celu zweryfikowania ich poprawności (a zatem m.in. należy zamieścić tu opis architektury sieci, wartości współczynników użytych w kolejnych eksperymentach, sposób inicjalizacji wag, metodę uczenia itp. oraz informacje o danych, na których prowadzone były badania). Przy opisie należy odwoływać się i stosować do opisanych w sekcji drugiej wzorów i oznaczeń, a także w jasny sposób opisać cel konkretnego testu. Najlepiej byłoby wyraźnie wyszczególnić (ponumerować) poszczególne eksperymenty tak, aby łatwo było się do nich odwoływać dalej.

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.

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 napotakane 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. Znów, tak jak poprzednio, najistotniejsze są wnioski o charakterze uniwersalnym.

Literatura

Na końcu należy obowiązkowo podać cytowaną w sprawozdaniu literaturę, z której grupa korzystała w trakcie prac nad zadaniem (przykład na końcu szablonu)