

Data oddania: \_\_\_\_\_

Ocena: \_\_\_\_\_

Radosław Grela 216769  
Jakub Wachała 216914

## Zadanie 1: ekstrakcja cech, miary podobieństwa, klasyfikacja\*

### 1. Cel

Celem naszego zadania było stworzenie aplikacji do klasyfikacji tekstów za pomocą metody  $k$ -NN ( $k$  najbliższych sąsiadów) oraz różnych metryk i miar podobieństwa, a następnie porównać kategorie z tymi wygenerowanymi przez aplikację.

### 2. Wprowadzenie

Głównym zagadnieniem projektowym, z którym mieliśmy do czynienia w ramach zadania 1 była klasyfikacja statystyczna tekstów na podstawie wektora wyekstrahowanych cech. Do przeprowadzenia eksperymentu zaimplementowaliśmy algorytm *k-najbliższych sąsiadów*.

Algorytm  $k$ -najbliższych sąsiadów ( $k$ -NN - *k-nearest neighbors*) to jeden z algorytmów zaliczanych do grupy algorytmów leniwych. Jest to taka grupa algorytmów, która szuka rozwiązania dopiero, gdy pojawia się wzorzec testujący. Przechowuje wzorce uczące, a dopiero później wyznacza się odległość wzorca testowego względem wzorców treningowych. [1].

Algorytm ten działa w taki sposób, że dla każdego wzorca testowego obliczana jest odległość za pomocą wybranej wetryki względem wzorców treningowych, a następnie wybierana jest  $k$  najbliższych wzorców treningowych.

---

\* Github: <https://github.com/Bonniu/KSR>

Wynik wyznaczony jest jako najczęstszy element wśród nich. W naszym zadaniu odległość ta jest równa skali podobieństwa tekstów.

## 2.1. Ekstrakcja cech

Do ekstrakcji cech charakterystycznych tekstu utworzyliśmy wektor cech, który opisuje tekst za pomocą 10 cech. Liczba słów zawsze jest liczona po zastosowaniu stop-listy oraz stemizacji, bez znaków przestankowych.

- $C_1$  - Stosunek słów kluczowych do wszystkich słów w pierwszych 10% tekstu. Obliczona jest za pomocą wzoru:

$$C_1 = s_{k10}/s_{10} \quad (1)$$

gdzie  $s_{k10}$  to liczba słów kluczowych, a  $s_{10}$  to liczba wszystkich słów w pierwszych 10% tekstu.

- $C_2$  - Stosunek słów kluczowych do wszystkich słów w ostatnich 10% tekstu. Obliczona jest za pomocą wzoru:

$$C_2 = s_{k90}/s_{90} \quad (2)$$

gdzie  $s_{k90}$  to liczba słów kluczowych, a  $s_{90}$  to liczba wszystkich słów w ostatnich 10% tekstu.

- $C_3$  - Stosunek słów kluczowych do wszystkich słów w dokumencie. Obliczona jest za pomocą wzoru:

$$C_3 = s_k/s \quad (3)$$

gdzie  $s_k$  to liczba słów kluczowych, a  $s$  to liczba wszystkich słów w dokumencie.

- $C_4$  - Stosunek słów kluczowych, których ilość liter  $\in (0,4]$  do wszystkich słów w dokumencie. Obliczona jest za pomocą wzoru:

$$C_4 = s_k/s \quad (4)$$

gdzie  $s_k$  to liczba słów kluczowych, których ilość liter  $\in (0,4]$ , a  $s$  to liczba wszystkich słów w dokumencie.

- $C_5$  - Stosunek słów kluczowych, których ilość liter jest  $\geq 8$  do wszystkich słów w dokumencie. Obliczona jest za pomocą wzoru:

$$C_5 = s_k/s \quad (5)$$

gdzie  $s_k$  to liczba słów kluczowych, a  $s$  to liczba wszystkich słów w dokumencie.

- Z jakiego zakresu/zbioru cecha przyjmuje wartości przed normalizacją.  
 - Czy „długość” oznacza liczbę liter, a może słów? Czy jest obliczana przed czy po stemizacji i/lub zastosowaniu stoplisty? jeśli znaków, to czy znaki przestankowe także są zliczane? - W przypadku cech mniej intuicyjnych (a najlepiej wszystkich) - mile widziany przykład jak liczyć (może być jeden tekst dla wszystkich cech pod ich opisem). - Jakie znaczenie ma ta cecha tekstu dla jego rozpoznania? Czy np. im tekst dłuższy, tym bardziej związany z etykietą USA lub CANADA? (istotne!)

### 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 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] <http://home.agh.edu.pl/~horzyk/lectures/miw/KNN.pdf> [dostęp 17.03.2020]