

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

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

Przemysław Lis 229940

Paweł Cichocki 150848

## Projekt 1. Klasyfikacja dokumentów tekstowych

Opis projektu ma formę artykułu naukowego lub raportu z zadania badawczego/doświadczonego/obliczeniowego (wg indywidualnych potrzeb związanych np. z pracą inżynierską/naukową/zawodową). Kolejne sekcje muszą być numerowane i zatytułowane. Wzory są numerowane, tablice są numerowane i podpisane nad tablicą, rysunki są numerowane i podpisane pod rysunkiem. Podpis rysunku i tabeli musi być wyczerpujący (nie ogólnikowy), aby czytelnik nie musiał sięgać do tekstu, aby go zrozumieć.

**Wybrane sekcje (rozdziały sprawozdania) są uzupełniane wg wymagań w opisie Projektu 1. i Harmonogramie Zajęć na WIKAMP KSR jako efekty zadań w poszczególnych tygodniach.**

### 1. Cel projektu

Celem projektu jest sparsowanie danych i przeanalizowanie do jakiego kraju odnosi się dany artykuł. Do wykonania zadania wykorzystamy algorytm  $k$ -NN aby należycie dopasować kraj. Oczekujemy że nasz program po analizie danych algorytmem będzie w stanie poprawnie dopasować kraj o którym mowa lub z którego pochodzi artykuł.

### 2. Klasyfikacja nadzorowana metodą $k$ -NN

$k$ -NN jest to algorytm regresji nieparametrycznej. Założeniem algorytmu jest że podobne problemy mają podobne rozwiązania. Algorytm sprawdza n

najbliższych sąsiadów wystąpienia i w zależności od wyniku klasyfikuje jego położenie. Jeżeli w sąsiedztwie naszego artykułu badawczego będzie najwięcej węzłów danego typu, to wtedy zostanie on odpowiednio dopasowany do danego typu. Parametrem wejściowym jest plik tekstowy, który następnie będzie zaklasyfikowany do odpowiedniego kraju.

## 2.1. Ekstrakcja cech, wektory cech

1. Najwięcej wyrazów w pierwszych 5 zdaniach.

$$\sum_{i=0}^5 s_i, \text{ gdzie } s - \text{słowo}$$

2. Stosunek liczby wystąpień słów kluczowych do długości tekstu.

$$\frac{\sum_{i=0}^n s_i}{d}, \text{ gdzie } s - \text{słowo kluczowe, } d - \text{długość tekstu}$$

3. Długość artykułu.

$$\sum_{i=0}^n l_i, \text{ gdzie } l - \text{litera}$$

4. Średnia długość słowa.

$$\frac{\sum_{i=0}^n l_i}{\sum_{j=0}^n s_j}, \text{ gdzie } l - \text{litera, } s - \text{słowo}$$

5. Liczba unikalnych słów.

$$\sum_{i=0}^n s_i, \text{ gdzie } s - \text{słowo unikalne}$$

6. Liczba słów zaczynająca się dużą literą.

$$\sum_{i=0}^n s_i, \text{ gdzie } s - \text{słowo zaczynające się na wielką literę}$$

7. Liczba słów nie przekraczająca 3 znaków.

$$\sum_{i=0}^n \text{len}(s_i) \leq 3, \text{ gdzie } s - \text{słowo}$$

8. Liczba słów.

$$\sum_{i=0}^n s_i, \text{ gdzie } s - \text{słowo}$$

9. Liczba słów dłuższych niż 8 znaków.

$$\sum_{i=0}^n \text{len}(s_i) > 8, \text{ gdzie } s - \text{słowo}$$

10. Litera na którą zaczyna się najwięcej słów z wielkiej litery

## 2.2. Miary jakości klasyfikacji

**Accuracy** - to miara która określa dokładność klasyfikacji. Jest to stosunek ilości poprawnie zaklasyfikowanych artykułów do wszystkich artykułów. Określona wzorem:

$$accuracy = \frac{\text{liczba poprawnie zaklasyfikowanych artykułów}}{\text{liczba wszystkich artykułów}}$$

**Precision** - to miara określająca precyzję klasyfikacji elementów do danej klasy. Liczona jest tak jak accuracy, lecz w obrębie pewnej klasy.

Określa się wzorem:

$$precision = \frac{\text{liczba poprawnie zaklasyfikowanych artykułów do klasy X}}{\text{liczba wszystkich artykułów zaklasyfikowanych do klasy X}}$$

**Recall** - jest to miara oznaczająca odsetek poprawnie zaklasyfikowanych artykułów do danej klasy.

Określa się wzorem:

$$recall = \frac{\text{liczba poprawnie zaklasyfikowanych artykułów do klasy X}}{\text{liczba wszystkich artykułów klasy X}}$$

**F1** - jest to miara uśredniająca wynik. Liczona jest poprzez średnią harmoniczną precision oraz recall.

Określa się wzorem:

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

## 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 [2]. 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.). **Podaj metryki i miary podobieństwa nie z literatury (te wystarczy zacytować linkiem), ale konkretne ich postaci stosowane w zadaniu. Jak zakresy wartości przyjmują te miary i metryki, co oznaczają ich wartości? Podaj przykładowe wartości dla przykładowych wektorów cech.**

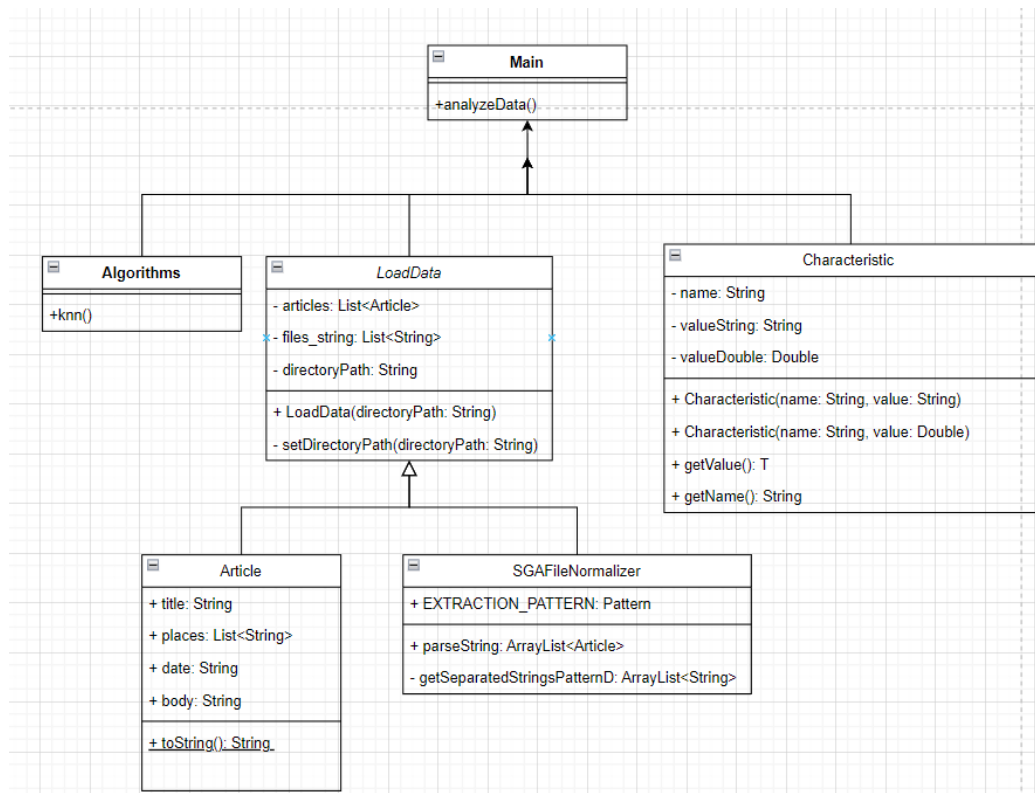
Sekcja uzupełniona jako efekt zadania Tydzień 04 wg Harmonogramu Zajęć na WIKAMP KSR.

## 4. Budowa aplikacji

### 4.1. Diagramy UML

Program będzie posiadał klasę Main która będzie zarządzać całym programem. Będą do niej dołączone klasy takie jak Algorithms, LoadData czy

Characteristic. Do załadowania plików posłużą nam klasa LoadData, która będzie łaadować artykuły do klasy Article za pomocą paternu z klasy SGAFFileNormalizer. klasyfikatora.



Rysunek 1. Diagram UML całego projektu

## 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] 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.