

Data oddania: _____

Ocena: _____

Kamil Kiszko-Zgierski 224328
Hubert Gawłowski 224298

Projekt 1. Klasyfikacja dokumentów tekstowych

1. Cel projektu

Celem projektu jest sklasyfikowanie zbioru dokumentów tekstowych. Klasyfikacja będzie prowadzona metodą k -NN. Wykonana zostanie ekstrakcja cech oraz zbadany zostanie ich wpływ na dokładność klasyfikacji.

2. Klasyfikacja nadzorowana metodą k -NN

Krótki opis metody k -NN: zasada działania, wymagane parametry wejściowe, format i znaczenie wyników/rezultatów. Opis własny z przypisami do literatury – minimum teorii potrzebnej do zadania, tak by inżynier innej specjalności zrozumiał dalszy opis [1].

Sekcja uzupełniona jako efekt zadania Tydzień 02 wg Harmonogramu Zajęć na WIKAMP KSR.

2.1. Ekstrakcja cech, wektory cech

Zostały wyekstrahowane następujące cechy, które posłużą w celu klasyfikacji tekstów:

1. Zapis cyfr - za pomocą ciągu cyfr otrzymamy informacje o np. numerach telefonu, które to są charakterystyczne w zależności od kraju.
2. Waluty - wyciągnięcie z tekstów nazw najczęściej używanych walut. Kraje, z których pochodzą porównywane teksty używają różnych walut.

3. Częstość występowania dat - zliczenie, jak często w podanych tekstach występują daty. Zakładamy, że w zależności, którego kraju dotyczy tekst, częstość wystąpień zapisów datowych będzie się różnić.
4. Format zapisu dat - w zależności od kraju format zapisu dat różni się.
5. Ogólna liczba słów - zliczenie wszystkich słów występujących w tekście. Uważamy, że w zależności od tego, jakiego kraju tekst dotyczy ich długość może być różna
6. Częstość słów rozpoczynających się wielką literą - słowa takie będą oznaczały najczęściej nazwy własne np. imiona, nazwiska, nazwy budynków. Pisząc o jednym kraju może być używane więcej takich słów, a o innych mniej. Z tej grupy wykluczymy jednak wyrazy składające się wyłącznie z wielkich liter, o których mowa będzie w punkcie następnym.
7. Częstość słów pisanych wielką literą - będą to najczęściej skróty. Uważamy, że w zależności o jakim kraju będzie tekst ilość skrótów może się różnić.
8. Układ SI/imperialny - zdecydowanie w krajach anglojęzycznych częściej w tekstach stosowany będzie układ imperialny, natomiast w pozostałych - układ SI.
9. Częstość występowania cytatów - uważamy, że pisząc o niektórych krajach autorzy częściej będą przytaczali różnego rodzaju wypowiedzi.
10. Słowa kluczowe - sporządzona zostanie lista słów kluczowych dla danego kraju (np. elementy lokalizacyjne)
11. Najczęściej występujące słowa - znalezienie słów, które najczęściej występują w tekstach o danym kraju

Wektor cech będzie się prezentował następująco: $v = [c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}]$.

2.2. Miary jakości klasyfikacji

Miary jakości klasyfikacji (Accuracy, Precision, Recall, F1). We wprowadzeniu zaprezentować minimum teorii potrzebnej do realizacji zadania, tak by inżynier innej specjalności zrozumiał dalszy opis.

Stosowane wzory, oznaczenia z objaśnieniami znaczenia symboli użytych w doświadczeniu. Oznaczenia jednolite w obrębie całego sprawozdania. Opis zawiera przypisy do bibliografii zgodnie z Polską Normą, (zob. materiały BG PŁ).

Sekcja uzupełniona jako efekt zadania Tydzień 03 wg Harmonogramu Zajęć na WIKAMP KSR.

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.).

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

4. Budowa aplikacji

4.1. Diagramy UML

Diagramy UML i zwięzłe opisy: idei aplikacji, modułu ekstrakcji i modułu klasyfikatora.

Sekcja uzupełniona jako efekt zadania Tydzień 03 wg Harmonogramu Zajęć na WIKAMP KSR.

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.