Informatyka, studia dzienne, I st. semestr VI

Komputerowe systemy rozpoznawania **2018/2019**
Prowadzący: dr inż. Marcin Kacprowicz wtorek, 16.15

Data oddania: _____ Ocena: _____

Piotr Traczyk 123123
Bartosz Jurczewski 210209

Zadanie 1: Ekstrakcja cech, miary podobieństwa, klasyfikacja*

1. Cel

Celem zadania było stworzenie aplikacji do klasyfikacji tekstów metodą k-NN, korzystając z różnych sposób ekstrakcji wektorów cech oraz istniejących miar podobieństwa porównać kategorie do tych przypisanych przez aplikację i zbadać jakie parametry najbardziej wpływają na rozpoznwanie tekstu.

2. Wprowadzenie

Zagadnieniem, jakim zajmowaliśmy się w ramach projektu jest klasyfikacja statystyczna, która jest rodzajem algorytmu statystycznego przydzielającego elementy do klas, bazując na cechach tych elementów. W ramach przeprowadzanego eksperymentu zaimplementowaliśmy klasyfikator k-najbliższych sąsiadów.

Algorytm k najbliższych sąsiadów, nazywany także algorytmem k-NN, należy do grupy algorytmów leniwych, czyli takich, które nie tworzą wewnętrznej reprezentacji danych uczących, lecz szukają rozwiązania dopiero w momencie pojawienia się wzorca testującego. Przechowuje wszystkie wzorce uczące, względem których wyznacza odległość wzorca testowego [2]. Metoda

* Repozytorium github: `https://github.com/jurczewski/KSR`

k-NN wyznacza k sąsiadów, do których badany element ma najmniejszą odległość w danej metryce, a następnie wyznacza wynik w oparciu o najczęstszy element, wśród k najbliższych. W przypadku naszego projektu odległość definiujemy jako skalę podobieństwa tekstów.

W ramach zadania zostały użyte 2 metody ekstrakcji cech:

— Inverse document frequency - metoda polegająca na wyznaczeniu, czy dane słowo występuje powszechnie we wszystkich dokumentach. Jest to logarytmicznie skalowana odwrotna część dokumentów zawierających wybrane słowo (uzyskana poprzez podzielenie całkowitej liczby dokumentów przez liczbę dokumentów zawierających ten termin). Obliczana jest z poniższego wzoru:

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

— Term frequency - metoda polegająca na zliczeniu częstości występowania danego słowa w dokumencie. Obliczana jest z poniższego wzoru:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k nk, j}$$

Do obliczenia odległości tekstów posłużyliśmy się 3 metrykami:

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

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

— 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:

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

— 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:

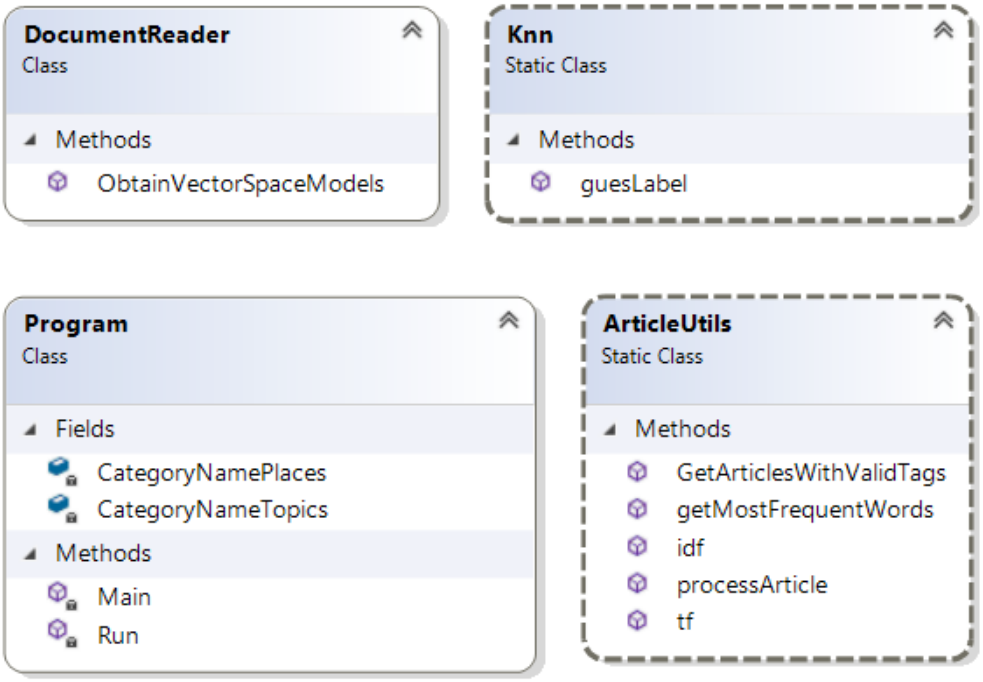
$$d_{ch}(x, y) = \max_i |x_i - y_i|$$

3. Opis implementacji

Program został w całosci stworzony w języku C# (.NET Framework, v4.6.1).

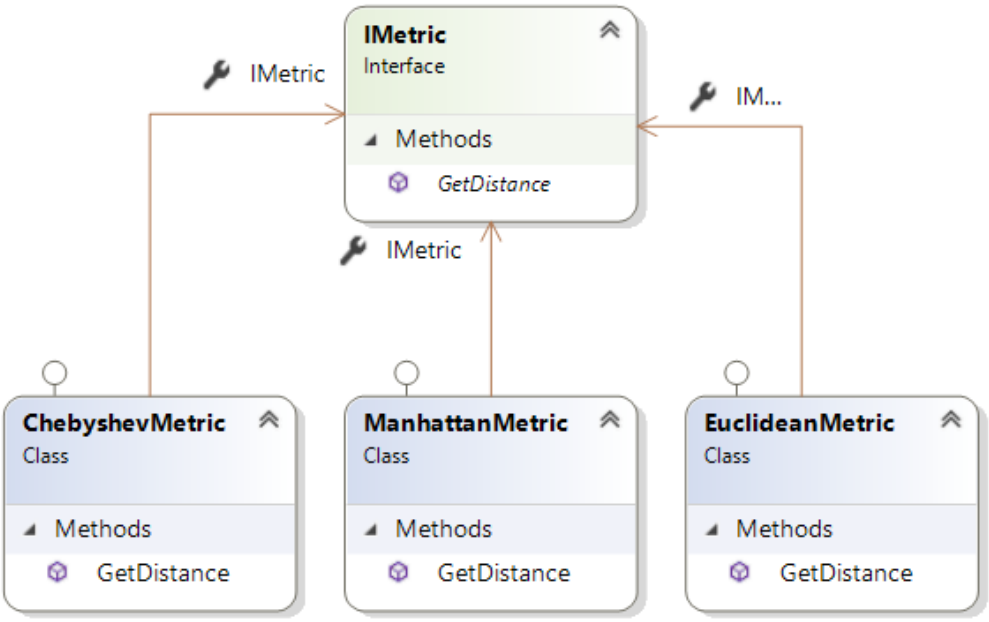
W programie znajdują się następujące klasy:

- DocumentReader - klasa odpowiedzialna
- Knn - implementacja algorytmu K-nn
- Program - klasa główna
- ArticleUtilis - klasa pomocnicza



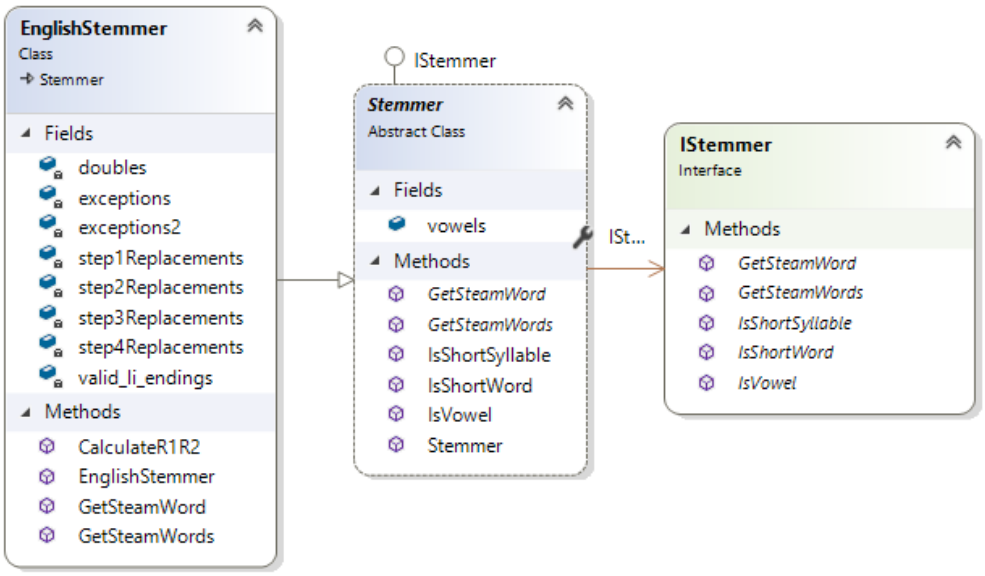
Rysunek 1. Diagram UML wygenerowany dla klas ogólnych.

- IMetric - interfejs dla metryk, klasy implementujące owy interfejs:
  - ChebyshewMetric - klasa odpowiedzialna za metrykę Czebyszewa
  - ManhatattanMetric - klasa odpowiedzialna za metrykę uliczną
  - EuclideanMetric - klasa odpowiedzialna za metrykę euklidesową



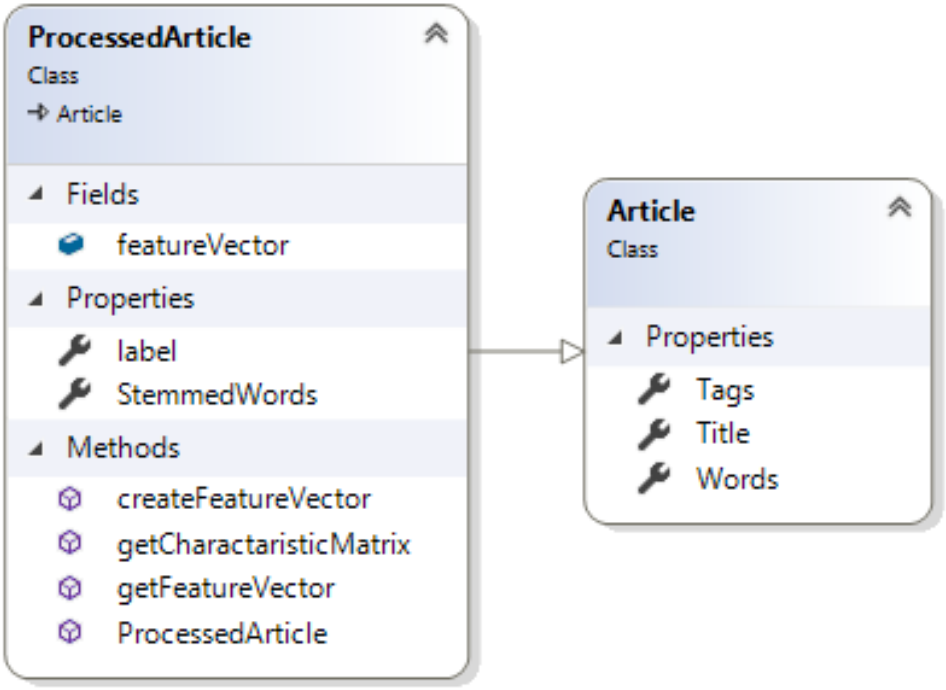
Rysunek 2. Diagram UML wygenerowany dla klas dotyczących metryk.

- IStemmer - interfejs reprezentujący stemmer
- Stemmer - reprezentacja stemera
- EnglishStemer - stemer dla języka angielskiego



Rysunek 3. Diagram UML wygenerowany dla klas dotyczących procesu stemizacji.

- Article - klasa reprezentująca pojedyńczy artykuł
- ProcessedArticle - klasa dziecicząca po Article, reprezentuje przetworzony artykuł



Rysunek 4. Diagram UML wygenerowany dla klas reprezentujących artykuły.

4. Materiały i metody

Klasyfikacja tekstów została wykonana wszystkimi dostępnymi metodami ekstrakcji cech dla wszystkich trzech metryk. Dla każdego przypadku testowego dokonano klasyfikacji tekstu dla k $\in \{2, 3, 5, 7, 10, 15, 20\}$ najbliższych sąsiadów. Wyniki porównano z faktyczną etykietą danego artykułu.

Klasyfikacja dotycząca lokalizacji przeprowadzana była jedynie na danych, których pole places przyjmowało jedną z wartości: west-germany, usa, france, uk, canada, japan. Zbiór treningowy stanowił 60% artykułów, zaś zbiór testowy 40% artykułów.

Klasyfikacja dotycząca tematów przeprowadzana była jedynie na danych, które pole topics przyjmowało jedną z wartości: gold, cocoa, sugar, coffe, grain. Zbiory testowe stanowiły 60% artykułów.

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. Wszyst-

kie 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

[1] Methods for the linguistic summarization of data - aplications of fuzzy sets and their extensions, Adam Niewiadomski, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008