Komputerowe Systemy Rozpoznawania, Informatyka, studia STACJONARNE, I st.
Projekt 1. – Klasyfikacja dokumentów tekstowych (ekstrakcja cech, miary podobieństwa tekstów)

Na ocenę 4.5:

1. Stworzyć aplikację w technologii JDK LTS (najnowsza wersja). **Główną funkcjonalnością aplikacji jest klasyfikacja zbioru dokumentów tekstowych metodą *k*-NN**. Aplikacja zawiera 2 moduły:
   1. **moduł ekstrakcji cech** operujący na dostarczonym zbiorze tekstów. Cechy pojedynczego tekstu, jako wartości rzeczywiste, tekstowe (i ew. inne) reprezentują każdy tekst w postaci wektora wartości wybranych cech. Należy wybrać minimum 10 cech i tak dobranych, aby były **niezależne od liczby obiektów/tekstów w bazie** i reprezentowały pojedyncze obiekty tylko w sposób niezmienny (a nie względem pozostałej części zbioru). Wśród cech muszą znaleźć się **min. dwie, których wartości wyrażone są tekstami** (a nie liczbami), także **niezależne od liczby obiektów/tekstów w bazie**. Zbiór tekstów do pobrania wraz z opisem: http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection
   2. **moduł klasyfikatora** *k*-NN operujący **wyłącznie** na wektorach cech reprezentujących teksty (a nie np. na dokumentach bezpośrednio, czyli przed ekstrakcją cech). Klasyfikator przyjmuje następujące parametry:
      - wartość *k*,
      - proporcje podziału zbioru wektorów na zbiór uczący i testowy (np. 60/40, 30/70, 50/50), w sposób deterministyczny, tzn. dla kilku kolejnych doświadczeń dany zbiór uczący/testowy zawiera dokładnie te same elementy (czyli wektory/teksty),
      - zbiór cech, na podstawie których dokonuje się dana klasyfikacja (nie każda klasyfikacja musi brać pod uwagę wszystkie cechy wyekstrahowane z tekstów),
      - metrykę i/lub miarę podobieństwa zastosowaną w metodzie *k*-NN (patrz niżej).

      Aplikacja posiada dowolny interfejs użytkownika do wprowadzania parametrów klasyfikacji oraz prezentacji wyników. Aplikacja wykonana w technologii Java.

2. Wykonać zadanie klasyfikacji tekstów, które w kategorii **places** posiadają etykiety: **west-germany**, **usa**, **france**, **uk**, **canada**, **japan** i są to ich jedyne etykiety w tej kategorii

3. W procesie klasyfikacji należy rozważyć następujące metryki i miary:
   1. (M1) Metryka euklidesowa
   2. (M2) Metryka uliczna
   3. (M3) Metryka Czebyszewa
   4. Wybrana miara podobieństwa tekstów (uwaga: miara nie zastępuje metryki, ale służy do porównywania, np. obliczania dystansu, tych cech, których wartości wyrażone są tekstami, należy ją uwzględnić przy oblicznaiu odległości/metryk pomiędzy wektorami) – patrz wykład.
4. W wynikach klasyfikacji należy każdorazowo podać wartości następujących miar jakości:
   1. *Accuracy* – dla całego zbioru dokumentów
   2. *Precision* – dla całego zbioru dokumentów oraz dla wybranych klas
   3. *Recall* – dla całego zbioru dokumentów oraz dla wybranych klas

4. Miara F1 – średnia harmoniczna miar *Precision* i *Recall*.
5. Porównać wyniki klasyfikacji metody *k*-NN dla 10 różnych wartości parametru *k* (wyznaczyć zależność Accuracy od *k* przy stałych wartościach innych parametrów).
6. Przy wybranej stałej wartości *k* wyznaczyć zależność Accuracy od pięciu wartości proporcji podziału zbioru (przy pozostałych parametrach stałych).
7. Wyznaczyć zależność Accuracy od wyboru metryki/miary (przy pozostałych parametrach stałych).
8. Na podstawie dowolnego wyboru 4-ch podzbiorów cech wskazać, które cechy potencjalnie mają najmniejszy, a które największy wpływ na wyniki klasyfikacji, zwłaszcza na Accuracy (przy innych wartościach stałych).

9. Dodatkowo na ocenę 5.0:
    1. Opracować własną miarę podobieństwa i/lub metrykę **oraz** porównać wyniki klasyfikacji tradycyjnie zrealizowaną *k*-NN zmodyfikowaną o nową miarę/metrykę dla różnych wartości parametru *k* (opracowana miara/metryka **powinna poprawiać uzyskiwane wyniki klasyfikacji**), lub
    2. Wybrać i zaimplementować inną metodę klasyfikacji w oparciu o te same cechy, której wyniki będą lepsze niż dla *k*-NN.

Sprawozdania należy opracować zgodnie z podanym szablonem, zob. WIKAMP KSR, Sprawozdanie-projekt-1-klasyfikacja.