Inteligentna analiza danych (laboratoria)

Projekt 1 – ekstrakcja cech, miary podobieństwa, klasyfikacja Na ocenę 5.0:

* Stworzyć aplikację do klasyfikacji zbioru tekstów metodą k-NN zawierającą:
  + Moduł ekstrakcji cech operujący na dostarczonym zbiorze tekstów. Ekstrahowane cechy muszą być tak dobrane, aby były niezależne od liczby obiektów/tekstów w bazie i reprezentowały pojedyncze obiekty w sposób niezmienny (a nie względem pozostałej części zbioru). Zbiór tekstów do pobrania wraz z opisem: <http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection>
  + Moduł klasyfikatora k-NN operujący na zbiorze wektorów cech reprezentujących teksty. 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 obiekty/teksty),
    - zbiór cech, na podstawie których dokonuje się dana klasyfikacja (nie każda klasyfikacja musi brać pod uwagę wszystkie cechy wyekstrahowane z tekstów)
    - metrykę lub miarę podobieństwa zastosowaną w metodzie k-NN (patrz niżej).
* Należy 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
* W procesie klasyfikacji należy rozważyć następujące metryki i miary:
  + (M1) Metryka euklidesowa
  + (M2) Metryka uliczna
  + (M3) Metryka Czebyszewa
* W wynikach klasyfikacji należy każdorazowo podać wartości następujących miary jakości:
  + Accuracy
  + Precision
  + Recall
* Porównać wyniki klasyfikacji metody k-NN dla 10 różnych wartości parametru k (wyznaczyć zależność Accuracy od k, przy stałych wartościach innych parametrów).
* Przy wybranej stałej wartości k wyznaczyć zależność Accuracy od pięciu wartości proporcji podziału zbioru (przy pozostałych parametrach stałych).
* Wyznaczyć zależność Accuracy od wyboru metryki/miary (przy pozostałych parametrach stałych).
* 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).

Dodatkowo na ocenę 6.0:

* Opracować własną miarę podobieństwa i/lub metrykę.
* Porównać wyniki klasyfikacji metody k-NN dla literaturowej i własnej miary podobieństwa/metryki oraz dla różnych wartości parametru k (opracowane miary/metryki **powinny poprawiać uzyskiwane wyniki klasyfikacji**).
* Wykonać klasyfikację tekstów w ramach innej kategorii niż opisana powyżej, z minimum dwiema etykietami.