Zadanie 1 - ekstrakcja cech, miary podobieństwa, klasyfikacja.

Zadanie to składa się z następujących elementów:

Na ocenę 4:

- Stworzyć szkielet aplikacji do klasyfikacji metodą k-NN na tyle uniwersalny aby był niezależny od typu obiektów, które podlegają klasyfikacji. Uniwersalność ta ma być osiągnięta w ten sposób, iż osoba korzystająca z aplikacji może dostarczyć zarówno ekstraktory cech, jak i miary podobieństwa dla stosowanego typu obiektów.
- Dla zadanego zestawu danych tekstowych zaimplementować dwa istniejące sposoby ekstrakcji wektorów cech. Odległość w wybranej metryce pomiędzy wyekstrahowanymi wektorami cech stanowić będzie miarę podobieństwa tekstów. Należy rozważyć następujące metryki.
    - (M1) Metryka euklidesowa
    - (M2) Metryka uliczna
    - (M3) Metryka Czebyszewa
- Dla zadanego zestawu danych tekstowych zaimplementować dwie istniejące miary podobieństwa. Miary podobieństwa tekstów zwyczajowo definiuje się w oparciu o miary podobieństwa słów, dlatego wystarczy rozważyć dwie różne miary podobieństwa słów.
- W obu powyższych przypadkach należy skorzystać z klasycznych metod, które można znaleźć w literaturze.
- Porównać wyniki klasyfikacji metody k-NN dla powyższych miar podobieństwa i różnych wartości parametru k.

Dodatkowo na ocenę 5:

- Dla zadanego zestawu danych tekstowych opracować własny sposób ekstrakcji cech.
- Dla zadanego zestawu danych tekstowych opracować własną miarę podobieństwa.
- Porównać wyniki klasyfikacji metody k-NN dla powyższych miar podobieństwa i różnych wartości parametru k (opracowane nowe metody powinny poprawiać uzyskiwane wyniki wcześniej klasyfikacji).

Zestawy danych do pobrania wraz z opisem znajduje się pod adresem:

http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection

Należy wykonać następujące zadania klasyfikacji:

- Klasyfikacja tekstów, które w powyższym zestawie danych w kategorii **places** posiadają etykiety: **west-germany**, **usa**, **france**, **uk**, **canada**, **japan** i są to ich jedyne etykiety w tej kategorii. W celu wyznaczenia zbioru treningowego i testowego należy wybrać odpowiednio 60% i 40% tekstów w kolejności ich występowania w powyższym zestawie (wpierw dane treningowe, a potem testowe).
- Klasyfikacja tekstów w ramach innej kategorii niż opisana powyżej z minimum dwiema

etykietami. Wybór zbioru treningowego i testowego jest dowolny (rozsądny).

- Klasyfikacja własnego zestawu tekstów. Należy opracować bazę z minimum 100 krótkimi teksami (wystarczą pojedyncze zdania) oraz z minimum 2 etykietami. Wybór zbioru treningowego i testowego jest dowolny (rozsądny)

Sprawozdania należy opracować zgodnie z podanym przez prowadzących szablonem.