Akademia Górniczo-Hutnicza w Krakowie
Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki

Inżynieria wiedzy i uczenie maszynowe

Konspekt zajęć laboratoryjnych
prowadzonych w Katedrze Informatyki
Studia Drugiego Stopnia
Drugi rok

Bartłomiej Śnieżyński

Laboratorium nr 9

Temat
System Weka – klasyfikacja tekstu

Wymagane wiadomości wstępne z wykładu
Problem klasyfikacji

Konfiguracja komputera
Podczas laboratorium wykorzystywany będzie system Weka.

Linki
http://www.cs.waikato.ac.nz/ml/weka/
http://archive.ics.uci.edu/ml/
Uwaga! Jeśli są problemy z dostępem do strony UCI można użyć proxy, np.
https://www.proxysite.com

Plan laboratorium

1. Uruchomić system Weka
2. **Przygotować plik arff z dokumentami**
   2.1. Stwórz plik z 2 atrybutami: text i class, pierwszym typu string, drugim typu {yes, no}
   2.2. Dane powinny składać się z następujących przykładów:
   'The price of crude oil has increased significantly', yes
   'Demand for crude oil outstrips supply', yes
   'Some people do not like the flavor of olive oil', no
   'The food was very oily', no
   'Crude oil is in short supply', yes
   'Use a bit of cooking oil in the frying pan', no
3. Zastosować filtr nienadzorowany „StringToWordVector" i oglądnąć efekty.
   3.1. Ile atrybutów zostało wygenerowanych?
   3.2. Jak zmieni się liczba atrybutów jeśli parametr minTermFreq = 2?
4. Klasyfikacja J48
   4.1. Wygeneruj klasyfikator J48
   4.2. Przetestuj go na poniższym zbiorze, zastępując kategorie znakiem „?", używając „FilteredClassifier" z ustawionymi „StringToWordVector" i J48 oraz ustawiając w More options parametr Output predictions = PlainText.
   Oil platforms extract crude oil
Canola oil is supposed to be healthy
Iraq has significant oil reserves
There are different types of cooking oil
   4.3. Sprawdź jak wygląda klasyfikator i jak zostały zaklasyfikowane przykłady testowe. Czy to ma sens?
5. Klasyfikacja większych zbiorów
   5.1. Powtórzyć eksperyment z danymi "ReutersCorn" oraz "ReutersGrain".
   5.2. Powtórzyć eksperyment z klasyfikatorem NaiveBayesMultinomial.
   5.3. Który klasyfikator wypada lepiej?
6. Ustawienia filtra
   6.1. Domyślnie atrybuty przyjmują wartość 0 lub 1. Sprawdź pozostałe możliwości oglądając wygląd drzewa i jego skuteczność:
      - outputWordCounts (zwraca liczbę wystąpień);
      - IDFTransform i TFTransform (jeśli oba są prawdziwe, to atrybut przyjmuje wartość TF × IDF, por. np. http://www.tfidf.com/);
      - stemmer (sprowadzanie do formy podstawowej);
      - useStopList (usuwanie niewiele mówiących słów, np. rainbow: http://www.cs.cmu.edu/~mccallum/bow/rainbow/);
      - tokenizer (np. n-gramy zamiast pojedynczych słów).
   6.2. Czy jakieś ustawienia poprawiają skuteczność klasyfikatora NaiveBayesMultinomial?