QuantUp
Artur Suchwałko
+48-511-175-949
artur@quantup.pl
http://www.quantup.pl

Program szkolenia

Text mining w praktyce z wykorzystaniem R

Prowadzący: dr inż. Artur Suchwałko

1. Przegląd: text mining and jego zastosowania (w tym sentiment analysis / opinion mining)
2. Podstawy przetwarzania tekstu w R
   - pakiet "stringr" – nowoczesne przetwarzanie tekstu w R
   - wprowadzenie do wyrażeń regularnych
   - data i czas w R (w tym pakiet "lubridate")
3. Przetwarzanie wstępne tekstu
   - wczytywanie danych tekstowych w różnych formatach
   - konwersja kodowania (pakiet "iconv")
   - usuwanie nieistotnych słów (stopwords)
   - stemming, lematyzacja (w tym użycie stemmera dla języka polskiego Morfologik)
   - normalizacja
4. Reprezentacja numeryczna dokumentów tekstowych
   - reprezentacja "bag of words"
   - macierz "document-term"
   - ocena ważności słów w macierzy "document-term" poprzez przekształcenia częstości występowania słów (w tym podejście "tf-idf")
   - określenie podobieństwa między słowami i dokumentami (w tym odległość Levenshteina)
5. Analiza dokumentów tekstowych
   - wizualizacja i redukcja wymiaru (PCA)
   - modelowanie predykcyjne z wykorzystaniem metod klasyfikacji (drzewa klasyfikacyjne, SVM, inne)
   - klasyfikacja bayesowska (jak w przypadku filtrów antyspamowych)
   - regresja
   - znajdowanie grup podobnych dokumentów: analiza skupień (k-means, PAM, metody hierarchiczne)
   - Latent Semantic Indexing z wykorzystaniem Singular Value Decomposition (opcjonalne)
6. Text mining z R w praktyce
   - analizy opisowe (w tym word clouds)
   - praca z pakietem "tm"
   - automatyczna klasyfikacja dokumentów tekstowych z pakietem "RTextTools"
   - śledzenie historii występowania słów i fraz