

Kod szkolenia: **ANA/TXT**

Tytuł szkolenia: **Analiza danych tekstowych i języka naturalnego (Python)**

Dni: 3



Partner merytoryczny

Opis:

Adresaci szkolenia

Dane tekstowe stanowią co najmniej 70% wszystkich danych generowanych w systemach informatycznych, a dodatkowo są to dane rzadko wykorzystywane w celu analizy i odkrywania wiedzy. Szkolenie ma przybliżyć problemy przetwarzania i analizy danych tekstowych. Szkolenie skierowane jest do:

- programistów, pragnących zastosować w swoich systemach metody odkrywania wiedzy z danych tekstowych
- dla analityków, którzy chcą rozbudować swój warsztat analityczny o narzędzie analizy danych tekstowych
- osób zainteresowanych zastosowaniem narzędzi statystycznych, metod uczenia maszynowego w pracy z danymi tekstowymi

Wymagana podstawowa wiedza z programowania w dowolnym języku (np. Python, R, matlab itp).

Cel szkolenia

Nauczenie szeregu narzędzi do pracy z danymi tekstowymi, przedstawienie szeregu przykładów użycia pokrywających większość tematów tej dziedziny.

Mocne strony szkolenia

Dużo przykładów użycia do wykorzystania w życiu/pracy, szerokie zapoznanie słuchacza z dziedziną analizy danych tekstowych, i możliwościami jej wykorzystania w pracy

Wymagania

Minimalne doświadczenie z programowaniem, doświadczenie w analizie danych.

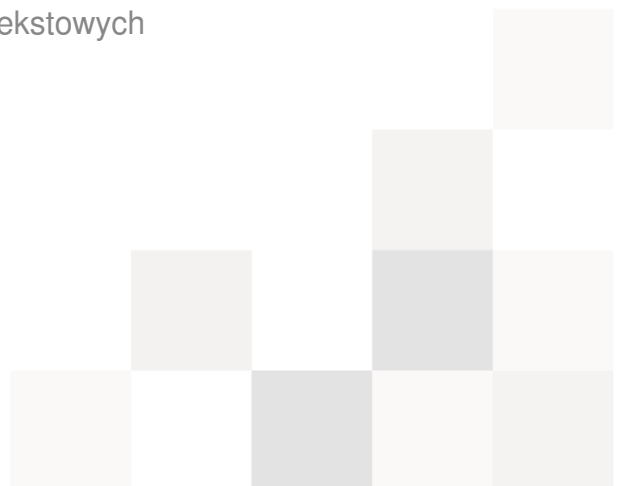
Parametry szkolenia

3*8 godzin (3*7 godzin netto) wykładów i warsztatów (z wyraźną przewagą warsztatów).

Wielkość grupy: maks. 8-10 osób.

Program szkolenia:

1. Praca z danymi tekstowymi
 - Dane tekstowe - ich charakterystyka, trendy
 - Analiza danych tekstowych, eksploracja tekstu, a przetwarzanie języka naturalnego
 - Krajobraz dziedziny - spacer po różnych obszarach i ich zastosowaniach
 - Języki programowania do pracy nad analizą danych tekstowych
 - Data Scientist - zawód, który głównie pracuje z danymi tekstowymi
2. Pozyskiwanie i wstępne przetwarzanie danych
 - Wprowadzenie do Python
 - Pakiet Pandas
 - Pakiet NLTK
 - Pakiet scikit-learn
 - Czytanie danych
 - istniejące korpusy
 - z pliku tekstowego
 - z katalogu plików
 - z Internetu
 - Czyszczenie i normalizacja danych
 - usuwanie nieistotnych słów tzw. stop words
 - usuwanie znaków specjalnych, przestankowych oraz liczb
 - sprowadzanie do małych liter
 - stemming/lematyzacja
 - Przykłady czytania danych z dobrze zdefiniowanych API (np. Twitter)
 - Pobieranie danych ze stron internetowych (Web scraping)
 - Parsowanie HTML z użyciem Python
3. Wizualizacja danych tekstowych
 - Opis problemu
 - Dostępne pakiety wizualizacji w Python
 - Przekrój metod wizualizacji danych tekstowych
 - word length counts plot,
 - word frequency plots,
 - word clouds,
 - correlation plots,
 - letter frequency plot,
 - letter position,
 - heatmap



4. Eksploracja danych tekstowych

- Budowanie macierzy Term-Document
- Wyszukiwanie słów kluczowych charakterystycznych dla dokumentów
- Mierzenie podobieństwa między dokumentami i słowami kluczowymi
 - Miara Cosinusowa
 - Miara Jaccarda
- Grupowanie tekstów
 - Metody data-centric
 - Hierarchical Agglomerative Clustering,
 - K-means,
 - Metody description-centric
 - Carrot2
- Klasyfikacja na przykładzie detekcji spamu
 - K Nearest Neighbours,
 - SVM,
 - Naive Bayes

5. Uczenie maszynowe w analizie języka naturalnego

- Znakowanie tekstu częściami mowy
- Analiza wydźwięku (sentiment analysis)
 - podejście słownikowe,
 - podejścia oparte na metodach uczenia maszynowego.
- Rozpoznawanie nazw własnych (ang. Named Entity Recognition)
- Semantyczne podobieństwo słów i tekstów
- Wykrywanie fraz (np. rzeczownikowych czy czasownikowych)
- Drzewa rozkładu
 - Penn TreeBank
 - Składnica

