sages

Course code: **ANA/TXT**

Course title: **Foundations of Text Mining and Natural Language Processing**

Days: 3

Description:

Course intended for

Text Mining constitute at least 70% of all data generated in IT systems. Such data is rarely used for analytical purposes or knowledge discovery. This course covers the problems related to the processing and analysis of Text Mining. The course is addressed to:

- programmers who wish to use the knowledge discovery methods using Text Mining in their systems,
- analysts who wish to develop their analytical workshop by a Text Mining analysis tool,
- those interested in using statistical tools and machine learning methods when working with Text Mining.

Basic programming knowledge in any language is required (for example Python, R, matlab etc.).

Course objective

Participants will learn a number of tools designated for working with Text Mining. A number of examples of their use will be presented which cover the majority of topics from that domain.The basic languages in working with texts will be presented: R, Python and Java.

Course strengths

Many examples of practical application at work will be provided. The participants become familiar with Text Mining analysis and the possibilities of using it at work.

Requirements

Minimum programming experience, experience in data analysis.

Course parameters

3 working days, 3*7 working hours, group 8-10 people. The course contains presentations and coding workshops.

Course curriculum:

1. Working with Text Mining
    - Text Mining - characteristics, trends
    - Text Mining analysis and discovering knowledge from Text Mining
    - Domain presentation - discussion of various areas and their applications
    - Programming languages designated for working on Text Mining analysis
    - Data Scientist - a profession comprising mainly of working with Text Mining
2. Initial data processing and simple statistical tools
    - Introduction to R
    - 'Tm' package for working with texts
    - Reading data
        - existing corpuses, for example crude, acq.
        - from file folder
        - from text file
        - from the Internet
    - Cleaning and normalisation of data
        - removing insignificant words,("stop words")
        - removing bullet points and digits
        - changing letters to lowercase
        - stemming/lemmatisation
    - Creating Term-Document matrix

- Finding common terms
- Finding associations
- Removing rare terms
- Measurement of similarity between documents and terms
  - Cosinus measure
  - Jaccard index
- Visualisation of term significance in the form of word clouds
- Tagging text with parts of speech
- Examples of initial text processing on series of StackOverflow entries, crude, acq corpuses, or data from the Internet
- Examples of reading data from properly defined API (for example TwiiterR)
- Web scrapping using R with the example of downloading an aggregating NHL statistics
- HTML parsing using R

3. Advanced data processing and visualisation
   - Sentiment analysis
     - vocabulary approach
     - based on Bayesian probability methods
   - Name entity recognition
   - Detecting phrases (for example noun or verb phrases)
   - Distribution trees
     - Penn TreeBank
     - Repository
   - Methods of data visualisation in R

        - word length counts plot
        - word frequency plots
        - word clouds
        - correlation plots
        - letter frequency plot
        - letter position
        - heatmap
    - Grouping texts using different methods
        - Data-centric methods
            - Hierarchical Agglomerative Clustering
            - K-means
        - Description-centric methods
            - Carrot2 and Yippy
            - SnSRC
    - Classification on the basis of spam detection
        - K Nearest Neighbours
        - SVM
        - Naive Bayes
    - Semantic similarity between texts
4. Text Mining processing - practical examples taken from commercial projects
    - Python and NLTK in a few steps using examples: detecting patterns in Text Mining, creating your own vocabularies, full text search using Apache Lucene, coexistence measure as the basis for measuring semantic similarity.
    - Creating bag-of-words using parts of speech and your own names filter

- Induction of word meaning and grouping results according to meanings
- Creating data extractors in Java, for example ScholarExtractor
- Extraction of keywords from texts in Java
- Publication classification according to OSJ taxonomy in Java
- Searching for similar studies on the basis of their competencies recorded in doc(x)/pdf files (text processing using Apache Tika and extraction of symbols from texts in order to build a structured vector representation, Jaccard index as an alternative for Cosinus index)
- Semantic development using Java and knowledge resources (for example Wikipedia and BabelNet)

Sages sp. z o.o., ul. Nowogrodzka 62c, 02-002 Warszawa
tel. +48 22 203 56 00 fax: +48 22 203 56 01
e-mail: office@sages.io web: www.sages.io