COURSE DESCRIPTION CARD - SYLLABUS

Course name
Text Mining (Text Data Exploration) [S1DSwB1>EDT]

Course

Field of study
Data Science in Business

Year/Semester
3/5

Area of study (specialization)
–

Profile of study
general academic

Level of study
first-cycle

Course offered in
Polish

Form of study
full-time

Requirements
elective

Number of hours

Lecture
15

Laboratory classes
0

Other
0

Tutorials
15

Projects/seminars
0

Number of credit points

3,00

Coordinators

dr hab. Grzegorz Pawłowski
grzegorz.pawlowski@put.poznan.pl

Lecturers

Prerequisites

Students should have a basic knowledge of statistics, linear algebra, and probability calculus, as well as the ability to work with data. Familiarity with basic programming (preferably Python) and fundamental data analysis skills is recommended. A general understanding of databases and SQL, as well as basic machine learning concepts, will also be useful.

Course objective

The aim of the course is to equip participants with theoretical foundations and practical skills in text mining, enabling effective analysis, interpretation, and utilization of information contained in large text datasets using NLP tools and machine learning techniques.

Course-related learning outcomes

Knowledge:
Characterizes fundamental text mining methods, including tokenization, normalization, vectorization, and semantic analysis [DSB1_W01].
Identifies key natural language processing (NLP) techniques such as sentiment analysis, topic modeling, and information extraction [DSB1_W02].

Explains the application of text mining methods in the context of machine learning, big data, and recommendation systems [DSB1_W03].

Skills:
Analyzes text data using appropriate NLP and text mining methods, such as TF-IDF, Word Embeddings, document classification, and clustering [DSB1_U02].
Designs and implements text processing models using open-source tools and big data environments [DSB1_U03].
Utilizes text mining techniques for opinion analysis, document classification, and building information retrieval systems [DSB1_U08].

Social competences:
Considers ethical and legal aspects of text data processing, including privacy issues and the reliability of analysis [DSB1_K01].
Takes responsibility for the quality and interpretation of text mining results in business and scientific applications [DSB1_K04].

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Lecture:
Students take two quizzes, each graded on a point-based system-50 points per quiz. The final grade is the sum of points from both quizzes. The first quiz takes place midway through the course, while the second is at the end. The passing threshold is a total of 50 points from both quizzes.
Laboratories:
Students receive two formative assessments. The first assessment involves preparing a report on solving a selected business problem related to text mining, as assigned by the instructor, worth up to 50 points. The second assessment is a quiz, also worth up to 50 points. The passing threshold is a total of 50 points from all activities.

Programme content

The course covers comprehensive topics in natural language processing, including tokenization, normalization, and text representation using methods such as TF-IDF and Word Embeddings. It focuses on text analysis techniques, classification, topic modeling, sentiment analysis, and the application of neural networks and deep learning in NLP. Participants will explore methods for information extraction, semantic analysis, and result visualization, as well as practical applications of open-source tools in big data environments, while also considering the ethical and legal aspects of text data processing.

Course topics

Introduction to Text Mining
Fundamentals of Natural Language Processing (NLP)
Text Tokenization and Segmentation Techniques
Text Normalization: Stop-word Removal, Stemming, Lemmatization
Text Representation: Vectorization, TF-IDF, Word Embeddings
Word Frequency Analysis and N-grams
Text Classification - Supervised and Unsupervised Methods
Sentiment and Emotion Analysis in Text
Topic Modeling and Topic Detection
Information Extraction: Named Entity Recognition
Language Models and Text Generation
Neural Networks and Deep Learning in NLP
Semantic Analysis and Word Relationships
Duplicate Detection and Document Clustering
Application of Text Mining in Opinion Analysis
Information Retrieval Techniques and Recommender Systems
Text Processing in Big Data Environments
Open-Source Tools and Libraries for Text Mining
Visualization of Text Analysis Results

Ethical and Legal Aspects of Text Data Processing

Teaching methods

Lectures: problem-based lecture, case study presentation
Laboratories: analysis of real-world data, practical group tasks, case study

Bibliography

Basic:
Spinczyk, D., Dzieciątko, M. (2016). Text Mining: metody, narzędzia, zastosowania. PWN
Weiss, D. (2006). Descriptive clustering as a method for exploring text collections. Unpublished doctoral dissertation, Poznan University of Technology, Poznan, Poland.
Russell, A., Klassen, M. (2019). Data Mining. Eksploracja danych w sieciach społecznościowych. Wydanie III, Helion
Rajnish, R., Srivastava, M. (2023). Web Data Mining z użyciem języka Python. Odkrywaj i wyodrębniaj informacje ze stron internetowych za pomocą języka Python. APN Promise

Additional:
Zastosowania statystyki i data mining w badaniach naukowych, Red. Jakubowski, J., Wątroba, J. (2017), StatSoft Polska. Kraków

Breakdown of average student's workload

	Hours	ECTS
Total workload	75	3,00
Classes requiring direct contact with the teacher	30	1,00
Student's own work (literature studies, preparation for laboratory classes/ tutorials, preparation for tests/exam, project preparation)	45	2,00