University of Gdańsk

Subject card

Subject name and code	Text and Social Network Mining, PG_00178724		
Field of study	Informatics and Econometrics		
Date of commencement of studies	October 2026	Academic year of realisation of subject	2027/2028
Education level	Master's studies	Subject group	Obligatory subject group in the field of study Optional subject group Subject group related to scientific research in the field of study
Mode of study	part-time studies	Mode of delivery	at the university
Year of study	2	Language of instruction	Polish
Semester of study	3	ECTS credits	5.0
Learning profile	academic	Assessment form	exam
Conducting unit	Department of Statistics -> Faculty of Management -> Rector		
Name and surname of lecturer (lecturers)	Subject supervisor	dr hab. Krzysztof Najman	
	Teachers		

Lesson types

Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
Number of study hours	16.0	0.0	16.0	0.0	0.0	32

E-learning hours included: 0.0

Learning activity and number of study hours

Learning activity	Participation in didactic classes included in study plan	Participation in consultation hours	Self-study	SUM
Number of study hours	32	2.0	91.0	125

Subject objectives

Familiarizing students with theoretical and practical aspects of text data mining.
Familiarizing students with the process of text mining analysis, searching unstructured documents and practical applications of text mining.
Familiarizing students with theoretical and practical aspects of building and analyzing social networks as structures describing relationships between entities (people, organizations, objects).

Learning outcomes	Course outcome	Subject outcome	Method of verification
	[IiEMU2_W06] Possesses a structured understanding of the processes, methods, and tools necessary for the design, creation, development, and provision of suitable conditions for informatics, econometrics or statistics tools	The student knows the principles of construction, verification and evaluation of selected methods of Text Mining and social network analysis.	[SW4] test/exam - oral or written [SW2] presentation/project/paper/ report
	[IiEMU2_U03] Is able to obtain and verify data from properly selected sources and to collect, process, and visualize it using modern econometrics, informatics or statistics tools	The student can collect, record, verify, and evaluate data necessary for Text Mining and social network analysis.	[SU2] presentation/project/paper/ report
	[IiEMU2_U06] Can utilize structured and detailed knowledge of management, quality sciences, economics, and finance to address dilemmas and develop innovative solutions for complex or unusual problems that arise in professional settings	The student is able to build and apply selected models of Text Mining and social network analysis in practice.	[SU2] presentation/project/paper/ report
	[IiEMU2_W08] Possesses a comprehensive understanding of the methods, conditions, directions, and dilemmas involved in applying advanced econometrics, informatics or statistics tools in response to dynamic environmental changes	The student knows the theoretical foundations of Text Mining analysis and social network analysis.	[SW4] test/exam - oral or written [SW2] presentation/project/paper/ report

Subject contents	**Definition, essence, purpose, scope and benefits of using text mining** (text data mining). **Areas and examples of text mining applications**. **Text mining analysis process**. Defining the purpose, scope and costs of analysis, determining sources of unstructured data, transforming source documents into a text file. Lexical process - marking parts of speech. Extraction of expression roots (stemming) automatic finding of roots or word stems, assigning words to their base form. Parsing - excluding words with low information value. Tagging - selection of morphosyntactic description. Lemmatization - identification of lexemes. Construction of stop lists, removal of irrelevant words (tags). Pruning - removal of unnecessary words: an approach based on removing expressions that occur most frequently (most frequent) and least frequently (least frequent). Construction of a frequency matrix of words in a document, text filtering and extraction of facts (document-expression matrix, term-document matrix). Representation of text documents, vector space model (VSM) and graph representation: standard, simple, n-distance, simple n-distance, absolute frequency, relative frequency. Using BERT-type language models to represent text data. **Searching unstructured documents**: full text search FTS (Full Text Searching). Term-by-document frequency matrix. Weighting words. Methods for determining the number of occurrences of a given expression in a set of documents (collections). **Topic modeling of text documents**. Introduction to probabilistic models: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA). Interpreting topics as distributions of words and documents as distributions of topics.Extracting topics using methods based on BERT-type language models, such as BERTopic. **Determining the emotional nature of statements** (sentiment analysis). Lexicon-based and statistical-based approaches: learning classification models (e.g. Naive Bayes Classifier) from labeled data sets. Sentiment assessment on various types of text representations. **Definition, essence and characteristics of social networks** as structures describing relationships between entities (persons, organizations, objects). Network components: nodes (actors) and edges (relations). Examples of applications of social networks in social, information and business contexts. The importance of networks in the analysis of social structures, information flow, group dynamics and social influence. **Basic concepts graph theory**: directed and undirected graph, weighted graph, graph connectivity, path, cycle. Converting relational data to network format. Representation of relations in a graph using: a simple list of relations (linked list), a weighted list of relations (edge list), an adjacency matrix (adjacency matrix), an incident matrix (incidence matrix). Principles of constructing graphs from text data. **Techniques for presenting network structure** using graphical tools. Aesthetics and readability of visualizations: application of different layouts, coloring, scaling of nodes and edges. Interpretation of visualizations as a tool supporting analysis of results and extraction of conclusions. **Identification and interpretation of the position of actors** (nodes) in the structure of a social network. Centrality measures: degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. Application of centrality indicators to detect influential users and assess their importance in the context of the flow of information, opinions and resources in the network. Analysis of the role of actors in the network. **Study of topological properties of networks**: connectivity, density, average path length, average degree of vertices. Using metrics to assess network structure, analyze atypical areas and nodes critical to its integrity. Assess network resilience to disruptions and assess its effectiveness in the context of information propagation. **Analysis of node groups**. Extraction of key structures such as the largest components, cliques and n-cliques. Review of community detection methods, including Girvan-Newman, Louvain algorithms and modularity methods. Application of community detection techniques to user segmentation and identification of cohesive groups in social networks. **Use of Markov models to predict message coverage**, identify potential paths of message propagation in the network, detect anomalies and predict further user behavior in the network. Simulation of information propagation using a random graph walk.
Prerequisites and co-requisites	Fundamentals of linear algebra, statistics, and multivariate statistical analysis.
Assessment methods and criteria	

Subject passing criteria	Passing threshold	Percentage of the final grade
semester project	51.0%	60.0%
written exam	51.0%	40.0%

Recommended reading	Basic literature	Feldman R., Sanger J., 2006, Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press New York Berry M.W., Kogan J., 2010, Text Mining, Applications and Theory, A John Wiley and Sons, Ltd. Solka J. L., 2008, Text Data Mining: Theory and Methods, Statistic Survey Majkowska, A., Migdał-Najman, K., Najman, K., i Raca, K. 2021. Identification of the words most frequently used by different generations of twitter users. Data analysis and classification: methods and applications, Springer, 2747. R. A. Hanneman, M. Riddle, Introduction to social network methods. Riverside, CA: University of California, Riverside, 2005. P.J. Carrington, J. Scott, S. Wasserman (eds.), Models and Methods in Social Network Analysis. Cambridge University Press, 2005. M. Kamola, P. Arabas, Sieci społeczne i technologiczne. Jak zrozumieć, jak wykorzystać, wyd. PWN, 2018 M. A. Russell, M. Klassen, Data mining. Eksploracja danych w sieciach społecznościowych, wyd. Helion, 2019
	Supplementary literature	P.Lula, Statystyczne modelowanie zawartości dokumentów tekstowych, WUEK, 2018.
	eResources addresses	
Example issues/ example questions/ tasks being completed		
Work placement	Not applicable	

Document generated electronically. Does not require a seal or signature.