

Subject card

Subject name and code	Text Mining, PG_00157278						
Field of study	Informatics and Econometrics						
Date of commencement of studies	October 2024		Academic year of realisation of subject		2025/2026		
Education level	Master's studies		Subject group		Obligatory subject group in the field of study 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	4		ECTS credits		1.0		
Learning profile	academic		Assessment form		credit		
Conducting unit	Department of Statistics -> Faculty of Management -> Rector						
Name and surname of lecturer (lecturers)	Subject supervisor		dr hab. Krzysztof Najman				
	Teachers		dr hab. Krzysztof Najman				
Lesson types	Lesson type	Lecture	Tutorial	Laboratory	Project	Seminar	SUM
	Number of study hours	0.0	8.0	0.0	0.0	0.0	8
	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	8		15.0		30.0	53
Subject objectives	Familiarize students with theoretical and practical aspects of text data mining.Familiarize students with the process of text mining analysis, searching unstructured documents and practical applications of text mining.						

Learning outcomes	Course outcome	Subject outcome	Method of verification
	[liEMU2_W04] The student has an in-depth knowledge of advanced mathematical, statistical, econometric and IT methods that enable the acquisition, processing and analysis of data reflecting the functioning and growth of the national economy and its components, as well as the phenomena and processes occurring in their environment.	Has in-depth knowledge of advanced data analysis methods using selected Text Mining methods.	[SW1] oral statement/ conversation/discussion [SW2] presentation/project/paper/ report
	[liEMU2_U02] The student is able to proficiently acquire detailed information about economic processes and phenomena through direct observation, planned experimentation or database queries, as well as collect and process it using modern information technology tools.	The student is able to collect statistical data from primary and secondary sources for the purposes of implementing social research using Text Mining models.	[SU1] oral statement/conversation/ discussion [SU2] presentation/project/paper/ report
	[liEMU2_K02] The student can communicate freely with the public on specialised topics in the field of computer science and econometrics in and outside the workplace, communicate his knowledge and share his skills through various media. Culturally participates in discussions, is not afraid to ask questions and knows how to give constructive criticism.	Is able to discuss, communicate and share knowledge regarding the theory and practice of using Text Mining models.	[SK1] oral statement/conversation/ discussion [SK2] presentation/project/paper/ report
	[liEMU2_K04] The student can work in a team, co-create it, effectively manage and supervise it; efficiently adapts his behavior and conduct to his role in it; is ready to take responsibility for the team and bear the consequences; understands the necessity of systematicity and consistency in action; is open to other team members and critical of himself.	The student is able to work responsibly in a team, perform various roles within it, critically evaluate his/her own work and that of other team members.	[SK2] presentation/project/paper/ report
	[liEMU2_K03] The student is able to communicate freely with the public inside and outside the workplace, transfer his knowledge and share his skills through various media.	Is able to communicate and transfer knowledge and skills in the field of Text Mining.	[SK1] oral statement/conversation/ discussion [SK2] presentation/project/paper/ report
	[liEMU2_W05] The student has an in-depth knowledge of socio-economic data sources, their databases and how to create them.	The student knows the sources and methods of searching for secondary data used in social research using Text Mining methods.	[SW1] oral statement/ conversation/discussion [SW2] presentation/project/paper/ report
	[liEMU2_K01] The student understands the need for continuous completion and deepening of acquired knowledge. The student inspires and organizes others' learning processes.	Understands the need to constantly update knowledge in the field of Text Mining.	[SK1] oral statement/conversation/ discussion [SK2] presentation/project/paper/ report

Subject contents	Definition, essence, purpose, scope and benefits of using text mining (text data mining). Relationships of text mining with related fields, interdisciplinary nature of text mining, discussion of relationships with data mining, machine learning, natural language processing (NLP), machine translation (MT), information retrieval systems (IR), information extraction (IE), statistics and computer science. Areas and examples of text mining applications: obtaining information from documents, identifying messages containing specific content, pattern classification, patternless classification, identification of relationships, visualization, generating answers to questions. Text mining analysis process. Defining the purpose, scope and costs of the analysis, defining the sources of unstructured data, transforming source documents into a text file (preprocessing): transforming documents into text, removing formatting characters, unifying the way characters are encoded, tokenization identifying basic text phrases, dividing the text into sentences, tokens, words. Lexical process marking parts of speech. Extraction of expression roots (stemming) automatic finding of roots or word stems, assigning words their base form. Parsing excluding words with low information value. Tagging selecting a morphosyntactic description that is appropriate in the specific context of using a given form. Lemmatization morphological analysis limited to finding the base form of a word (lexeme identification). Stop list (stop list) removal of irrelevant words (tags) and collection of semantically irrelevant parts of speech in a separate list. Pruning (removal of unnecessary words): an approach based on removing expressions that occur most frequently and least frequently. Methods of storing information in a document in a computer system: based on the frequency of occurrence of individual words in the document, striving to give information obtained from documents a structured form. Input data format: .txt, .html, less often .doc or .pdf. Frequency matrix of words in a document, text filtering and fact extraction (term-document matrix). Representation of text documents - vector representation (vector space model VSM) and graph representation: standard, simple, n-distance, simple n-distance, absolute frequency, relative frequency. Searching unstructured documents: full-text search FTS (Full Text Searching), using a special type of indexes, the so-called full-text indexes. Term-by-document frequency matrix. Weighting words in the matrix: Frequency Weight method for determining the frequency of occurrence of specific phrases in a document: binary, logarithmic, without modification (none) and Term Weight method for determining the number of occurrences of a given expression in a set of documents (collection), methods: Entropy, IDF (Inverse Document Frequency), GF-IDF (Global Frequency - Inverse Document Frequency), Normal, Chi-Squared, Mutual Information, Information Gain.		
Prerequisites and co-requisites	Basic knowledge of linear algebra, statistics, multivariate statistical analysis.		
Assessment methods and criteria	Subject passing criteria	Passing threshold	Percentage of the final grade
	final project	51.0%	100.0%
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. Maimon O., Rokach L., 2005, The Data Mining and Knowledge Discovery Handbook, Springer Solka J. L., 2008, Text Data Mining: Theory and Methods, Statistic Survey	
	Supplementary literature	.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.