Syllabus
Course Program

Analysis and Synthesis of Natural Language Processing

Specialty
113 Applied mathematics

Institute
Educational and Scientific Institute of Computer Science and Information Technology

Educational program
Intelligent Data Analysis

Department
Computer Mathematics and Data Analysis

Level of education
Master's level

Course type
Special (professional), Selective

Semester
1

Language of instruction
English

Lecturers and course developers

Valentyna Pikalova

valentyna.pikalova@khpi.edu.ua
Candidate of Pedagogical Science (Information and Communication Technologies in Education), Assistant Professor of Computer Mathematics and Data Analysis Department

Work experience - more than 20 years. The author of more than 50 scientific, educational, and methodological works.

More about the lecturer on the department's website

General information

Summary

Natural language processing is a subfield of linguistics, computer science, and artificial intelligence that uses algorithms to interpret and manipulate human language.
In the course, you will learn how to design NLP applications that perform question-answering and sentiment analysis, create tools to translate languages, summarize text, and even build chatbots.

Course objectives and goals

The discipline is aimed at acquiring the necessary competencies in the field of natural language processing, machine learning and artificial intelligence

Format of classes

Lectures, laboratory classes, consultations, self-study. Final control in the form of an exam.

Competencies

GC 3. Ability for continuous learning, acquiring new knowledge and skills, including in areas other than professional ones.
GC 4. Ability to identify, pose, and solve problems in professional activities.

GC 7. Ability to work with information, find and use information from various sources necessary for solving professional tasks.
SC 1. Ability to formulate a mathematical problem, relying on the language of the subject area, verifying the correctness of the formulation, including under conditions of uncertainty.
SC 2. Ability to choose, develop, and investigate mathematical, analytical, or numerical methods for solving practical problems that ensure the required accuracy and reliability of the result.
SC 5. Ability to conduct mathematical and computer modeling and computational experiments, collect, visualize, analyze, and process obtained data, solve formalized problems using specialized software tools.
SC 7. Ability to search, study, and analyze scientific and technical information, domestic and foreign experience related to the application of mathematical methods for the study of processes and systems.
SC 10. Ability to choose, develop, investigate, and apply mathematical models and methods for intelligent data analysis under conditions of uncertainty.
SC 12. Ability to develop and operate specialized software tools for intelligent data analysis of texts, signals, and images.

Learning outcomes

LO 1. Demonstrate knowledge and understanding of the fundamental and applied mathematics concepts, principles, and theories and apply them in practice.
LO 2. Ability to formalize problems formulated in the language of a specific subject area, choose a rational method for solving them, solve problems using analytical or numerical methods, assess the accuracy and reliability of the results, and interpret them.
LO 5. Develop algorithms that are efficient in terms of accuracy, stability, speed, and resource consumption for numerical investigation of mathematical models and data analysis, decision-making.
LO 7. Apply modern programming technologies and software development, implement numerical and symbolic algorithms.
LO 12. Know and understand modern methods for solving mathematical problems of statistical and intelligent data analysis, forecasting, etc.
LO 14. Ability to apply existing and develop new algorithms and software tools for statistical and intelligent analysis of uncertain data.
LO 15. Ability to apply existing and develop new algorithms and software tools for data processing, text, signals, and images.

Student workload

The total volume of the discipline is 150 hours. (5 ECTS credits): lectures – 32 hours, laboratory work – 32 hours, independent work – 86 hours.

Course prerequisites

"Programming", "Mathematical analysis", "Linear algebra", "Probability theory", "Mathematical statistics", "Functional analysis", "Mathematical statistics", "Optimization methods", "Methods of machine learning"

Features of the course, teaching and learning methods, and technologies

When teaching this discipline, such teaching and learning methods as gamification and peer-to-peer are used. LMS (learning management systems) systems are used in the learning process.

Program of the course

Topics of the lectures

Topic1: Logistic Regression for Sentiment Analysis of Tweets
Topic2: Naïve Bayes for Sentiment Analysis of Tweets
Topic3: Vector Space Models
Topic4: Word Embeddings and Locality Sensitive Hashing for Machine Translation
Topic5: Auto-correct using Minimum Edit Distance
Topic6: Part-of-Speech (POS) Tagging

Topic7: N-gram Language Models
Topic8: Word2Vec and Stochastic Gradient Descent
Topic9: Sentiment with Neural Nets
Topi10: Language Generation Models
Topic11: Named Entity Recognition (NER)
Topic12: Siamese Networks
Topic13: Neural Machine Translation with Attention
Topic14: Summarization with Transformer Models
Topic15: Question-Answering with Transformer Models
Topic16: Chatbots with a Reformer Model

Topics of the workshops

Workshops are not provided within the discipline.

Topics of the laboratory classes

Topic1: Use a simple method to classify positive or negative sentiment in tweets
Topic2: Use a more advanced model for sentiment analysis
Topic3: Use vector space models to discover relationships between words and use principal component analysis (PCA) to reduce the dimensionality of the vector space and visualize those relationships
Topic4: Write a simple English-to-French translation algorithm using pre-computed word embeddings and locality sensitive hashing to relate words via approximate k-nearest neighbors search
Topic5: Create a simple auto-correct algorithm using minimum edit distance and dynamic programming
Topic6: Apply the Viterbi algorithm for POS tagging, which is important for computational linguistics
Topic7: Write a better auto-complete algorithm using an N-gram model (similar models are used for translation, determining the author of a text, and speech recognition)
Topic8: Write your own Word2Vec model that uses a neural network to compute word embeddings using a continuous bag-of-words model
Topic9: Train a neural network with GLoVe word embeddings to perform sentiment analysis of tweets
Topic10: Generate synthetic Shakespeare text using a Gated Recurrent Unit (GRU) language model
Topic11: Train a recurrent neural network to perform NER using LSTMs with linear layers
Topic12: Use so-called 'Siamese' LSTM models to compare questions in a corpus and identify those that are worded differently but have the same meaning
Topic13: Translate complete English sentences into French using an encoder/decoder attention model
Topic14: Build a transformer model to summarize text
Topic15: Use T5 and BERT models to perform question answering
Topic16: Build a chatbot using a reformer model

Self-study

The course involves the completion of individual tasks, the results of which are monitored and assessed by teachers. Students are also recommended additional materials (videos, articles) for self-study.

Course materials and recommended reading

1. R.N. Rao. Machine Learning in Data Science Using Python. - Dreamtech Press, 2022. - 956 p. ISBN 978-939-154-046-3
2. P. Chatterjee, M. Yazdani, F. Fernández-Navarro, J. Pérez-Rodríguez. Machine Learning Algorithms and Applications in Engineering. – New York: Taylor & Francis, 2023. – 314 p. ISBN 978-036-756-912-9
3. A. Burkov. Machine Learning Engineering. – True Positive Inc., 2020. – 310 p. ISBN 978-199-957-957-9
4. M. Kubat. An Introduction to Machine Learning. - Springer Cham, 2021. – 458 p. ISBN 978-303-081-935-4
5. M. Gori, A. Betti, S. Melacci. Machine Learning. - Morgan Kaufmann, 2024. - 580 p. ISBN 978-032-389-859-1 https://doi.org/10.1016/C2020-0-03158-0

Assessment and grading

Criteria for assessment of student performance, and the final score structure

Description of the final score structure, course requirements, and necessary steps to earn points, especially paying attention to self-study and individual assignments.

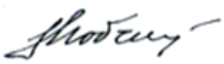
Grading scale

Total points	National	ECTS
90–100	Excellent	A
82–89	Good	B
75–81	Good	C
64–74	Satisfactory	D
60–63	Satisfactory	E
35–59	Unsatisfactory (requires additional learning)	FX
1–34	Unsatisfactory (requires repetition of the course)	F

Norms of academic integrity and course policy

The student must adhere to the Code of Ethics of Academic Relations and Integrity of NTU "KhPI": to demonstrate discipline, good manners, kindness, honesty, and responsibility. Conflict situations should be openly discussed in academic groups with a lecturer, and if it is impossible to resolve the conflict, they should be brought to the attention of the Institute's management.
Regulatory and legal documents related to the implementation of the principles of academic integrity at NTU "KhPI" are available on the website: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

Approval

Approved by	Date, signature 31.08.2023		Head of the department Olena AKHIIEZER
	Date, signature 31.08.2023		Guarantor of the educational program Leonid LYUBCHYK