Syllabus
Course Program

Processing and analysis of textual information

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
Bachelor's level

Course type
Special (professional), Selective

Semester
8

Language of instruction
Ukrainian

Lecturers and course developers

Костюк Ольга Василівна

Olha.Kostiuk@khpi.edu.ua
Candidate of Technical Sciences, Associate Professor of the Department of Computer Mathematics and Data Analysis of KhPI National Technical University

Author of scientific and educational and methodical works. Leading lecturer in the disciplines: "Decision theory", "Infrastructure and management of big data", "Introduction to the specialty and engineering activity"

General information

Summary

The main goal of the educational discipline "Processing and analysis of textual information" is to acquire basic knowledge in the field of processing and analysis of textual information, as well as to acquire skills in solving problems that arise during the development of linguistic data processing systems. The content of the course is aimed at familiarizing students with the basic concepts of natural language processing, the classical approach to text processing, pre-processing methods, approaches to morphological, syntactic and semantic analysis. Modern models of text information processing based on neural networks will also be considered in the course.

Course objectives and goals

The purpose of this course is to provide participants with in-depth knowledge and skills in the field of processing textual information. Participants will gain an in-depth understanding of the basic methods of analyzing, processing, and extracting information from texts, as well as learn to use modern tools and

software tools to automate these processes. The course is aimed at preparing participants for the effective use of text information processing in various fields, including natural language, social media analytics, automated document processing, etc. Upon completion of the course, participants will have the necessary skills to apply modern word processing techniques to practical tasks and research.

Format of classes

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

Competencies

GC 1. Ability to learn and master modern knowledge.
GC 2. Ability to apply knowledge in practical situations.
GC 6. Capability of abstract thinking, analysis and synthesis.
GC 7. Ability to search, process and analyse information from various sources.
GC 8. Knowledge and understanding of the subject area and understanding of professional activities.
SC 3. Ability to choose and apply mathematical methods for solving applied problems, modelling, analysis, design, management, forecasting, decision-making.
SC 10. Ability to conduct mathematical and computer modelling, data analysis and processing, computational experiment, solving formalized problems using specialized software.
SC 20. Ability to develop and operate software tools for intelligent analysis of measurement and observation data, texts, signals and images.
SC 22. Ability to use information technologies for statistical and intellectual data analysis, forecasting, decision-making, information retrieval and knowledge extraction.

Learning outcomes

LO 3. Formalize tasks formulated in the language of a particular subject fields; formulate their mathematical formulation and choose rational method of solution; solve the resulting problems with analytical and numerical methods, evaluate the accuracy and reliability of the results obtained.
LO 12. Solve individual engineering problems and/or tasks that arise in at least one subject area: sociology, economy, ecology, and medicine.
LO 22. To know and understand the methods of solving mathematical problems of intellectual information retrieval and knowledge extraction.
LO 24. Be able to apply existing and develop new algorithms and software tools for processing measurement and observation data, texts, signals and images.

Student workload

The total volume of the course is 120 hours (4 ECTS credits): lectures – 20 hours, laboratory classes – 20 hours, self-study – 80 hours.

Course prerequisites

To successfully pass the course, you must have knowledge and practical skills in the following disciplines: "Algorithmization and programming", "Computer discrete mathematics", "Probability theory", "Mathematical statistics", "Neural network technologies", "Methods and tools of machine learning" .

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

Lectures and practicals are conducted online using modern software. The Colab software environment is used for practical (laboratory) work, the focus is on the practical application of various models and methods or frameworks for processing text information.

Program of the course

Topics of the lectures

Topic 1. Introduction to natural language processing (NLP)
Basic concepts. Fields of application of NLP: existing tasks, applications and real life examples.
Topic 2. Preprocessing of textual information

Basic terminology and algorithms that help prepare data for linguistic analysis: tokenization, stop word removal, stemming and lemmatization, morphological analysis.

Topic 3. Modeling of language, language processes

Statistical language models, n-grams and neural language models (eg Word2Vec, GloVe).

Topic 4. Information search and text classification

Basic understanding of vector space models and their implementation. What is the term "frequency", document inverse frequency (TF-IDF) and what TF-IDF is used for and how to calculate it. The concepts of document similarity and text classification, text classification algorithms (for example, Naive Bayes, Support Vector Machines) and examples of their use.

Topic 5. Analysis of moods and opinions

Document-level sentiment analysis, facet-based sentiment analysis, and sentiment analysis based on social media data.

Topic 6. Recognition of named objects and relationships between entities

Approaches to recognizing named entities, linking entities, and using knowledge bases (eg, DBpedia, Wikidata) to extract entities.

Topic 7. Thematic modeling

Discovering hidden topics in a collection of documents. Latent Dirichlet Allocation (LDA). Non-negative matrix factorization (NMF).

Topic 8. Text referencing

Methods of compressing texts into concise and coherent summaries. Extractive and abstract abstracting. Metrics for evaluating the quality of generated abstracts, such as ROUGE, BLEU and METEOR.

Topic 9. Machine translation

Models of automatic text translation from one language to another. Neural machine translation, deep learning architectures.

Topic 10. Question-answer systems, chat bots

Architectures of automatic question answering systems. Deep learning models such as BERT and informer-based architectures.

Topics of the workshops

Workshops are not provided for in the curriculum.

Topics of the laboratory classes

Topic 1. Initial analysis of textual information

Methods of preprocessing text information, using regular expressions for processing text information.

Topic 2. Statistical methods of processing textual information

Using NLTK library, Pymorphy, WordNet, n-grams. Methods of lemmatization, stemming and morphological analysis (part-of-speech tagging).

Topic 3. Search for linguistic properties in the text

Linguistic feature extraction from texts, such as POS tags, dependency parsing, and lemmatization. Using packages such as Spacy, NLTK and Stanza. Working with phrases and phrasal verbs.

Topic 4. Vector representation of words

Using and building vector word representations, including Fastext, GloVe and Word2Vec. Solving various problems using vector representations of words.

Topic 5. Extraction of named entities

Named Entity Recognition (NER) tasks, in particular, CV entity recognition and location retrieval using various approaches such as POS-based, Spacy packages, Stanza, and dependency searches.

Topic 6. Classification of texts

Text classification using classic machine learning methods. Implementation of the sentiment analysis task.

Topic 7. Topic modeling

Implementation of classic topic modeling algorithms (LDA) on your data.

Topic 8. Building a machine translation system.

Implementation of a machine translation system for the Ukrainian language based on classical "sequence to sequence" methods and using neural network approaches.

Topic 9. Building your chatbot

Using a pre-trained model based on transformers in the telegram application. Or retraining such a model on your data and using it as a chat bot in Telegram.

Self-study

Self-study involves an individual task on a topic chosen by the student. Students are also recommended additional resources (videos and articles) for independent study.

Non-formal education

Within the framework of non-formal education according to the relevant Regulation (z0328-22), the educational component or its separate topics can be taken into account in case of independent completion of professional courses/training, obtaining civic education, online education, professional internship, etc.
In particular, individual topics of this component may be taken into account upon successful completion of the following courses:
- Topic 4. Information search and text classification
https://www.coursera.org/projects/analyze-text-data-yellowbrick
- Topic 5. Analysis of moods and opinions
https://www.coursera.org/learn/text-mining-analytics
- Topic 10. Question-answer systems, chat bots
https://www.coursera.org/learn/transformer-models-and-bert-model

Course materials and recommended reading

Books:
"Speech and Language Processing" by Daniel Jurafsky and James H. Martin.
"Natural Language Processing in Action" by Lane, Howard, and Hapke.
"Foundations of Statistical Natural Language Processing" by Christopher D. Manning and Hinrich Schütze.
"Deep Learning for NLP and Speech Recognition" by Palash Goyal, Sumit Pandey, Karan Jain.

Online courses:
Coursera: Natural Language Processing Specialization
Udemy: Natural Language Processing with Deep Learning in Python

Scientific articles:
"Attention is All You Need" by Vaswani et al. (Transformer model).
"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" by Devlin et al.
"Word2Vec" by Mikolov et al.
"GloVe: Global Vectors for Word Representation" by Pennington et al.

Blogs and educational materials:
The Annotated Transformer
Jay Alammar's Visualizing Neural Machine Translation
Chris Olah's Understanding LSTM Networks

Scientific journals and conferences:
Association for Computational Linguistics (ACL): https://www.aclweb.org/
Conference on Empirical Methods in Natural Language Processing (EMNLP):
https://www.emnlp2022.org/

Additional resources:
Hugging Face Transformers - A library for working with state-of-the-art natural language processing.
AllenNLP - An open-source NLP research library, built on PyTorch.

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 29.08.2024		Head of the Department Olena AKHIIEZER
	Date, signature 29.08.2024		Guarantor of the Educational Program Olena AKHIIEZER