COURSE DESCRIPTION CARD - SYLLABUS

Course name
MACHINE LEARNING METHODS IN NATURAL LANGUAGE PROCESSING [S5ITIT>MUMWPJN]

Course

Proposed by Discipline
–

Year/Semester
3/5

Level of study
Doctoral School

Course offered in
English

Form of study
full-time

Requirements
elective

Number of hours

Lecture
4

Laboratory classes
0

Other
0

Tutorials
0

Projects/seminars
0

Number of credit points

1,00

Coordinators

dr hab. inż. Mikołaj Morzy prof. PP
mikolaj.morzy@put.poznan.pl

Lecturers

dr hab. inż. Mikołaj Morzy prof. PP
mikolaj.morzy@put.poznan.pl

Prerequisites

Algorithmic literacy: Ability to understand simple pseudo-code containing variables, loops, and conditional logic (if-then-else). No prior programming experience is required. Statistical foundations: Familiarity with core statistical concepts, including probability, mean, and standard deviation. Interdisciplinary application: An aptitude for applying abstract patterns to research problems and thinking creatively about how to use textual data.

Course objective

This course is a comprehensive introduction to modern Natural Language Processing (NLP), focusing on the machine learning techniques that transform unstructured text into valuable insights. We will cover the essential pipeline of text analysis, from foundational principles to the state-of-the-art models shaping the field today. We will begin with fundamental processing techniques like tokenization, lemmatization, and classic representation methods such as TF-IDF. The course then progresses to the neural network approaches that define modern NLP, including word embeddings and the revolutionary Transformer architecture. A major focus will be on the current paradigm of Large Language Models (LLMs) and their role as foundational models, covering practical techniques like prompt engineering and fine-tuning. All concepts will be grounded in practical applications such as sentiment analysis, named entity recognition (NER), and machine translation.

Course-related learning outcomes

Knowledge
A PhD student who graduated from doctoral school knows and understands:
1) current achievements in the combined fields of machine learning and natural language processing, they understand basic principles of algorithms used to extract useful knowledge from unstructured text, [P8S_WG/SzD_W01]
2) the current developmental trends in machine learning and natural language processing, and can identify research questions in their scientific domains that can be addressed using machine learning and natural language processing. [P8S_WG/SzD_W02]

Skills
A PhD student who graduated from doctoral school can:
1) has the knowledge of machine learning and natural language processing to collect new data and new insights in their respective scientific disciplines, [P8S_UW/SzD_U01]
2) design new distributed representations of data in their scientific disciplines using the paradigm of encoder-decoder neural network architecture. [P8S_UW/SzD_U03]

Social competences
A PhD student who graduated from doctoral school is ready to:
1) acknowledge the importance of natural language processing methods by designing a research question involving their own discipline that can be addressed using machine learning and natural language processing. [P8S_KK/SzD_K03].

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Students are required to submit a detailed Project Proposal for an NLP-based Scientific Experiment relevant to their own academic or scientific discipline. The proposal must outline a novel and methodologically sound application of Natural Language Processing (NLP) techniques to investigate a specific scientific question.

The resulting document must function as a complete plan for the experiment, covering the following mandatory sections:
- Formulation of a Testable Hypothesis: A clear, concise, and scientifically relevant hypothesis that the proposed NLP experiment is designed to test.
- Corpus Availability and Description: A description of the available corpora (textual datasets) needed for the experiment, including an estimate of their size, quality, and accessibility (public, proprietary, or requiring acquisition).
- Data Acquisition and Preparation Plan: A detailed plan outlining what textual data would need to be extracted, gathered, or created to execute the experiment, and the necessary steps for data cleaning and pre-processing.
- Selection and Justification of NLP Methods: A description of the specific NLP methods (e.g., word embeddings, topic modeling, semantic parsing, sentiment analysis) selected for the analysis, with a robust justification for why they are the most appropriate tools for testing the hypothesis.
- Contribution to Discipline: A persuasive explanation of how the anticipated results from the proposed NLP experiment would contribute to or advance research in the student's specific scientific field.

The project proposals will be assessed based on the following criteria:

- Scientific clarity of hypothesis (25%): The clarity, precision, and scientific relevance of the formulated hypothesis to the student's discipline.
- Corpus and data feasibility (25%): The quality, size, and availability of the proposed text corpora, and the appropriateness and feasibility of the plan for data extraction and preparation.
- Methodological appropriateness (25%): The suitability and originality of the chosen NLP methods for defining semantic relationships, extracting relevant features, or effectively testing the formulated hypothesis.
- Disciplinary impact (25%): The depth and significance of the explained contribution the proposed NLP experiment will make to the student's specific area of research.

When verifying learning outcomes through exams, tests, etc., the following assignment of grades to percentage result ranges will be used:

- Below 50%: F (Fail)
> 50.0% to 60.0%: E (Sufficient/Pass)
> 60.0% to 70.0%: D (Satisfactory)
> 70.0% to 80.0%: C (Good)
> 80.0% to 90.0%: B (Very Good)
> 90.0% to 100%: A (Excellent)

Programme content

- Tokenization, Lemmatization, & Stemming: Breaking down text into its basic components (words or sub-words) and simplifying them to their root forms for analysis.
- TF-IDF (Term Frequency-Inverse Document Frequency): A classic statistical method used to measure a word's importance to a document within a larger collection (corpus).
- Word Embeddings: Representing words as dense numerical vectors that capture their semantic relationships, meaning, and context.
- The Transformer Architecture: A powerful neural network design that uses a "self-attention" mechanism to weigh the influence of different words in a sequence.
- Large Language Models (LLMs): Massive Transformer-based models trained on vast amounts of text data to understand, generate, and reason about human language.
- Prompt Engineering: The art and science of crafting effective inputs (prompts) to guide an LLM toward generating a specific and desired output.
- Fine-Tuning: The process of adapting a large, pre-trained model by continuing its training on a smaller, specialized dataset for a particular task.
- Sentiment Analysis: Automatically classifying the emotional tone or opinion expressed in a piece of text (e.g., positive, negative, neutral).
- Named Entity Recognition (NER): Identifying and categorizing key entities in text, such as names of people, organizations, locations, and dates.
- Machine Translation: The task of automatically translating text from a source language to a target language using sequence-to-sequence models.

Course topics

This lecture provides a comprehensive journey through the evolution of NLP methods in machine learning, beginning with fundamental text processing techniques like Tokenization, Lemmatization, and Stemming to normalize raw language. We will then examine classical statistical approaches, such as TF-IDF (Term Frequency-Inverse Document Frequency), which quantifies word importance before transitioning to modern vector-based representations, where Word Embeddings encode complex semantic meaning into dense numerical vectors. The core of modern NLP is the Transformer Architecture, which we will dissect to understand its revolutionary "self-attention" mechanism, paving the way for the development of massive, versatile models known as Large Language Models (LLMs). The final part of the lecture focuses on practical application, covering specialized tasks like Sentiment Analysis and Named Entity Recognition (NER), exploring advanced techniques such as Fine-Tuning to adapt LLMs for specific domains, and introducing the critical new skill of Prompt Engineering to effectively interact with these powerful systems, with a brief overview of sequence-to-sequence applications like Machine Translation.

Teaching methods

Lecture: multimedia presentation including illustrations and examples. LIve coding sessions.

Bibliography

Basic

1. Jurafsky, Daniel, and James H. Martin. Speech and language processing 3rd edition draft. 2019.
2. Goldberg, Yoav. Neural network methods in natural language processing. Morgan & Claypool Publishers, 2017.
3. Eisenstein, Jacob. Introduction to natural language processing. MIT press, 2019.

Additional

4. Tunstall, Lewis, Leandro Von Werra, and Thomas Wolf. Natural language processing with transformers.

O'Reilly Media, Inc, 2022.
5. Manning, Christopher, and Hinrich Schutze. Foundations of statistical natural language processing. MIT press, 1999.
6. Bender, Emily M. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. Morgan & Claypool Publishers, 2013.

Breakdown of average student's workload

	Hours	ECTS
Total workload	25	1,00
Classes requiring direct contact with the teacher	4	0,00
Doctoral student's own work (literature studies, preparation for laboratory classes/tutorials, preparation for tests/exam, project preparation)	21	1,00