Course: Natural language processing

2023/24, Spring semester

Lecturer: Prof. Dr. Marko Robnik-Šikonja
Assistants: Assist. Prof. Dr. Slavko Žitnik, Assist. Aleš Žagar, Assist. Boshko Koloski

Course objectives: Learn about the theory and main practical approaches in natural language processing and understanding. Use modern large language models and statistical techniques for language processing.

Student's obligations:

- five web quizzes
- assignments
- written exam

Grading

The practical work encompasses work with natural language processing tools and large language models. It is graded through assignments, which have to be finished on time. The assignments are done in groups of three students. The topics of the assignments are set at the start of the semester. The results of the assignments shall be described in a paper and publicly presented in front of the class.
The exam is in the form of a written test. The preconditions for the written exam are successfully passed quizzes and assignments. During the written exam, students are allowed to use one A4 sheet of paper. The prerequisite for the (optional) oral exam is to pass the written exam. In case of any doubts about the score of assignments, quizzes, or written exams, the oral exam is obligatory. The final grade is the sum of assignment scores and the written exam. The contribution of each work to the final grade and the conditions to pass the exam are listed below:

Work	% of total	subject to
Five quizzes	0%	≥ 50% altogether
Three assignments	50%	≥ 25%
Written exam	50%	≥ 25%

The grades are valid in the current year. The students not passing the exam in the current year have to redo the quizzes and assignments next year.

Syllabus

The syllabus is based on a selection of modern natural learning techniques and their practical use. The lectures introduce the main tasks and techniques and explain their operation and theoretical background. The knowledge gained during practical sessions, and seminars is applied to practical tasks using open-source tools. Students investigate and solve assignments based on real-world research and industrial problems, mainly dealing with English and Slovene.

1. Introduction to natural language processing: motivation, language understanding, ambiguity, traditional, statistical, and neural approaches.
2. Text preprocessing and normalization: regular expressions for search and replacement, grammars for syntax analysis, string similarity, Levenhstein distance, advanced normalization techniques, and lemmatization.
3. Language resources: corpora, dictionaries, thesauri, networks and semantic databases, WordNet.
4. Text similarity: measures, clustering approaches, cosine distance, language networks, and graphs.
5. Text representation: sparse and dense embeddings; language models; word, sentence, and document embeddings.
6. Deep neural networks for text: recurrent neural networks, CNNs for text, transformers.
7. Neural embeddings: word2vec, fastText, ELMo, BERT, cross-lingual embeddings.
8. Large pretrained language models: BERT, GPT, and T5 families, multimodal models.
9. Large generative language models, prompt engineering, and retrieval augmented generation.
10. Shallow computational and lexical semantics: part-of-speech tagging, dependency parsing, named entity recognition, semantic role labeling, FrameNet.
11. Word senses and disambiguation.
12. Affective computing: sentiment, emotions, etc.
13. Text summarization: text representations, extractive methods, query-based methods, abstractive summarization, text simplification, evaluation.
14. Question answering and reading comprehension: methods and evaluation.
15. Semantic tasks: natural language inference, commonsense reasoning, paraphrasing.
16. Neural machine translation and its evaluation.
17. Semantic representations: knowledge graphs.

Literature (all freely available)**:**

1. Jurafsky, David and Martin, James H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 3rd edition draft, 2024.*
   This is the primary course literature; available on authors' webpages
2. Simon J.D. Prince: *Understanding Deep Learning. MIT Press, 2023*
   https://udlbook.github.io/udlbook/ This recent book intuitively explains many important concepts and techniques in deep learning.
3. Bird, Steven, Ewan Klein, and Edward Loper*. Natural language processing with Python.* O'Reilly Media, Inc., 2009.
   The book was updated in 2019, based on NLTK library for Python 3