Course: Natural language processing

2021/22, Spring semester

Lecturer: Prof. Dr. Marko Robnik-Šikonja
Assistant: Assist. Prof. Dr. Slavko Žitnik

Course objectives: learn about the theory and main practical approaches in natural language processing and understanding. Use modern statistical approaches and techniques for language processing, develop machine learning models for text mining.

Student's obligations:

- five web quizzes
- assignments
- written exam or oral exam in case of epidemics

Grading

The practical work encompasses work with natural language processing tools and is graded through assignments, which have to be finished on time. The assignments are done in groups of up to three students. The topics of the assignments is set at the start of the semester. The assignments' results 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 precondition for the (optional) oral exam is to pass the written exam. In case of any doubts about the score of assignments, quizzes, or written exam, the oral exam is obligatory. The final grade is the sum of assignment scores and written exam. The contribution of each obligation to the final grade and conditions to pass the exam are listed below:

Obligation	**% of total**	**subject to**
Five quizzes	0%	≥ 50% alltogether
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.

In case of epidemics, all the exams will be oral.

Syllabus

The syllabus is based on a selection of modern statistical natural learning techniques and their practical use. The lectures introduce the main tasks and techniques, explain their operation and theoretical background. During practical sessions and seminars, the gained knowledge is applied to practical tasks using open source tools. Students investigate and solve assignments based on real-world research and commercial problems, mostly 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, 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. Shallow computational and lexical semantics: part-of-speech tagging, dependency parsing, named entity recognition, semantic role labelling, FrameNet.
9. Word senses and disambiguation.
10. Affective computing: sentiment, emotions.
11. Text summarization: text representations, extractive methods, query-based methods, abstractive summarization, evaluation.
12. Question answering and reading comprehension: methods and evaluation..
13. Machine translation: statistical and neural machine translation, evaluation.
14. Semantic representations: knowledge graphs for commonsense reasoning.

Literature:

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, 2022.*
   This is the primary course literature; available on authors' webpages
2. Jacob Eisenstein. *Natural Language Processing, MIT press, 2019*
3. Bird, Steven, Ewan Klein, and Edward Loper. *Natural language processing with Python. O'Reilly Media, Inc., 2009.*
   Freely available book, updated in 2019, based on NLTK library for Python 3
4. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning. MIT press, 2016*