GHENT UNIVERSITY

Course Specifications

Valid in the academic year 2023-2024

Natural Language Processing (E061341)

Course size *(nominal values; actual values may depend on programme)*

Credits 6.0 **Study time 180 h**

Course offerings and teaching methods in academic year 2023-2024

A (semester 2)	English	Gent	practical	15.0h
			lecture	35.0h
			group work	10.0h

Lecturers in academic year 2023-2024

Demeester, Thomas	TWO5	lecturer-in-charge
Develder, Chris	TWO5	co-lecturer

Offered in the following programmes in 2023-2024	crdts	offering
Bridging Programme Master of Science in Bioinformatics(main subject Engineering)	6	A
Master of Science in Bioinformatics(main subject Engineering)	6	A
Master of Science in Computer Science	6	A
Master of Science in Computer Science Engineering	6	A
Master of Science in Computer Science Engineering	6	A
Master of Science in Statistical Data Analysis	6	A
Exchange Programme in Computer Science (master's level)	6	A

Teaching languages

English

Keywords

Natural language processing, machine learning, artificial neural networks, statistical methods

Position of the course

In various application domains, a substantial fraction of data comprises natural language text, e.g., web pages, news articles, magazines, blogs, tweets, Facebook posts, chat messages, etc. This purely textual information comprises useful information that would be a lot more valuable if it were interpretable for further automated computer processing (e.g., by conversion to structured information in databases). Unlocking the value of that information by developing techniques to interpret human language using computer algorithms is what Natural language processing (NLP) aims at.

NLP is a field that combines computer science, artificial intelligence and linguistics and is generally aimed at enabling computers to solve tasks that involve understanding and/or generating natural language. These tasks that have to deal with human language are omnipresent in our daily lives and range from basic search (in web search engines), to automatic question answering or machine translation.

In this course, we focus on the processing of language in written form (and thus will not deal with, e.g., speech processing). More specifically, this course serves as an introduction to (1) Classical NLP, also referred to as Statistical NLP (SNLP) since it heavily relies on statistics and machine learning, as well as (2) Neural NLP, i.e., methods based on neural networks. In SNLP, computers are not directly programmed to process language, but typically learn how to process (or generate) language based on the statistics of an (often vast) corpus of natural language. More recent models based on neural networks are typically trained end-to-end on

the textual data, typically avoiding specific feature engineering and/or explicit definition of subtasks in Classical NLP (e.g., PoS tagging, dependency parsing) as pipelined steps towards the final end goal.
Given the extremely rapid evolution of the field of neural NLP, both in terms of research publications and open source models and software, we will encourage students to look into recent literature and share some of their insights with the group. In particular, after an in-depth introduction of the core building blocks of transformers and seminal model architectures, the students will read a recent transformer-based paper and summarize it in a short pitch to their peers.

The purpose of the course is to empower students with both theoretical and practical knowledge of the most important concepts and techniques in NLP, so that they know (1) the major NLP tasks (including text classification, sequence tagging, syntactic parsing, language modelling, machine translation), (2) the essential methods and frameworks (including both statistical NLP and neural network based methods), as well as (3) the practical implementations thereof (including modern software tools, e.g., Pytorch).

Contents

- Introduction: what is NLP; basic NLP tasks, sample applications;
- Classical NLP
  - Techniques for text classification (including Naïve Bayes, logistic regression)
  - N-gram language models
  - Sequence tagging (including Part-of-speech tagging)
  - Constituency and dependency parsing
- Neural NLP
  - Recurrent sequence models
  - Neural language models
  - Word and sentence representations
  - Sequence to sequence models (including machine translation)
  - Transformers: building blocks, seminal models, and recent advances
  - Recent paradigms (including pre-training and fine-tuning, prompt engineering)

Initial competences

- Basic programming in Python
- Basic machine learning concepts:
  - Supervised learning (train/dev/test approach)
  - Neural network basics (multilayer perceptron, gradient-based training)

Final competences

1 Know the basic NLP tasks and the methods to address them (e.g., text preprocessing, language modeling, parsing, sequence tagging, text classification, sequence-to-sequence tasks)
2 Explain, apply and evaluate methods for NLP-based applications such as named entity recognition, machine translation, sentence classification or information extraction.
3 Have insights in models based on learned representations (ranging from static word embeddings to pre-trained transformer models) and compatible neural network building blocks, to leverage these learned representations to solve custom problems in NLP.
4 Explain and understand various types (e.g., intrinsic and extrinsic) and measures of evaluation.
5 Implement and evaluate an NLP application using Python.
6 Hands-on experience in following up recent literature and deploying newly released models.

Conditions for credit contract

Access to this course unit via a credit contract is determined after successful competences assessment

Conditions for exam contract

This course unit cannot be taken via an exam contract

Teaching methods

Group work, Lecture, Practical, Independent work

Extra information on the teaching methods

- The course will be offered through weekly theory lectures on topics in Classical and Neural NLP. By the end of the term, the students will be divided into small groups, to present the key ideas of a recent paper.
- Practical sessions will be organized as guided self-study sessions (i.e., remote support via MS Teams).

Learning materials and price

Lecture slides will be published on the electronic learning platform. The material will be complemented by chapters from textbooks (such as Jurafsky), academic papers and other references.

References

- Speech and Language Processing, D. Jurafsky and J.H. Martin
- Neural Network Models in Natural Language Processing, Y. Goldberg
- Natural Language Processing - A Machine Learning Perspective, Y. Zhang, Z. Teng
- Deep Learning, I. Goodfellow, Y. Bengio and A. Courville
- Foundations of Statistical Natural Language Processing, C.D. Manning and H. Schütze
- Articles from specialised recent literature; available on request if necessary.

Course content-related study coaching

The teachers and their assistant(s) will be available during and in between lectures for additional clarification (upon making an appointment).

Assessment moments

end-of-term and continuous assessment

Examination methods in case of periodic assessment during the first examination period

Written assessment

Examination methods in case of periodic assessment during the second examination period

Written assessment

Examination methods in case of permanent assessment

Assignment

Possibilities of retake in case of permanent assessment

examination during the second examination period is not possible

Extra information on the examination methods

- Periodic evaluation: written exam, aimed at evaluating the understanding of and ability to apply the concepts presented in the course.
- Permanent evaluation:
  - Practicals: aimed at applying the theory in practice. Students will implement solutions to NLP tasks, as well as evaluate and interpret the results. Students may have to read through scientific papers in preparation for the assignments to be performed.
  - Paper presentation: will be assessed in terms of scientific content and clarity of presentation; the score for the presentation will be weighted as a single practical session.
  - There is no second chance for permanent evaluation; the score for practicals will be transferred in case of second chance (for the written exam only then). Note that given the score calculation as detailed below, a total score of more than 8/20 needs to be obtained for the practicals to pass the course.

Calculation of the examination mark

- Written exam: 75%
- Permanent evaluation (practicals and paper presentation): 25%
- Additional requirement to pass: a score of at least 9/20 for both parts ('written exam' and 'permanent evaluation'). If the student scores 8/20 or lower for one of the components, any total score of ten or more will be reduced to the highest failing mark (9/20).

Facilities for Working Students

Timing of the practicals can be changed for working students.