236299: Introduction to Natural Language Processing

Syllabus:
Introduction to natural language processing. The course presents methods for representing human language at different levels, moving from simple bag-of-words representations, through sequential models, to syntactic and semantic structures. At each level, the course develops several machine learning models and algorithms for solving various natural language processing tasks, such as text classification, sequence labeling, syntactic parsing, and semantic parsing. The course discusses models such as Naïve Bayes, logistic regression, hidden Markov models, various deep neural networks, and appropriate learning algorithms for estimating their parameters. The course emphasizes implementation of classical and modern natural language processing models.

Staff:
Yonatan Belinkov, instructor (belinkov@technion.ac.il)

Learning outcomes:
By the end of this course, the student will be able to:

1. Model the standard levels of linguistic structures using formal grammars or statistical and computational models.
2. Identify and carry out proper experimental methodology for training and evaluating natural language processing systems.
3. Manipulate probabilities and estimate parameters of structured models using supervised training methods.
4. Implement simple models of language, and employ and adapt them in service of solving natural language processing problems.

Pre-requisites:
The official pre-requisite is 234247: Algorithms 1. However, we will make extensive use of machine learning algorithms, so a course in machine learning or deep learning – such as 236756: Introduction to machine learning or 236781: Deep Learning on Computational Accelerators – is highly recommended. We will also mention formal language concepts so it is recommended to have taken 234129: Introduction to set theory and automata for CS. Finally, this is a hands-on course with problem sets and lab exercises, so familiarity with Python and associated tools (Jupyter/Colab notebooks) will be assumed.

Coursework:

This course is focused on in-class labs and out-of-class problem sets. The course is divided in four segments, where each segment starts with an overview lecture and continues with a series of labs. We will use both the lecture time and the recitation time for the labs.

Before each lab, students prepare by reading textbooks and other course materials. *Preparation at home is crucial* for understanding the material. In the lab, students work through a series of small exercises in (randomly assigned) pairs. These exercises are intended to reinforce the concepts introduced in the readings, as well as provide hands-on experience with them. The exercises include both programming problems and pencil-and-paper exercises. Students submit the lab to a grading server for automatic feedback. The course staff provides guidance during the lab.

Each segment includes a problem set, to be completed in pairs (can be chosen by students). The problem sets require substantial implementation. All programming will be done in Python (and PyTorch) in Jupyter notebooks, either locally or via Google Colab.

There will be a mid-term quiz and a final exam.

Language of instruction:
The course materials are all in English. The course staff will use English if there is a request from non-Hebrew speaking students, in particular in the introductory lectures in the beginning of each segment. Students can feel free to ask questions and communicate in Hebrew or any language they are comfortable with. The submission of labs and problem sets, which are in the form of Jupyter notebooks, will also be in English (and, obviously, Python).

Collaboration policy:
Students are encouraged to discuss all aspects of the course (reading material, labs, and problem sets), as such discussions can be a useful learning experience. However, except where specifically stated otherwise, students should complete all assignments *individually* or in the *assigned pairs*. High-level discussions are allowed, and encouraged, but working together directly on implementing your solutions is disallowed. Using other people's code, whether from other course participants or elsewhere, is inappropriate. In labs that are completed in pairs, the pair members will work closely together. High-level discussions may take place across pairs, but sharing code outside the pair is disallowed.

Textbooks:
We will mainly use the following two textbooks:

- Dan Jurafsky and James H. Martin, Speech and Language Processing (3rd edition).
- Jacob Eisenstein, Natural Language Processing.

Grading:
The course grade will be comprised **approximately** as follows:

- Problem sets 25%
- Labs 25%
- Mid-term quiz 25%
- Final exam 25%

Schedule:
The course meets weekly on Mondays at 10:30-13:30 at TBD. We will use the three hour slot for the interactive labs, with a short break. A more detailed schedule will be made available later.
We will **meet in person** with no remote participation, subject to change per Technion regulations.

Expectations from students:
This course is somewhat different from what you might be used to. There will be *quite a lot of reading to do at home*, and class time will be devoted primarily to hands-on exercises (programming and pencil-and-paper). We will assume that students have done the reading at home and it will be essential for being able to complete the exercises successfully.
In addition, much of the work will be done in small groups, where students discuss and converge on a solution. This means that your fellow class mates depend on you and you depend on them. Therefore, you are expected to *actively participate* in every class, having done the reading at home, as well as work diligently on homework problem sets. Failure to collaborate will result in deduction of points and may lead to removal from the class. Attendance is mandatory, except for approved absences according to Technion policies. We will permit one unapproved absence. Each additional absence will deduct 5 points from the final grade.

Support:
We encourage asking questions (and posting answers!) via Piazza: https://piazza.com/technion.ac.il/winter2023/236299. We will also post announcements through Piazza so make sure to enroll yourself and follow regularly. Please use Piazza also for direct communication with the staff via private messages.

While Piazza is the preferred way to get support, the course staff will also hold weekly office hours:

- Yonatan Belinkov Sunday 13:00-14:00