MATH 60621A: Natural language processing, Winter 2024

Mondays 12:00 PM – 3:00 PM, Building CSC, Room Fonds Cogeco

Instructors (Until March 10th)
Dr. Eva Portelance (she/her)
eva.portelance@hec.ca
Office - CSC 4.834 Office hours: Thursdays 1:00 PM – 2:00 PM

(After March 10th)
Dr. Marcus Mosbach (he/him)
marius.mosbach@mila.quebec
Office - CSC 4.834 Office hours: TBD

1 Overview

This course will cover the fundamental tools and models used for analysing and generating natural language data in a computational setting. We will learn about the core principles behind contemporary natural language processing methods and about the structure of natural language data and how it may be used for downstream tasks such as topic modeling, document search and classification, text generation, and text summarisation.

2 Topics covered include:

- The nature of text data
- Preprocessing: tokenization and lemmatization
- Bag-of-words and naive classification
- N-gram language models (markov models)
- Hidden markov models and part-of-speech tagging
- Distributed representations, vector semantics, and topic models
- Recurrent neural language models, LSTMs and language generation
- Transformers and masked language modeling
- Encoder models and semantic search
- Encoder-decoder models and text summarization

By the end of this course, students will:

- Understand the nature of text data and how statistical algorithm may be applied to it;
- Learn how both statistical and neural network based language models are structured and how to implement them;
- Learn to think critically about current NLP methodologies and to identify the limitations and reasonable applications of each algorithm covered;
- Develop practical programming skills and concepts that support the above goals.

3 Prerequisites

The course expects previous experience with the Python programming language, though no previous experience with machine learning specific libraries is required.

4 Course structure

Lectures will be delivered in person, and will not be recorded. Slides for lectures will always be provided before class and students are expected to take their own notes. The first half of each course will be a presentation of new course materials. This presentation will then be followed by a 15 minute break. After break, students will work together in class on each week's practical exercises.

There will be two types of assessments: (1) practical assignments consisting solely of applied coding questions on which students will work either alone or in teams of 2-3 people (each assessed individually); (2) exams covering written theoretical and ethical questions.

Students are encouraged to attend my office hours to ask questions in person, or to discuss course materials and related topics.

5 Group work policy

Students can work either alone or in teams of 2-3 people (in fact they are encouraged to) on assignments, if they so desire. If they choose to work with others, they must write down the names of the students with whom they collaborated on their assignment file in the indicated fields. They are still required to individually submit their assignments and if they fail to do so, they will receive a zero grade.

6 Late assignment policy

Students will be deducted 15% of their assignment grade per late day, no exceptions. You will be given assignments 3 weeks in advance before their due date — plan accordingly, there is no reason they should be late.

7 Email policy

Please allow 2 work days for all email responses. I will never send or respond to emails outside of working hours (8:00-17:00).

8 Evaluation

- **Assignments (50%)**: 2 assignments, each worth 25%
- **Midterm exam (25%)**
- **Final exam (25%)**

9 Resources

Recommended chapter readings will come from:

- Dan Jurafsky and James H. Martin. (2024). *Speech and Language Processing (3rd Edition draft)*. https://web.stanford.edu/ jurafsky/slp3/

10 Course schedule

Note that the exact dates below are subject to change, depending on how quickly we make progress through topics in the course.

Date	Content
January 6	Class introduction and the nature of text data

January 13	Regular expressions, tokenization and lemmatization *Recommended readings: Chapter 2, up to section 2.7*
January 20	Bag-of-words and naive classification *Recommended readings: Chapter 4*
January 27	N-gram language models, hidden markov models and part-of-speech tagging *Recommended readings: Chapter 3 and Appendix A* *Notes:* Assignment 1 released
February 3	Distributed representations, vector semantics and topic models *Recommended readings: Chapter 6*
February 10	Logistic regression and basic neural networks for classification *Recommended readings: Chapter 5 and Chapter 7*
February 17	Recurrent neural language models and LSTMs *Recommended readings: Chapter 8* **Assignment 1 Due**
February 24	[Break]
March 3	**Midterm Exam**
March 10	Transformers and masked language modeling *Recommended readings: Chapter 9 and Chapter 11* *Notes:* Assignment 2 released
March 17	Text encoding and neural semantic search *Recommended readings: Chapter 14*
March 24	Encoder-decoders and text summarization or translation *Recommended readings: Chapter 13*
March 31	The current context: LLMs, Chatbots, and prompting *Recommended readings: Chapter 10 and Chapter 12* **Assignment 2 Due**
April 7	Guest lecture or catch up
TBD	**Final Exam**