Course program and reading list

Semester 2 Year 2024

School: Efi Arazi School of Computer Science M.Sc.

Large language Models and Information Theory

Lecturer:

Dr. Alon Kipnis alon.kipnis@runi.ac.il

Course No.:	Course Type :	Weekly Hours :	Credit:
3968	Seminar	3	3

Course Requirements :	Group Code :	Language:
Final Paper	242396801	Hebrew

Prerequisites

Prerequisite:

52 - Calculus I
53 - Calculus II
54 - Linear Algebra I
55 - Linear Algebra II
56 - Discrete Mathematics
59 - Data Structures
69 - Logic And Set Theory
417 - Introduction To Computer Science
3798 - Information Theory **OR** 3523 - Natural Language Processing

Course Description

Seminar in information theory and natural language processing (NLP).

Students in this seminar will be required to prepare and present a mini-research project

falling within a list of prescribed topics at the interaction of information theory and natural language processing.

The presentation is typically a combination of material and code.

Example topics:

- Data compression using large language models
- Authorship attribution
- Style transfer
- Intelligence via compression of information
- Limitations of large language models
- Information measures for achieving tasks in NLP

Course Goals

Exploring applications of data compression and information measures to achieve goals in natural language processing. For example:

-
  - Neural language-based text compressor
  - Neural language-based keyword extractor
  - Neural language-based summarizers

Attaining familiarity with up-to-date research and applications of language models.

Grading

Tentative policy (the final grading policy will be presented in the first class)

20% attendance (at least 80% attendance is required to obtain a full attendance grade)

40% presentation preparation

30% presentation delivery

10% post-presentation material preparation for archiving

Bonuses:

+15% for the publication of an article based on the project in a popular science venue (like Medium)

+30% for the publication of an article based on the project in an academic venue

(in last year's iteration of the class, one paper resulting from the class project was published at ICML)

Lecturer Office Hours

Monday 14:00-15:00

Reading List

- Daniel Jurafsky and James Martin. "Speech and Language Processing". Available online: https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf
- Chapters 1-5 of "Elements of Information Theory" by M. Thomas and T. Cover. 2006. (Second Edition)
- Patrick von Platen. 2020. "How to generate text". (https://huggingface.co/blog/how-to-generate)
- Bellard, F. (2021). "NNCP v2: Lossless Data Compression with Transformer".
- Aston Zhang, Alexander J. Smola, Zachary Lipton, Mu Li. (2023). "Dive into Deep learning" (https://d2l.ai/chapter_introduction/index.html)
- Andrej Karpathy. "Let's build GPT: from scratch, in code, spelled out." https://www.youtube.com/watch?v=kCc8FmEb1nY