HSLU Lucerne University of Applied Sciences and Arts

Introduction

NLP
Andreas Marfurt

Information Technology
20.02.2025

FH Zentralschweiz

About me

- MSc/BSc in computer science from ETH
- PhD from EPFL in NLP (summarization)
- Lecturer at HSLU since September 2022
- Teaching NLP and MLOps

Course TAs

- Angelina Parfenova (angelina.parfenova@hslu.ch)
- Christof Bless (christof.bless@hslu.ch)
- Maria Andueza Rodriguez (maria.anduezarodriguez@hslu.ch)
- Pascal Wullschleger (pascal.wullschleger@hslu.ch)

Course info

- 9 ECTS (= a lot of work)
- Language: English
- Lectures
  - Thursdays & Fridays, 09:05 – 11:25
  - Room 203
  - Streamed on Zoom (guest lectures TBD)
  - (Unedited) Zoom recordings uploaded to Ilias
- Project discussions
  - Fridays 12:50 – 15:10
  - Room 303 (with 3 exceptions, check myCampus)
  - Not streamed/recorded

Lectures

- Lecture structure
  - Theory
  - In-class exercises: Jupyter notebooks
    - Implementation for concepts from lecture
    - Collaborative: Discuss with your neighbor
    - Can be solved on your laptop
    - Can use GPUHub or Google colab
- Please ask questions if you're not sure you understood!
  - Maybe an example could help?
  - Terminology unclear?
  - Other students almost always have the same questions!

Project discussions

- Split the class
  - Not too many people for a discussion
- One topic per week
  - All relevant for your projects
- Discussion of
  - Challenges
  - Solution options
  - Trade-offs

Group 1 (12:50 - 13:55)	Group 2 (14:05 - 15:10)
Bozovic, Luka	Bürge, Jonas
Dollfus, André	Christen, Michael
Dubach, Fabian	Fender, Jakob
Frischknecht, Yannick Sascha	Furlan, Aaron
Gansner, Pascal	Gashi, Flora
Hermann, Nicola	Gonzalez, Diego
Hodel, David	Helfenstein, Nevin
Horlacher, Sofia	Kyburz, Luca
Jacober, Jonas	Le, Nicola
Kronenberg, Luke	Mariani, Pedro
Muri, Tristan	Niederer, Luca
Schurtenberger, David	Rajakone, Jamian
Sharma, Aditi	Schmid, Timon
Sherbetjian, Sevan	Soler, Lars
Sparapani, Leonardo	Walter, Tamino
Suter, Michelle	Weder, Nico
Trede, Maiko	Windlin, Teresa

Schedule

Week	Thu 1 - Lecture	Fri 1 - Lecture	Fri 2 - Project discussions
1	Admin, project intro, preliminaries	NLP pipeline, tokenization, libraries	Experiment tracking
2	Vector space model, topic modeling	Word embeddings	Preprocessing
3	RNN 1	RNN 2	Data loading, batching
4	LSTM, GRU	Advanced RNNs	Input/output format
5	Attention	Transformer 1	Training (checkpointing, early stopping)
6	Transformer 2	Transformer 3	Hyperparameter tuning
7	Transformer paper preparation	Transformer paper discussion	Evaluation (metrics, error analysis)
8	Intermediate presentations	Intermediate presentations	Intermediate presentations
9	Pretraining 1	Good Friday (free)	Good Friday (free)
10	Pretraining 2	LLMs	Debugging, testing
11	Text generation 1	Text generation 2	Interpretation & expectations
12	Text classification 1	Text classification 2	Jupyter notebooks
13	Guest lecture/research topics	Guest lecture/research topics	Presentations
14	SwissText conference (free)	Final presentations	Final presentations
15	Ascension (holiday)	Exam preparation (free)	Exam preparation (free)

Grade

- Testat: Hand in first project
  - NLP will exceptionally not run in HS25
- Project (50% of grade)
  - More on next slides
- Exam (50% of grade)
  - In exam session
  - Pen-and-paper
  - Closed book
  - All course material is relevant (including in-class exercises, projects and guest lectures)

Projects

- 2 single-person projects
  - Discussion among students is fine, use of AI tools allowed
  - Think for yourself, no docs/code sharing
- Hand in: Jupyter notebook, presentation slides
- Jupyter notebook
  - Documentation
  - Code (save notebook with cell outputs)
  - Experiment tracking
- Presentation: in group sessions
  - On-site
  - 5 minutes per student
  - 5 students per group (randomly shuffled)
  - Followed by group discussion and questions (10 minutes)

Project 1

- Topic: Word embeddings & RNNs
- Testat: Hand in project (decide before submission)
- Jupyter notebook: LLM feedback (no grade)
  - You write a rebuttal and list “things to improve” (included in project 2 grade)
- Presentation (25% of project grade)

Project 2

- Topic: Transformers
  - Randomly initialized
  - Pretrained
  - LLM
- Notebook graded by TAs & me (50% of project grade)
- Presentation (25% of project grade)

Feedback from previous years

- A lot of material in this course
- More work than other courses, but also learned a lot
- Need to stay on top of the lectures
  - Later content builds on previous weeks
  - Hard to catch up
  - You are responsible to make sure you understood!

- My view: students learn a lot and deliver great projects
  - I aim to prepare you for a successful Bachelor's thesis and future career

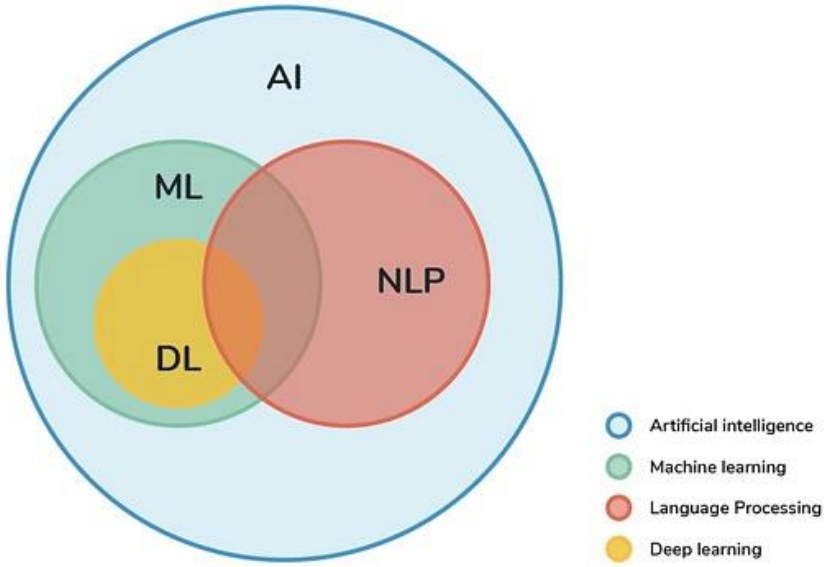
About the slides

Source

https://www.pinterest.ch/pin/374150681539780873/

What is NLP?

- NLP = Natural Language Processing
- Modality: Text

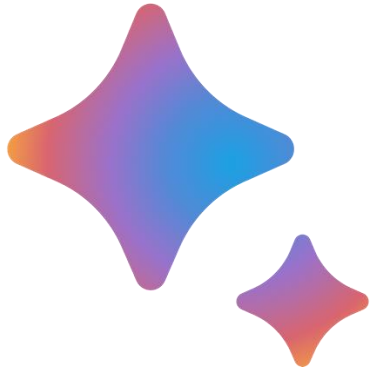


https://www.projectpro.io/article/machine-learning-vs-nlp/493

Why NLP?

- Automatically process large volumes of text
- Human knowledge is stored in text (encyclopedias, textbooks, novels, blogs, tweets, …)
- A lot of unstructured information on the web
- Important applications:
  - Web search
  - Sentiment analysis
  - Translation
  - Summarization
  - Chatbots

NLP is everywhere now…



Hugging Face

GitHub Copilot

Gemini

DeepL

deepseek

Images from https://images.google.com