PhD Course on Natural Language Processing and Large Language Models for Research Data Exploration and Analysis

Course Introduction / Description:

Generative AI is a type of artificial intelligence that can create and generate new content. As part of Generative AI, **Large Language Models** [LLMs] are deep learning-based transformer architectures (e.g., GPT-4/Generative Pre-trained Transformer-4), which are considered a significant breakthrough in the Natural Language Processing (NLP) and AI field and have shown a substantial potential to transform organizations and society in several ways. An example of an LLM is ChatGPT, which has recently gained widespread attention for its exceptional language generation skills and has demonstrated tremendous capabilities across various domains and tasks such as question-answering and passing examinations (such as Uniform Bar Exam, etc.), thereby even challenging our wisdom and cognition.

The primary purpose of this course is to provide knowledge and a deep understanding of various concepts, techniques, and methods that serve as a foundation for LLMs such as ChatGPT. Starting from the basic NLP concepts, this course will delve into deep learning architectures for NLP and Generative AI and then into applications and data analysis using LLMs. Subsequently, the course will focus on the opportunities, challenges, and risks associated with these Generative AI models like ChatGPT and their implications for organizations and society. The following is the outline for the course.

The course is mainly designed for PhD students who want to use NLP and text analysis in their research using LLMs. It also contains hands-on exercises on these topics using the Python programming language. The PhD students are expected to have some basic understanding of either Python or R programming languages and some familiarity with running Python scripts using Jupyter Notebooks. The following is the course outline.

1. The course starts with some fundamental concepts of machine learning (ML) and NLP. It then focuses on using these techniques for data analysis, using supervised and unsupervised approaches, such as text classification and topic modeling.
2. Second, it presents the high-level architectures of deep learning, generative models, and LLMs and elaborates on why LLMs like ChatGPT have achieved so many analytical capabilities.

3. Third, it will provide a detailed account of these models' capabilities, possible applications to various fields, and how they will impact society and organizations in the future.
4. Fourth, it will present how LLMs can be used for text analysis by examining some of the techniques, such as text summarization, text classification, and code generation.
5. Finally, it will discuss these models' diverse societal impacts and challenges, especially in terms of inequity, misuse, and legal and ethical considerations.

Course Learning Outcomes:

After completing this course, the participants should be able to:

1. Demonstrate the fundamental understanding of NLP and how they can be used for the analysis of text corpora.
2. Explain the fundamental principles of generative AI and LLMs and how they can be used for data analysis.
3. Compare various approaches to using Generative AI and LLMs, demonstrating their practical relevance through real-world applications and case studies.
4. Describe the key challenges and opportunities, including issues related to reliability, hallucination, and ethical considerations in using Generative AI and LLMs in various domains.

Pre-requisites:

Some basic understanding of either Python or R programming languages and ability to run Python scripts using Jupyter Notebooks.

Pedagogy:

Face to Face teaching

Session Plan:

Session	Time	Topic & Objective	Study Material
	Day-01: Tuesday, 11-03-2025		
00	10:00 – 10:15 [RRM + SR]	Course Introduction and Practicalities	
01	10:15-12:00 [SR]	Fundamentals of Machine Learning and Natural Language Processing (NLP) • Types of Machine Learning • Performance Measure • Basic Text Processing and Tokenization, • Word normalization and Parts-of-speech tagging	Slides, articles and other reading materials
Lunch Break			
02	13:00 – 14:45	Supervised approaches for NLP: text classification sentiment analysis, Naïve Bayes Classifier	Slides, articles and other reading materials

Session	Time	Topic & Objective	Study Material
	[RRM]		
03	15:00 - 17:00 [RRM]	Case Studies and Hands-on Session: • Analyzing text from discussion forums on Type-2 diabetes for domain-specific text classification • Sentiment analysis and text classification for movie reviews using NLTK	Jupyter notebooks and Python scripts
	Day-02: Wednesday, 12-03-2025		
04	09:00 - 11:30 [RRM]	Unsupervised and Deep Learning approaches for NLP: • Topic modeling • Word Vectors/Word Embeddings	Slides, articles and other reading materials
Lunch Break			
05	12:30 – 14:30 [RRM]	Case Studies and Hands-on Session: • Analyzing newspaper headlines using Topic modeling • Building word embeddings for your own text corpus	Jupyter notebooks and Python Scripts
06	14:45 - 17:00 [SR]	Introduction to Generative AI and Large Language Models (LLMs): transformers architecture, attention mechanism and generating text with transformers	Slides, articles and other reading materials
	Day-03: Thursday, 13-03-2025		
07	08:30 – 10:00 [RRM]	Configuring and fine-tuning LLMs for specific applications, e.g. text classification, text summarization.	Slides, articles and other reading materials
08	10:15 - 11:30 [RRM]	Case Studies and Hands-on Session: Hands-on: text summarization and text classification using LLMs using Google Cloud and Gemini LLM	Jupyter notebooks and Python Scripts
Lunch Break			
09	12:30 - 13:30 [SR]	LLMs use cases, challenges, opportunities, and ethical considerations	Slides, articles and other reading materials
10	13:30 – 14:30 [RRM + SR]	Wrap-up: Discussion about exam projects! Feedback and reflections on the course	

Evaluation Criteria:

Sr. No.	Component	Individual / Group	Weightage
1	Final project	Individual/group	100%
Total			100%

Profile of Instructors:

Raghava Mukkamala:

Raghava Mukkamala is an associate professor at the Department of Digitalization, Copenhagen Business School (CBS), Denmark. Raghava is also the programme director for the Master's Programme in Data Science at CBS and teaches several courses in Deep Learning and Natural Language Processing. His research primarily centered around Data Science, Blockchain Technologies, and Cybersecurity. His current research focuses on developing novel computational methods to analyze social media discourse, misinformation, and hate speech by combining formal/mathematical modeling techniques with advanced machine learning algorithms. As part of a pro-bono research collaboration with the United Nations High Commissioner for Refugees (UNHCR), he works on domain-adaptation and finetuning Large Language Models to identify hate speech and bias against refugees. Even though most of his research is mainly published in IEEE and ACM journals, he has also published several papers in FT-50/AJG-4*/ABDC-A* journals like the Journal of the Association for Information Systems (JAIS) and the Journal of Management Information Systems (JMIS). Raghava holds a Ph.D. in Theoretical Computer Science and an M.Sc. in Information Technology from IT University of Copenhagen, Denmark.

Link to homepage: https://www.cbs.dk/en/staff/rrmdigi

Sippo Rossi: