| | |
|---|---|
| ***Titolo del seminario*** | Large Language Models for Information Management |
| ***Settore Scientifico disciplinare di riferimento*** | Secs-S01, Inf-01 |
| ***Docenti proponenti*** | Prof. Claudio Conversano<br>Dott. Maurizio Romano<br>Dott. Marco Ortu<br>Prof. Diego Reforgiato Recupero (Dip. di Matematica e Informatica)<br>Prof. Roberto Tonelli (Dip. di Matematica e Informatica)<br>Dott. Mirko Marras (Dip. di Matematica e Informatica) |
| ***Docenti*** | Dott. Andrea Cadeddu, Linkalab<br>Dott. Luca Secchi, Linkalab<br>Dott. Vincenzo De Leo, Linkalab e Università di Cagliari |
| ***Semestre nel quale viene impartito*** | Secondo AA 2023/2024 |
| ***Crediti*** | 2 |
| ***Giorni, Orari, Aula*** | 6, 13 e 20 maggio dalle 14:30 alle 18:30, totale 12 ore<br>Laboratorio M, Piano Terra, Palazzo delle Scienze |
| ***Prerequisiti*** | Discreta conoscenza delle funzionalità di base del linguaggio di programmazione Python e una conoscenza minima dell'invocazione dei metodi di una libreria software e la ricerca di informazioni sul funzionamento e sui parametri dei metodi di una libreria software nella documentazione online della stessa. |
| ***Obiettivi formativi*** | Le tecniche di Retrieval Augmented Generation (RAG) forniscono un modo per ottimizzare la produzione di un LLM con informazioni mirate senza modificare il modello alla sua base; quelle informazioni mirate possono essere più aggiornate rispetto al LLM, ma anche più precise nel caso di organizzazioni e settori specifici. Ciò significa che il sistema di intelligenza artificiale generativa può fornire risposte più appropriate ai prompt, e basare tali risposte su dati estremamente attuali. La RAG rappresenta quindi un ambito di profondo interesse relativamente alla gestione efficace del problema delle allucinazioni dei LLM.<br>Il seminario si pone l'obiettivo di introdurre tutti i concetti principali necessari per comprendere i meccanismi di base dei LLM, le tecniche principali di in-context learning e di information retrieval per rappresentare e ricercare le informazioni su base semantica e, infine, le caratteristiche principali dei sistemi RAG per la ricerca di documenti; a completamento della parte teorica verrà quindi proposta ai partecipanti una esercitazione pratica in prima persona sullo sviluppo di un sistema di interrogazione di un LLM Open Source con accesso a una base dati personale tramite un sistema RAG. |
| ***Contenuti*** | L'offerta didattica è organizzata in 5 moduli suddivisi in 3 lezioni da 4h:<br><br>Lezione n. 1: Modulo Base (MB - 2h), Modulo Prompt Engineering (MP - 2h)<br>Lezione n. 2: Modulo Prompt Engineering Open (MPO - 2h), Modulo Information Retrieval (MIR - 2h)<br>Lezione n. 3: Modulo Retrieval Augmented Generation (MRAG - 4h)<br><br>Per ogni modulo sono descritti di seguito i contenuti: |

| | |
|---|---|
| | **1) Modulo Base (MB) [2h - 6 maggio dalle 14:30 alle 16:30]**<br><br>Nel Modulo Base vengono introdotti tutti i concetti principali necessari per comprendere i meccanismi di base, come il meccanismo di attenzione, che hanno portato al successo dei LLM rispetto ai modelli precedenti, sono approfondite le tecniche di addestramento, le caratteristiche dei LLM e le modalità di personalizzazione dei modelli tramite l'utilizzo di basi di conoscenza esterna per il loro utilizzo in specifici domini o per la limitazione dei fenomeni di allucinazione.<br><br>I contenuti sintetici sono i seguenti:<br><br>● Transformer model<br>● Pre-training<br>● LLM<br>● Transfer Learning, In Context Learning, Fine Tuning, Knowledge Injection<br>● Modelli generativi basati su LLM<br><br>**2) Modulo Prompt Engineering (MP) [2h - 6 maggio dalle 16:30 alle 18:30]**<br><br>Nel Modulo Prompt Engineering vengono introdotti tutti i concetti principali necessari per comprendere quali sono le principali tecniche e caratteristiche di realizzazione di prompt efficaci per l'esecuzione di specifici task con i LLM, come gli approcci di tipo zero-shot e few-shot learning.<br><br>I contenuti sintetici sono i seguenti:<br><br>● In Context Learning con ChatGPT<br>● Introduzione ai sistemi Retrieval Augmented Generation (RAG) per la l'accesso del modello a basi di conoscenza specifiche e per la limitazione delle allucinazioni.<br><br>**3) Modulo Prompt Engineering Open (MPO) [2h - 13 maggio dalle 14:30 alle 16:30]**<br><br>Nel Modulo Prompt Engineering Open vengono estese le conoscenze e competenze acquisite nel modulo MP per permettere l'utilizzo di LLM Open Source, quali ad esempio Llama-2, come alternativa ai modelli GPT. Nel modulo viene inoltre illustrata la metodologia di sviluppo che consente di eseguire (ed eventualmente personalizzare tramite, ad esempio, fine-tuning) questi modelli in locale per poterli interrogare in maniera automatizzata tramite librerie dedicate.<br><br>Il modulo prevede un'esercitazione sul setup di un LLM Open Source<br><br>● *NOTE:*<br>1. *Richiede account Google per utilizzo di Python Notebook su Colab*<br>2. *Richiede Token HuggingFace per scaricamento LLM*<br><br>**4) Modulo Information Retrieval (MIR) [2h - 13 maggio dalle 16:30 alle 18:30]**<br><br>Nel Modulo Information Retrieval vengono introdotte in maniera dettagliata e approfondita le tecniche di rappresentazione e calcolo vettoriale utilizzate ida sistemi di Information Retrieval per rappresentare e ricercare le informazioni su base semantica.<br><br>Il modulo prevede un'introduzione teorica e pratica ai seguenti aspetti:<br><br>● Vettorizzazione testi (es. TF-IDF)<br>● Distanza vettoriali (Cosine Similarity)<br>● Modelli KNN<br>● Introduzione ai sistemi RAG per la ricerca di documenti |

| | |
|---|---|
| | **5) Modulo Retrieval Augmented Generation (MRAG) [4h - 20 maggio dalle 14:30 alle 18:30]**<br><br>Nel Modulo Retrieval Augmented Generation viene proposta ai partecipanti una esercitazione in prima persona sullo sviluppo di un sistema di interrogazione di un LLM Open Source con accesso a una base dati personale tramite un sistema RAG. I partecipanti avranno a disposizione un notebook con il codice in linguaggio Python già predisposto per lo scaricamento in locale del modello e una serie di contenuti informativi con le specifiche e le indicazioni necessarie per sviluppare in autonomia l'esercitazione ricercando nella documentazione fornita quello che serve per implementare la soluzione al problema proposto nel testo della esercitazione.<br><br>I contenuti sintetici sono i seguenti:<br><br>● Introduzione ai sistemi RAG per la generazione di testi basati su informazioni presenti in una specifica base di dati definita dall'utente<br>● Esercitazione sulla realizzazione di un sistema RAG con LLM Open Source su piattaforma Colab<br>● *NOTE:*<br>1. *Richiede account Google per utilizzo di Python Notebook su Colab*<br>2. *Richiede Token HuggingFace per scaricamento LLM* |
| ***Metodo d'insegnamento*** | Presentazioni di vari relatori e attività di tipo laboratoriale nella quale sarà data agli studenti la possibilità di provare le esperienze di in-context learning, fine-tuning, e knowledge injection con diversi LLM e di sviluppare un sistema RAG per la gestione delle informazioni nei LLM. |
| ***Verifiche dell'apprendimento / procedure di valutazione*** | Al termine del seminario gli studenti dovranno compilare un questionario di verifica dell'apprendimento. |
| ***Altre informazioni*** | Partecipazione limitata a **42** studenti dei seguenti percorsi formativi:<br>● Informatica applicata e data analytics (CdL IADA);<br>● Data Science (CdLM DSBAI);<br>● Informatica (CdL e CdLM in INF).<br><br>Per ogni corso di laurea saranno riservati 14 posti con la possibilità di compensare le minori iscrizioni di un percorso con quelle degli altri.<br>Il seminario si terrà in lingua italiana presso il Laboratorio M, piano terra, Palazzo delle Scienze.<br>L'ammissione è determinata dall'ordine cronologico di arrivo della domanda di iscrizione, da effettuarsi entro il giorno **22 aprile 2024** al seguente indirizzo:<br><br>https://docs.google.com/forms/d/e/1FAIpQLScUHR1JPItDMTyawsNcQn6bSvjs85E7vxxee3kvBNbmUHvHQg/viewform?usp=sf_link |