


IA generativa - Modelli linguistici massivi (Large Language Model - LLM)

Sviluppare nuove forme di contenuto utilizzando gli LLM

 A distanza



2 giorni (17 Ore)

Open : 1.990,00 € +IVA

WebCode: IT.1.17

Packaged in azienda : 4.380,00 € +IVA +10% di
Project Management (Quota riferita ad un gruppo
di 10 pax max)

Customized : Su richiesta

L'avvento dei **Large Language Models (LLM)**, come **ChatGPT**, è un passo cruciale per l'IA: le macchine possono oggi comprendere una quantità incredibile di contenuti e sviluppare nuove forme di servizi con altissimi livelli di efficienza. Le aziende e le organizzazioni non possono perdere l'occasione di trarre vantaggio da questi **importanti sviluppi** in termini di creazione di valore. È quindi necessario che sappiano comprendere i meccanismi alla base di questi modelli, scegliere il LLM giusto, conoscere i principali ecosistemi del settore, comprendere la personalizzazione in uso tramite **Prompt Engineering**, valutare la possibilità di arricchire questi modelli tramite fonti interne o strumenti), affrontare con serenità i processi di messa a punto per casi d'uso specifici e, infine, identificare le problematiche legate all'implementazione di tali modelli. Attraverso questo corso di formazione, i partecipanti saranno in grado di cogliere, sulla base di spiegazioni e **workshop su casi concreti**, ciascuna delle componenti significative di questi modelli dirompenti.

A chi è rivolto



Per chi

- Project Manager
- Sviluppatori
- Data Scientist



Prerequisiti

- Nessuno

Programma

1-Introduzione

- Un cambiamento di paradigma
- Cosa sta cambiando?
- Una pubblicazione fondamentale
- Una legge di scala per i modelli linguistici
- Evoluzione temporale dei LLM
- Nuovi ecosistemi
- L'era del Post Deep Learning
- La personalizzazione per Prompt
- Personalizzazione per arricchimento
- Personalizzazione tramite fine-tuning

2-Casi d'uso

- Agenti conversazionali e assistenti virtuali
- Generazione e debug del codice
- Analisi dei sentimenti e delle opinioni
- Classificazione e clustering del testo
- Sintesi di testi o corpo
- Traduzione
- Generazione di contenuti
- Altri casi d'uso significativi

Laboratorio: Prova su casi concreti

3- Fondamenta

- Elaborazione del linguaggio naturale (NLP)
- L'architettura dirompente di Transformers
- La tokenizzazione del testo
- Codifica di un Transformer
- Il livello di incorporazione
- Codifica di posizionamento
- Vettore di posizionamento
- Il meccanismo di attenzione multi-testa
- Punti chiave del meccanismo di attenzione
- La "specializzazione" delle teste di attenzione
- Calcolo dei punteggi di attenzione
- Addizione e normalizzazione delle uscite
- Il decodificatore trasformatore
- Decodifica di Trasformer
- Lo strato nascosto di auto-attenzione
- Lo strato di attenzione del decodificatore
- Gli strati superiori del decodificatore

4- In pratica

- Scegliere un LLM
- Criteri di selezione
- Tre classi di modelli
- Modelli a codificatore singolo
- Modelli BERT focalizzati
- Modelli a singolo decodificatore
- Focus sui modelli GPT
- Un'abbondanza di modelli derivati
- La battaglia per gli LLM
- La corsa verso LLM leggeri
- L'esempio di LaMa
- Tre approcci alla riduzione
- Ecosistemi chiave
- API per modelli chiusi
- HuggingFace e modelli aperti
- Ecosistemi applicativi di tipo LangChain
- LLMops e MLFlow

Laboratorio: Come affrontare gli ecosistemi chiave di un LLM

5- Implementazione

- Scelta del servizio / in house / ibrido
- Flusso di lavoro in house
- Flusso di lavoro del servizio
- Ecosistemi di formazione e inferenza
- Addestramento di un modello massivo
- Fase di valutazione del modello
- Impostazione degli iperparametri
- Distribuzione di un modello
- Messa a punto del modello
- Ingegneria dei prompt
- LLMs MLOps

Laboratorio: ambiente di implementazione di LLM

6- Prompt Engineering

- Configurazione dei parametri LLM
- Cos'è un token?
- Nozione di distribuzione di probabilità LLM
- Campionamento top-K e top-p
- Temperatura del modello
- Impostazione dei parametri nella pratica
- I componenti di un prompt
- Regole generali
- L'approccio Few-Shot Learning
- Da zero a uno al Few-Shot Learning
- Approccio Chain of thoughts
- Suggerimenti per la Chain of thoughts
- Approcci più avanzati
- ReAct Prompting
- Metodo ReAct

Laboratorio: Prompt Engineering su casi concreti

7- LLM Aumentati

- Oltre il prompt, arricchire i LLM
- Aggiungere capacità di memoria
- Memoria tampone
- Diversi meccanismi di memoria
- Le memorie nell'ecosistema LangChain
- Espansione della conoscenza
- Generazione aumentata del recupero (RAG)
- Partizione di testi esterni
- Proiezione semantica dei documenti (Embeddings)
- Database vettoriali
- Algoritmi di ricerca per database vettoriali
- Una galassia di strumenti possibili!

Laboratorio: Implementazione di un agente conversazionale

8- Impiego dei LLM

- Quando l'ingegneria immediata non è più sufficiente
- Che cos'è il fine tuning?
- Tre tecniche classiche
- Reinforcement Learning by Human feedback (RLHF)
- Dettagli della messa a punto supervisionata
- Le tre opzioni per la regolazione dei parametri
- Gli approcci PEFT (Parameter Efficient Tuning)
- Il metodo LoRA (Low Rank Adaptation)
- Una variante efficiente: QLoRA
- Che cos'è la messa in servizio di LLM?
- Registrazione del modello nel registro dei modelli
- Creazione di un endpoint per il modello
- Interrogazione dell'endpoint



Obiettivi del corso

- Comprendere l'utilità degli LLM
- Padroneggiare l'uso degli LLM e dell'IA generativa
- Sviluppare nuovi progetti con l'aiuto delle LLM



Esercitazioni

- Un approccio formativo basato sull'alternanza di fasi teoriche, workshop pratici, feedback e sessioni di discussione.
- Condivisione delle migliori pratiche da parte di consulenti esperti in Intelligenza Artificiale.



Date 2026



Ultimi posti



Edizione garantita

dal 26 mar al 27 mar

dal 16 lug al 17 lug

dal 14 set al 15 set

dal 26 nov al 27 nov