UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

COURSE SYLLABUS

Information Retrieval

2627-2-F1802Q127

Obiettivi

Conoscenza e capacità di comprensione
L'obiettivo del corso è fornire conoscenze e capacità di comprensione dei principali concetti, modelli e tecniche alla base della progettazione dei sistemi di Information Retrieval, noti anche come motori di ricerca.

Il corso introduce l'Information Retrieval come l'area che si occupa dell'accesso automatico a informazioni testuali in risposta ai bisogni informativi degli utenti, espressi attraverso query.

Gli studenti acquisiranno conoscenze sulle principali fasi delle pipeline alla base del funzionamento di un sistema di Information Retrieval, tra cui il preprocessing del testo, la rappresentazione del testo, l'indicizzazione, l'elaborazione delle query, il recupero, il ranking e la valutazione.

Comprenderanno i principali modelli di recupero utilizzati per stimare la rilevanza dei documenti rispetto a una query utente, inclusi approcci basati su corrispondenza esatta, lessicali, probabilistici, neurali e ibridi.

Il corso mostrerà come l'Information Retrieval costituisca una componente fondamentale di tecnologie moderne, tra cui motori di ricerca Web, sistemi di ricerca per domini specifici, Retrieval-Augmented Generation e sistemi conversazionali basati su Large Language Models.

Le attività di laboratorio supporteranno la parte teorica del corso attraverso la progettazione, l'implementazione e la valutazione pratica di applicazioni di Information Retrieval mediante software open-source.

Durante le attività di laboratorio, gli studenti avranno l'opportunità di affrontare un problema di ricerca reale e di presentare la propria soluzione. Questa attività permetterà agli studenti di applicare le conoscenze acquisite durante il corso e di mettere alla prova la propria capacità di analizzare un problema, compiere scelte tecniche di progettazione e sviluppare una soluzione originale.

Capacità di applicare conoscenza e comprensione
Attraverso le lezioni e le attività di laboratorio, gli studenti saranno incoraggiati e valutati rispetto alla loro capacità

di applicare le conoscenze acquisite ai temi trattati nel corso.

Autonomia di giudizio
Il corso mira a promuovere l'autonomia di giudizio e le capacità di analisi critica in relazione alle principali sfide dell'elaborazione del linguaggio naturale e della ricerca di informazioni, nonché ai relativi compiti principali. Tali competenze saranno ulteriormente sviluppate attraverso discussioni in aula e attività di laboratorio.

Abilità comunicative
Il corso favorisce lo sviluppo della capacità di comunicare contenuti tecnici, idee, problemi e relative soluzioni in modo chiaro e non ambiguo a diversi tipi di pubblico. Tali abilità saranno sviluppate durante il corso e valutate come parte dell'esame finale, in particolare attraverso la presentazione del progetto.

Capacità di apprendimento
Il corso è progettato per fornire sia conoscenze teoriche sia competenze pratiche, offrendo una solida base di partenza per ulteriori studi individuali sui principi della rappresentazione, dell'analisi e del recupero del testo.

Contenuti sintetici

- Il corso introduce i principali concetti dell'Information Retrieval e il ruolo dei sistemi di ricerca nell'accesso all'informazione testuale.
- Il corso presenta tecniche per il preprocessing del testo, la rappresentazione del testo e l'indicizzazione, incluse le rappresentazioni bag-of-words e gli indici invertiti.
- Saranno introdotti i principali modelli di Information Retrieval per stimare la rilevanza dei documenti rispetto alle query degli utenti, dal retrieval booleano e dai modelli a spazio vettoriale, con particolare attenzione a BM25, fino ai modelli probabilistici e neurali.
- Il corso tratterà inoltre la metodologia standard per la valutazione dei sistemi di ricerca, incluse le collezioni di test, i giudizi di rilevanza, le metriche di ranking e i test di significatività statistica.
- Saranno presentate le tecniche alla base dei motori di ricerca Web, insieme ad approcci per l'interazione con l'utente, la personalizzazione, la ricerca in domini specifici e il learning to rank per la ricerca Web.
- La parte finale del corso introdurrà i moderni sistemi di retrieval neurali e basati su Large Language Models, incluse le rappresentazioni neurali del testo, il dense retrieval, il retrieval neurale sparse e ibrido, il reranking, la Retrieval-Augmented Generation e la ricerca in sistemi agentici.

Programma esteso

1. Introduzione all'Information Retrieval e ai sistemi di ricerca. Principali concetti dell'Information Retrieval, motori di ricerca, biblioteche digitali, enterprise search, moderni sistemi di ricerca basati su AI ed esempi di applicazioni dell'Information Retrieval.
2. Preprocessing, rappresentazione e indicizzazione del testo. Tokenizzazione, normalizzazione, stemming, lemmatizzazione, stopword, n-grammi, metadati, campi dei documenti, rappresentazioni bag-of-words, indici invertiti, posting list, indici posizionali, compressione degli indici e pipeline di indicizzazione.
3. Query, rilevanza e retrieval exact-match. Tipi di query, unità documentali, bisogni informativi, intento dell'utente, rilevanza, giudizi di rilevanza, retrieval booleano, operatori booleani, query di frase, ricerca per prossimità e query strutturate.
4. Modelli di ranking lessicale. Frequenza dei termini, inverse document frequency, TF-IDF, modello a spazio vettoriale, similarità coseno, retrieval probabilistico, BM25, parametri di BM25, BM25 con campi e confronto tra TF-IDF e BM25.
5. Valutazione dei sistemi di Information Retrieval. Collezioni di test, topic, query, qrels, giudizi di rilevanza e misure standard di valutazione per l'Information Retrieval.

6. Interazione con l’utente, personalizzazione e ricerca Web. Riformulazione delle query, sessioni di ricerca, comportamento di click, log di ricerca, personalizzazione, ricerca esplorativa, crawling, indicizzazione Web, analisi dei link, PageRank, rilevamento dello spam, freshness e snippet.
7. Ricerca in domini specifici e Learning to Rank. Ricerca scientifica, ricerca biomedica, ricerca legale, enterprise search, ricerca e-commerce, terminologia di dominio, vincoli di dominio, learning to rank, inclusi approcci pointwise, pairwise e listwise, feature di ranking, LambdaMART e apprendimento basato sui click.
8. Rappresentazioni neurali e dense retrieval. Word embedding, embedding contestuali, sentence embedding, document embedding, rappresentazioni basate su transformer, bi-encoder e vector database.
9. Retrieval neurale sparse, dense, ibrido e multi-stage. Learned sparse retrieval, neural lexical matching, metodi in stile SPLADE, retrieval ibrido sparse-dense, fusione degli score, generazione dei candidati, neural reranking, cross-encoder, modelli a late interaction, ranking in stile ColBERT e trade-off tra efficienza ed efficacia.
10. Sistemi di ricerca basati su LLM, RAG e agenti. Riscrittura delle query, espansione delle query, generazione di query sintetiche, reranking basato su LLM, sintesi dei risultati, generazione delle risposte, architettura RAG, chunking, costruzione del contesto, grounding, generazione di citazioni, controllo delle allucinazioni, valutazione dei sistemi RAG, uso di strumenti, ricerca agentica, memoria, agenti Web e sfide di ricerca aperte.

Prerequisiti

È richiesta una conoscenza di base di statistica, algebra lineare e programmazione, ad esempio in Python.

Modalità didattica

Il corso sarà tenuto in lingua inglese e comprende sia lezioni sia esercitazioni di laboratorio.
Il corso prevede 40 ore di lezioni e 12 ore di esercitazioni di laboratorio.
La maggior parte delle lezioni avrà una durata di 2 ore, anche se alcune lezioni più intensive potranno durare 3 ore, in base all’argomento trattato.

Le lezioni saranno svolte attraverso una combinazione di modalità didattiche.

* La prima parte di ciascuna lezione sarà principalmente basata sulla didattica frontale e introdurrà concetti teorici, modelli formali e metodi tecnici.
* La seconda parte includerà attività didattiche interattive, come discussioni guidate, esempi, brevi esercizi e analisi di casi pratici, con l’obiettivo di incoraggiare la partecipazione attiva degli studenti e supportare l’applicazione dei concetti introdotti durante la lezione.

Durante le esercitazioni di laboratorio, gli studenti utilizzeranno software open-source per progettare, implementare e valutare componenti di sistemi di Information Retrieval.
Saranno organizzati seminari tenuti da esperti di riconosciuto profilo internazionale.
Alcune lezioni potranno essere svolte da remoto in formato interattivo.

Materiale didattico

- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008.
- Bhaskar Mitra and Nick Craswell. An Introduction to Neural Information Retrieval. Foundations and Trends in Information Retrieval, Volume 13, Issue 1, 2018.

Periodo di erogazione dell'insegnamento

Primo semestre.

Modalità di verifica del profitto e valutazione

- Esame scritto individuale composto da esercizi, domande aperte e domande chiuse relative ai contenuti del corso.
- La valutazione comprende inoltre un progetto di laboratorio, che potrà essere sviluppato individualmente o in gruppi di massimo tre studenti.
- L’esame scritto è finalizzato a valutare il livello di comprensione degli aspetti teorici e tecnici di base trattati durante il corso.
- L’obiettivo del progetto è utilizzare software open-source per sviluppare soluzioni tecnologiche a problemi di ricerca affrontati nel corso.
- Il progetto si concentrerà su scenari applicativi reali che richiedono agli studenti di progettare, implementare e valutare componenti di sistemi di Information Retrieval introdotti durante il corso.

Orario di ricevimento

Da concordare con il docente.

Sustainable Development Goals