UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Semantics

2425-1-FDS01Q010

Obiettivi

Lo scopo principale del corso è fornire agli studenti le conoscenze e competenze necessarie per comprendere e risolvere problemi di legati all'interpretazione semantica dei dati in applicazioni di data science, con particolare riferimento a problemi di **rappresentazione**, **integrazione**, e **arricchimento** di dati eterogenei e **analisi semantiche di testi**.

In particolare verranno presentati i due principali paradigmi semantici proposti nell'ambito dell'Intelligenza Artificiale:

- Semantica dichiarativa, basata su paradigmi logici-formali, con particolare attenzione sulla ormai diffusa astrazione dei **Grafi di Conoscenza**
- Semantica distribuzionale e basata sulla modellazione del linguaggio, con particolare attenzione a **modelli del linguaggio di grandi dimensioni**
  Infine verranno presentate metodologie neuro-simboliche in cui questi paradigmi sono opportunamente combinati per supportare applicazioni quali la preparazione dei dati, la costruzione di basi di conoscenza, la ricerca semantica, e la generazione di contenuti sulla base di informazioni esterne.

Gli argomenti che verranno trattati hanno un duplice scopo: 1) fornire un insieme di strumenti teorici e pratici per rappresentare, organizzare, pubblicare, interrogare, riconciliare, esplorare e interpretate dati e conoscenze in scenari applicativi reali (ampiamente discussi durante le lezioni frontali e affrontati durante le esercitazioni) utilizzando tecnologie semantiche e 2) acquisire le competenze necessarie per comprendere problemi di interoperabilità semantica nuovi e le tecniche necessarie per risolverli adeguatamente indipendentemente dalle particolari tecnologie di riferimento.

Contenuti sintetici

Il corso presenta strumenti computazionali per rappresentare, armonizzare e ricostruire la semantica dei dati utilizzati in applicazioni di data science, con particolare attenzione a:

- modelli e linguaggi elaborati nell'ambito del web semantico per supportare l'integrazione di dati eterogeni (**knowledge graph**, **ontologie**, **RDF, RDFS, OWL**);
- modelli per apprendere la semantica dai dati, con particolare riferimento a dati in formato testuale (**word embeddings**, **Large Language Models (LLM)**)
- tecniche neurali per la **riconciliazione di dati**;
- tecniche di elaborazione del linguaggio naturale per **estrarre informazioni strutturate da testi**;
- tecniche per **integrare knowledge graph e LLM**.

Programma esteso

1. **Data Semantics:** Semantica dei dati ed applicazioni di data analytics (big data, sorgenti web, formati eterogenei, integrazione di informazioni ed arricchimento semantico, connessione tra dati, knowledge graph)
2. **Knowledge Graph e Web Semantico:** rappresentazione e interogazione dei dati nel web semantico (RDF, SPARQL, tecnologie semantiche e architettture, rappresentazioni in ambito industriale mediante basi di dati a grafo). Esercitazione su interrogazione di Knowledge Graph pubblici con SPARQL; definizione di vocabolari condivisi mediante ontologie e linguaggi logico-formali (dai vocabolari condivisi alle ontologie, tassonomie, ontologie lessicali, ontologie assiomatiche, ragionamento automatico e semantica, RDFS, OWL). Esercitazione su modellazione di ontologie mediante RDFS e OWL.
3. **Semantica distribuzionale e apprendimento di rappresentazioni:** introduzione alla semantica distribuzionale e all'apprendimento di rappresentazioni distribuite (semantica distribuzionale); modelli per apprendere rappresentazioni distribuite da corpus testuali (word embeddings e word2vec, contextual word embeddings e Large Language Models - LLM). Esercitazione su LLM e attenzione. Seminario: modelli per comparare rappresentazioni distribuite differenti per applicazioni di computational social science e cultural analysis (allineamento tra word embeddings, analisi diacroniche, studi basati su word embeddings con WEAT e SWEAT).
4. **Riconciliazione semantica:** algoritmi di entity matching basati su reti neurali (deep matcher, Ditto, BERT-based matching, matching con large language models).
5. **Elementi di NLP - tecniche di estrazione di informazioni:** introduzione e presentazione di alcuni approcci all'estrazione di informazioni strutturate da testo e altri dati semi strutturati (named entity recognition, entity linking, estrazione di relazioni, semantic table interpretation). Esercitazione su named entity recognition e named entity linking.
6. **Tecniche di accesso alle informazioni mediate dalla semantica:** tecniche semantiche per l'esplorazione di informazioni (faceted search,retrieval augmented generation)

Prerequisiti

Conoscenze matematiche e informatiche insegnate nei corsi obbligatori del primo semestre.

Modalità didattica

Lezioni frontali ed esercitazioni con i personal computer degli studenti. Uso della piattaforma Moodle. Seminari su applicazioni delle tecnologie semantiche a problemi reali da parte di experti del mondo dell'industria.

Didattica Erogativa: ~32h (lezioni frontali)
Didattica Interattiva: ~12h (esercitazioni guidate)

Insegnato in Inglese

Materiale didattico

Knowledge Graphs: Fundamentals, Techniques, and Applications. Kejriwal, Mayank, Craig A. Knoblock, and Pedro Szekely. MIT Press, 2021.

The Web of Data. Aidan Hogan. 2020. Springer. Pages 1-680.

Verrà fornito agli studenti materiale aggiuntivo sotto forma di presentazioni e articoli scientifici per coprire gli argomenti più recenti non coperti dal libro di testo.

Periodo di erogazione dell'insegnamento

Semestre II

Modalità di verifica del profitto e valutazione

La valutazione finale è costituita dall'aggregazione dei punteggi ottenuti in due valutazioni indipendenti.

- La prima valutazione è basata su un **progetto d'esame,** effettuato individualmente o in gruppo, e finalizzato all'approfondimento di un argomento specifico trattato nel corso o collegato ad argomenti trattati nel corso; il progetto viene discusso attraverso una **presentazione orale supportata da slide** della durata di 20 min circa; è possibile, durante la presentazione, includere una breve demo del progetto svolto. *La valutazione si basa su: significatività del progetto rispetto agli argomenti trattati nel corso, rigore metodologico (nei limiti di quanto ragionevole chiedere per un progetto d'esame); padronanza dell'argomento approfondito dimostrata durante la presentazione orale*.
- La seconda valutazione è basata sulla **verifica della conoscenza degli argomenti affrontati durante il corso** mediante valutazione di esercizi (assignment) da completare individualmente e discussione orale. Gli assignment verranno valutati e discussi in sede d'esame, dopo la discussione del progetto.

Orario di ricevimento

Su richiesta

Sustainable Development Goals