UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Semantics

2122-1-F9101Q011

Obiettivi

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, riconciliazione e integrazione di dati eterogenei e ad analisi di testi che debbano tenere conto del significato delle parole in essi contenuti.

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, data linking, ontologie, RDF, RDFS, OWL);
- tecniche per integrare dati e vocabolari;

- tecniche per estrarre informazioni strutturate da testi;
- tecnique per supportare l'accesso a grandi quantità di conoscenze.

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.
3. **Rappresentazione della Conoscenza e Ragionamento Automatico:** 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, SWRL). Esercitazione su modellazione di ontologie mediante i linguaggi RDFS e OWL.
4. **Riconciliazione semantica:** riconciliazione di ontologie e vocabolari (ontology matching per allineare ontologie e tassonomie, terminologia e mapping, similarità semantica e combinazione di diverse funzioni di similarità, selezione dei mapping). Riconciliazione a livello dei valori o delle istanze (deduplicazione e record linkage, approcci probabilistici, metriche di distanza e misure di similarità, combinazione e apprendimento di misure di similarità complesse, strategie per la fusione di infromazioni eterogenee, misure di similarità basate su grafi). Esercitazione su riconcilizione di dati con l'aiuto di strumenti esistenti.
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). .
6. **Esplorazione di informazioni e conoscenze:** tecniche semantiche per l'esplorazione passiva e attiva di informazioni (faceted search, sistemi di raccomandazione).
7. **Elementi di NLP - 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); modelli per comparare rappresentazioni distribuite differenti (allineamento tra word embeddings, analisi diacroniche, studi basati su word embeddings con WEAT e SWEAT).

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.

?L'attività didattica sarà erogata in presenza, salvo indicazioni diverse, nazionali e/o di Ateneo, dovute al protrarsi dell'emergenza COVID-19.

Insegnato in Inglese

Materiale didattico

ITA: Tommaso Di Noia, Roberto De Virgilio, Eugenio Di Sciascio, Francesco M. Donini. Semantic Web: tra ontologie e Open Data, Apogeo, 2013.

ENG: Grigoris Antoniou, Paul Groth, Frank van Harmelen, *A Semantic Web Primer*, (Third Edition), MIT Press, 2012.

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

Giovedì 14.30-15.30