UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Data Semantics

2526-1-FDS02Q010

Aims

Knowledge and understanding (DdD 1)

By the end of the course, students will have acquired knowledge of:

- the fundamental principles of data semantics and their role in data science applications;
- the two main semantic paradigms: declarative semantics, based on logic-based models and knowledge graphs; and distributional semantics, based on representation learning from textual data and large language models (LLMs);
- neuro-symbolic methodologies that combine symbolic and neural approaches for semantic representation and processing;
- core Semantic Web technologies (RDF, RDFS, OWL, SPARQL) and the foundations of ontologies, taxonomies, and automated reasoning;
- word embeddings, pre-trained language models, large language models, and their applications;
- techniques for building, enriching, and making knowledge bases usable, including heterogeneous data reconciliation, entity and relation extraction, and question answering over data and documents.

Applying knowledge and understanding (DdD 2)

Students will be able to:

- model and query knowledge graphs using Semantic Web languages and tools;
- design and use ontologies to support semantic integration of data in application scenarios;
- apply distributional models and LLMs to text interpretation and knowledge extraction from unstructured data;
- implement solutions for semantic reconciliation, entity extraction, and natural language question answering;
- address real-world semantic interoperability problems by selecting appropriate methods and tools for the specific context;
- develop small-scale applications for data analysis, exploration, and content generation combining

knowledge graphs, NLP, and LLMs .

Other Dublin Descriptors (DdD 3, 4, 5)

- Making judgments: students will develop the ability to critically assess different semantic solutions and to select suitable methodologies for novel problems, also through project work and seminar discussions.
- Communication skills: through oral presentations of projects and interactive exercises, students will be able to clearly communicate complex concepts and results, including in interdisciplinary contexts.
- Learning skills: the course provides theoretical and practical tools that enable students to continue studying semantic technologies independently, including through reading scientific literature and advanced materials.

Contents

The course presents computational methods to represent, harmonize and interpret the semantics of data used in data science applications, with a particular focus on:

- models and languages developed within the semantic web to support the integration of heterogeneous data (**knowledge graph**, **ontologies**, **RDF, RDFS, OWL**);
- models to learn (semantic) representations from data, especially from text corpora (**word embeddings**, **Large Language Models**);
- techniques for the **integration of knowledge graphs and LLMs**.
- neural techniques for **data matching**;
- **information extration** techniques, with particular enphasis on entity extraction;
- **question ansering techqniques** and **retrieval-augmented generation**

Detailed program

1. **Data semantics:** the role of semantics in data analytics (big data, web sources, heterogeneous formats, information integration, semantic enrichment, data linking, knowledge graphs).
2. **Knowledge graphs and the semantic web:** representation and query of data in the semantic web (RDF, SPARQL, semantic technologies and architectures, corporate knowledge graphs with graph databases). Excercise on querying RDF knowledge graphs with SPARQL; definition of shared vocabularies with ontologies and logic-based languages ??(from shared vocabularies to ontologies, taxonomies, lexical ontologies, axiomatic ontologies, automatic reasoning and semantics, RDFS, OWL). Excercises on ontology modeling with RDFS and OWL.
3. **Distributional semantics and language models:** introduction to distributional semantics and distributed representations (distributional semantics); models for learning distributed representations from textual corpora (word embeddings and word2vec, Large Language Models - LLMs). Exercises on LLMs and attention. Seminar: models to compare different distributed representations (alignment between word embeddings, diachronic language studies, studies based on word embeddings with WEAT and SWEAT).
4. **Semantic reconciliation:** neural network-based entity matching algorithms (deep matcher, Ditto, BERT-based matching).
5. **Introduction to NLP - information extraction:** presentation of selected approaches to the extraction of structured information from texts and other semi-structured data (named entity recognition, entity linking, relationship extraction, semantic table interpretation). Esercitazione su named entity recognition e named entity linking
6. **Information and knowledge exploration:** semantic techniques for the exploration of information (semantic search, retrieval augmented generation).

Prerequisites

Mathematics and computer science as taught in the compulsory courses of the first semester.

Teaching form

Lectures and exercises with students' personal computers. Moodle e-learning platform. Seminars about the usage of semantics in real-world applications given by experts from the industry.

Teacher-centered lessons: ~32h
Interactive lessons: ~12h (hands-on sessions)

Textbook and teaching resource

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.

Daniel Jurafsky and James H. Martin. 2025. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition. (Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3)

Additional material such as presentations and articles is provided to cover novel topics that are not covered by the textbook.

Semester

Semester II

Assessment method

The final evaluation consists of the aggregation of the scores obtained in two independent assessments.

- The first assessment is based on an exam-tailored project, carried out individually or in groups and aimed at bringing the student to have in-depth knowledge and/or hands-on experience of a specific topic covered in the course or linked to topics covered in the course; the project is discussed through an oral presentation supported by slides lasting about 20 minutes; it is possible, during the presentation, to include a short demo of the project. The evaluation is based on: significance of the project for the topics covered in the course, methodological soundness (within the limits of what is reasonable to ask for an exam project); mastery of the in-depth topic demonstrated during the oral presentation.
- The second assessment is based on the evaluation of the knowledge acquired by the student on the topics

addressed during the course through the discussion of assignments that students must execute individually as homework. Assignments will be evaluated and discussed during the oral exam after the presentation of the project.

Office hours

On demand

Sustainable Development Goals

QUALITY EDUCATION | INDUSTRY, INNOVATION AND INFRASTRUCTURE