Knowledge Graphs with Large Language Models

Dr. Panos Alexopoulos

p.alexopoulos@gmail.com

Course Overview

In the past decade the concept of a knowledge graph has emerged as a pivotal framework for managing and utilizing interconnected data and information. Stemming from Google's proclamation that shifted the focus from searching for strings to understanding entities and relationships, the term encapsulates a network of interconnected entities and concepts that facilitates semantic data integration, sharing, and utilization within organizations.

At the same time, the advent of Large Language Models (LLMs) has revolutionized the field of natural language processing and artificial intelligence. These models, exemplified by architectures like GPT-4, have the capability to comprehend and generate human-like text on an unprecedented scale. By leveraging vast amounts of pre-existing linguistic data, LLMs excel in tasks such as language translation, text completion, and even creative content generation.

The synergy between knowledge graphs and LLMs represents a powerful combination that can significantly enhance various stages of knowledge graph development, such as schema design, knowledge acquisition and quality control, while reducing LLMs hallucinations and improving their accuracy, reliability, and explainability.

This course is designed to equip students with a thorough comprehension of the synergistic potential between knowledge graphs and LLMs. Through a mix of theory and practice, the students will explore all the necessary steps for initiating and executing a knowledge graph development project, receiving practical guidance on the effective utilization of LLMs at different stages of the process. The course will also cover techniques by which knowledge graphs can be used to augment the accuracy, reliability, and explainability of LLMs.

Key Outcomes

By the end of this course, students will be able to:

- Analyze and determine when a knowledge graph is the optimal solution for a specific data management or AI challenge, and develop a detailed implementation strategy.
- Design a knowledge graph schema that facilitates ease of development and aligns with specified requirements.
- Use state-of-the-art tools and methods to (semi-)automatically populate a knowledge graph from diverse data sources.
- Develop and apply mechanisms to assess and enhance the quality of a knowledge graph.
- Use Large Language Models (LLMs) to improve each step of the knowledge graph development process.
- Integrate knowledge graphs with LLMs to improve their accuracy and reliability.

Requirements and Prerequisites

Students should be familiar with entity-relational database design and have working knowledge of Python programming. Prior experience with Natural Language Processing is useful but not obligatory.

Course Materials

There is no required textbook. All course materials will be provided in class and be available for downloading.

Books

The following books provide a good supporting material for the topics covered by the course:

- “Semantic Modeling for Data”, P. Alexopoulos, O’Reilly Media, 2020
- “Building Knowledge Graphs: A Practitioner’s Guide”, J. Barassa, Jim Webber, O’Reilly Media, 2023
- “Semantic Web for the Working Ontologist”, D. Allemang, J. Hendler, ACM Books, 3rd edition, 2020
- “Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs”, S. Ozdemir, Addison-Wesley Professional; 1st edition, 2023

Software/Computing requirements

- The examples of the course will be implemented in Python on Google Colab (https://colab.research.google.com/)
- No special hardware is required

Grading

Students will be graded on their performance as follows:

- Assignments (60%)
- Final Exam (40%)

Late Assignments

Late assignments will either not be accepted or will incur a grade penalty unless due to documented serious illness or family emergency. Exceptions to this policy for reasons of civic obligations will only be made available when the assignment cannot reasonably be completed prior to the due date, you make suitable arrangements, and give notice for late submission in advance.

Attendance Requirements

Class attendance is essential to succeed in this course. An excused absence can only be granted in cases of serious illness or grave family emergencies and must be documented. Job interviews and incompatible travel plans are considered unexcused absences. Where possible, please notify the instructor in advance of an excused absence. Students are responsible for keeping up with the course material, including lectures, from the first day of this class, forward. It is the student's obligation to bring oneself up to date on any missed coursework.

Code of Ethics

Students may not work together on individual graded assignments unless the instructor gives express permission. Exercise integrity in all aspects of one's academic work including, but not limited to, the preparation and completion of all other course requirements by not engaging in any method or means that provides an unfair advantage. In any case of doubt, students must be

able to prove that they are the sole authors of their work by demonstrating their knowledge to the instructor.

Clearly acknowledge the work and efforts of others when submitting written work as one's own. Ideas, data, direct quotations (which should be designated with quotation marks), paraphrasing, creative expression, or any other incorporation of the work of others should be fully referenced. No plagiarism of any sort will be tolerated. This includes any material found on the internet. Reuse of material found in question and answer forums, code repositories, other lecture sites, etc., is unacceptable. You may use online material to deepen your understanding of a concept, not for finding answers.

Course Syllabus

The course comprises five units of three hours each

Unit 1: Understanding knowledge graphs and their relation with large language models

- What are knowledge graphs and why we build them
- Key factors influencing the ease or difficulty of building a knowledge graph
- Large Language Models (LLMs) and their interplay with knowledge graphs
- Knowledge graph development lifecycle
- Crafting a knowledge graph strategy

Unit 2: Designing the knowledge graph schema

- Building elements of a knowledge graph
- Knowledge graph schema design using Semantic Web languages (RDF, OWL, SKOS, SHACL) and Labeled Property Graphs.
- Conceptual modeling best practices, dilemmas, and pitfalls
- Addressing uncertainty and vagueness
- Using LLMs in schema design: what works and what doesn't

Unit 3: Populating the knowledge graph

- Understanding and defining knowledge graph population tasks
- Evaluating and selecting data sources and population tools
- Designing population strategies and pipelines.
- Populating knowledge graphs from structured data

- Populating knowledge graphs from text
- LLMs as knowledge providers and knowledge miners - what works and what doesn’t

Unit 4: Managing knowledge graph quality and evolution

- Knowledge graph quality dimensions and metrics
- Typical quality problems and dilemmas/trade-offs
- Automatically detecting quality problems in knowledge graphs
- Crafting a quality management strategy
- Addressing the challenge of knowledge graph maintenance and evolution
- Detecting, measuring, and monitoring concept drift
- Crafting a knowledge graph governance strategy.

Unit 5: Applying the knowledge graph

- Typical knowledge graph applications
- Entity linking and disambiguation
- Using knowledge graphs to ground Large Language Models