Open Data Project

OBJECTIVES

This is a research effort where students must put in practice the lessons learnt in this course. The project has a strong component based on critical reasoning and then a practical aspect to make you realise when to use graphs in real Data Science projects:

- Comprehend the benefits of graph modelling for automating data integration processes,
- Use either a property graph or a knowledge graph to model your data and exploit the resulting graph with data analysis (e.g., via graph embeddings) in real-world use cases (e.g., recommendations),
- Learn and understand what is SHACL and why it is getting so popular. Position and distinguish it from reasoning in RDFS/OWL and how it is used in real-world use cases.

The outcome of the project can simply be a document or be accompanied by a proof of concept system. More details below.

PRACTICE STATEMENT

There are three statements to choose from. Decide between them according to your genuine interests for your future professional career. Each statement is detailed in the next subsections. Nevertheless, all of them have two components: a research aspect and a practical aspect to understand the potential of graphs in real use cases. In all cases, you are expected to do a search on the topic, summarize what you found and exemplify with a real use case (as realistic as possible) via clear examples showing its feasibility.

1. Graph-Based Data Integration

In the course we saw a brief introduction about how to perform advanced data integration via knowledge graphs. This is one of the strongest points of usage of knowledge graphs in data science projects. The first option for the project is to further investigate it.

The recommended starting point for this topic is the "*Additional Material: Graph-driven Federated Data Management*" from the Graph-Based Data Integration lecture. The tasks are as follow:

- Read, understand and summarize the paper in, at most, 5 pages.
- Aside, you must create a detailed real use case (i.e., consider real data sources, construct the needed constructs according to the paper and showcase an example of interesting query once integrated). This part should be, at most 5 pages.

Nevertheless, you can adapt the topic to your interest. Another valid approach would be to explore the tools out there allowing to perform something similar and discuss its pros and cons with regard to the baseline explained in the course. Almost any company nowadays is building their solution to automate data integration based on graphs. Some pointers you may want to explore (may times this kind of solutions are directly called *data catalogs* but be sure they are based on graphs):

- Delta Lake by Databricks
- DatHub by LinkedIn
- Data Access Layer by Twitter
- DataPortal by Airbnb
- Databook by Uber
- Lexikon by Spotify
- Metacat by Netflix
- Artifact by Shopify
- Atlas by Apache
- Marquez by WeWork
- Amundsen by Lyft

2. Real use cases of property / knowledge graphs and their exploitation via data analysis

Another valid topic for the project is to put into practice graphs for real problems. This usage must span the exploitation phase via advanced data analysis:

- Learn and understand what graph embeddings are. Graph embeddings provide a vector representation of graphs (node-based or edge-based) that can be later used to perform data analysis on them. An starting point for this task could be: https://arxiv.org/abs/1709.07604. You need to summarize the main findings about graph embeddings in at most 5 pages.
- Put what you learnt into practice. Given a real graph (e.g., one of those generated in the labs or any other you may have open access to), briefly describe it to understand its purpose, propose an embedding strategy and run a chosen machine learning or data mining algorithm with a specific purpose (e.g., recommendation). You need to summarize the practical part in at most 5 pages.

3. SHACL Vs. Reasoning and it is use in real use cases

SHACL is gaining a big momentum. As briefly mentioned in the lectures, Shapes Constraint Language (SHACL) is a W3C recommendation that allows to constrain the graph shapes according to your interest. Most available tools support, at least, SHACL core and we can expect to be one of the most popular graph tools in the next years:

- Learn and study SHACL. You can start from: https://www.w3.org/TR/shacl/ but there are plenty of tutorials and available open materials. Summarize your learning about SHACL in at most 5 pages. Include a comprehensive comparison with regard to reasoning in RDFS/OWL with clear pros and cons.

- Put what you learnt into practice. Given a real graph (e.g., one of those generated in the labs or any other you may have open access to) use SHACL to constraint the graph and clearly justify the benefits of such approach. When doing such

critical reasoning, compare what could be done and not be done if you used RDFS/OWL reasoning instead. A suggested way to lead this part is by proposing an exploitation mechanism for the graph and highlight the benefits brought by SHACL (also compared to using reasoning instead or on top of). This part must be, at most, 5 pages long.

DELIVERABLES

By the deadline stated in this event, one person of the group must upload a document in pdf format. If you developed a proof of concept during the project for the practical part, include your Github link (or project page you are using; remember to include the require credentials to access it if needed) and a brief explanation in an additional page at the end of the document.

Thus, the document length must be 10 pages long, unless you included a proof of concept that then you can use up to 11 pages. These are hard constraints, and part of the project's objective is to synthesize and summarize effectively your findings.

PROJECT TEAMS

Use the team creator event to register your project. The team must be of 2 people and can be a repetition of a lab team. So be sure to choose a teammate you feel comfortable working with.

EVALUATION CRITERIA

The project will be evaluated according to the following criteria:

Conciseness

The document fits in 10 / 11 pages and explains all the main details requested.

Understandability

You provide enough details as to assess your solution.

Soundness

There are no contradictions about the choices made and the inherent advantages of the underlying theory chosen.

Proof of Concept

If you provide a PoC, you will be evaluated out of 12 (instead of 10). Whatever mark above 10 you get, it will be automatically transferred to the final exam as a surplus.