Group 38
Haoqi Yu, Bjørn Magnus Hoddevik

Approach to the problem

We began by looking into ways to convert CSV files into RDF triples given our RDFS produced in Assignment 2. We quickly found a python library called “rdfpandas”, which showed promise. However, it was abandoned for the more powerful, and therefore complex, “rdflib” library. After a quick prototyping session, with focus on the student csv file, we committed to this approach. Perhaps some more research into prebuilt tools would’ve saved some frustration.

By working iteratively, with one csv file at a time, we quickly got through the data but not without hiccups. We had to revise our work from Assignment 2 several times, as some attributes were simply missing or our ER-diagram from Assignment 2 had failed to satisfy the data. Moreover, there were some problems that were revealed when we began working on the queries. These problems were related to the way we were producing our RDF triples. We will discuss this in the next section.

We approached each query one at a time, starting with the first. We started small, and built up each query line by line, verifying that nothing went wrong for each step.

Decisions and assumptions

The most challenging part was how we would deal with weak entities and relations. We decided to treat both as a RDFS bag, although we are not certain that is the right term. We treated each ER-diagram entity as the main subject, and then every attribute, weak entity and relation was a predicate with their value as the object. While attribute objects were simple, the weak entities and relations were more tricky. We used BNodes, added them as an object connected to the entity subject. And then used the BNode as the main subject with the attributes of the weak entity or relation as objects.

We then decided to only add the relations to one side, meaning the relation “Teaches” was only added to teachers. There is no reverse of this. Which side the relation was added to was determined by which csv file had the information, meaning that if the csv had information about teachers, then the relation was added to teachers.

I am not aware of any assumptions we made, outside of assuming our RDFS definitions from Assignment 2 (after the aforementioned changes) captured all the relevant data without losing information.

Scripts used

We have attached the file “rdfs_and_csv_to_triples.py” which takes the input file “assignment_2_rdfs_revised.txt” and the given csv files, and outputs the finished graph in the file “graph.txt”. To run the scripts, we have also attached the “requirements.txt” file.

Queries and result

See next pages