DSCI 558: Building Knowledge Graphs

Homework 7: Structured Data & KG Embeddings

Released: Oct 22nd, 2020

Due: Nov 9th, 2020 @ 23:59

Ground Rules

This homework must be done individually. You can ask others for help with the tools, however, the submitted homework has to be your own work.

Summary

This homework spans two topics we recently covered. In the first task, you will extract data from a structured source (tabular data) and model it using the RDF Data Cube ontology. In the second task, you will use the AmpliGraph python library to generate knowledge graph embeddings.

Task 1: Structured Data (6pts)

In this task you will extract quantitative data from World Bank data source (`world-development-indicators.xlsx`) and model the data using the RDF Data Cube ontology. The data holds population numbers and population gender distribution percentages of more than 200 countries from 1960 to 2017.

- The general specifications and some modeling examples using the RDF Data Cube ontology can be found here: https://www.w3.org/TR/vocab-data-cube/

Task 1.1 (3pts)

There are two dimensions (time period, country) and three measures (population, female population percentage and male population percentage) in the data. Create RDF properties to represent them in the Data Cube ontology and write your extension to `world_bank_ont.ttl`.

- Hint: inherit from `qb:DimensionProperty` and `qb:MeasureProperty`.
- For the sake of simplicity, you can use `http://dbpedia.org/resource/Person` and `http://dbpedia.org/page/Percentage` as units for the measures above.

Task 1.2 (3pts)

Convert your data into RDF triples using the RDF Data Cube ontology with your extension. Your RDF triples should include a dataset entity (instance of `qb:Dataset`), and list of observations. Rename the file to `world_bank_data.ttl` for submission.

- You can test your RDF triples by executing the following SPARQL query and verifying with the result below.

Query:

```
PREFIX qb:    <http://purl.org/linked-data/cube#>
PREFIX my_ns: <http://dsci558.org/myfakenamespace#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?population ?female_population ?male_population
WHERE {
 ?obs a qb:Observation ;
      my_ns:refPeriod <http://reference.data.gov.uk/id/gregorian-interval/2014-01-01T00:00:00/P3Y> ;
      my_ns:refArea / rdfs:label "Canada" ;
      my_ns:measure_population ?population ;
      my_ns:measure_femalePopulation ?female_population ;
      my_ns:measure_malePopulation   ?male_population .
}
```

Result:

population	female_population	male_population
35535348	50.39765	49.60235

Task 2: Knowledge Graph Embeddings (4pts)

In this task, you will use the AmpliGraph library to generate KG embeddings using different embedding models and the given dataset file `GoT.csv`. We provide a python notebook (`kge.ipynb`), which contains code and instructions to accomplish this task.

Task 2.1 (3pts)

Run the code under task 2.1 in the notebook, for each type of the models: `TransE`, `DistMult` and `ComplEx`. You need to perform the following for each:

2.1.x.1. Train the model.

2.1.x.2. Evaluate it, include the performance scores in your final report.

2.1.x.3. Generate a visualization of the embeddings using Tensorboard (PCA-reduced, **two** components) using the code in the notebook. Include a screenshot of the plot in your report.

- It is safer to re-execute the whole process for each model type after performing a complete Kernel restart between each one of the sessions)

Task 2.2 (1pts)

Describe and explain (in your report, in a few sentences) what may cause the differences between the different scores you got in 2.1.x.2 between the different models (i.e., compare between the scores in 2.1.1.2, 2.1.2.2, 2.1.3.2)

Notes:

- Once you finish accomplishing all the tasks, change the notebook's filename to `Firstname_Lastname_hw07_kge.ipynb` (you will submit it).
- In order to explore the embedding space (in Tensorboard) with labels (instead of indexes), you should load the file `metadata.tsv` (using the 'Load' button under the 'Data' tab). The file is generated using the last command in the notebook in the resulting directory.
- You can find additional info on AmpliGraph here: https://docs.ampligraph.org/

Submission Instructions

You must submit (via Blackboard) the following files/folders in a single `.zip` archive named `Firstname_Lastname_hw07.zip`:

- `Firstname_Lastname_hw07_report.pdf`: pdf file with your relevant written answers
- `world_bank_ont.ttl`: your extension of the Data Cube ontology in task 1.1.
- `world_bank_data.ttl`: your RDF triples in task 1.2.
- `Firstname_Lastname_hw07_kge.ipynb`: The notebook to accomplish task 2
- `Firstname_Lastname_hw07_kge.pdf`: A printed version of the notebook. You can save your notebook to pdf using 'Print Preview' or 'Download it as PDF' in 'File menu'