Building & Mining Knowledge Graphs

(KEN4256)

Assignment 2: Mining relations from RDF knowledge graphs

Due date: 2 March 2020 (upload files to Student Portal before **noon**)

In this assignment, you will be supplied with an RDF knowledge graph contained in the file "human_kg_train.nt", which is a subgraph extracted from the Wikidata knowledge graph. This subgraph consists of information about people and entities belonging to related classes (e.g. Country, Occupation, Sex, Place). For the five relations listed below, we have removed **some** triples in which these relations occur. Your task will be to predict these missing connections **using only the information in this graph** (no external knowledge should be used, including other information in Wikidata).

For each relation type *p* given below, your general task will be to predict new subject-object pairs (*s,o*) such that <s> <p> <o> could be one of the missing triples in the supplied knowledge graph:

1. country of citizenship : <http://www.wikidata.org/prop/direct/P27>
2. sibling : <http://www.wikidata.org/prop/direct/P3373>
3. mother : <http://www.wikidata.org/prop/direct/P25>
4. place of death : <http://www.wikidata.org/prop/direct/P20>
5. spouse : <http://www.wikidata.org/prop/direct/P26>

Tasks:

1. Define hand-crafted SPARQL queries (based on "horn-like" rules) for each relation *p* above to insert the missing relations <s> <p> <o> into the given graph.

   Note: You may also supply queries based on **certain** rules (whose confidence score is always 100%) but keep in mind that there might not be any suitable ones for this particular graph. For each **uncertain** hand-crafted rule that you supply, you will give a confidence score for that rule (it will be a number less than 100%). We will consider

those rules with confidence scores equal to or higher than 60% as **high quality** rules. Those that have confidence scores of between [30 and 60%) will be considered **medium quality** rules and those that are between [0 and 30%) will be considered **low quality rules**. For each relation (1-5 above), you have to identify either one **high quality** rule OR two **medium quality** rules.

2. Use automated rule mining techniques that you studied in class to **learn** new rules which predict (s, o) pairs that could be related via the relation *p,* for each relation 1-5 above. Again, one **high quality** rule or two **medium-quality** rules per predicate should be specified. Formulate SPARQL queries for inserting the new triples into the graph and specify these queries in the answer sheet provided.

3. Apply the rules which perform the best (according to confidence scores) for each predicate in Tasks 1 & 2, to the given graph, to predict the missing relations. For each predicate *p*, supply a list of the top 10 (s,o) pairs such that <s> <p> <o> represents a missing relation (ranked based on the confidence score).

4. Use knowledge graph embeddings to predict the missing (s,o) pairs for each predicate *p* above. Clearly state in your report what method you used to train these embeddings, how you constructed them and how you applied them to predict the missing relations.

Submission instructions:

1. For Tasks 1 and 2 above, use the sample submission template called "BMKG_Assignment_2_handcrafted_rules.txt" to record your SPARQL queries (rules) for each predicate along with the confidence scores for each rule.
2. For Tasks 3 and 4 above, use the submission template called "BMKG_Assignment_2_predictions.xlsx" to record your top 10 (subject,object) pairs with highest confidence scores after applying your rules.
3. You will also submit a **written report** (**max 4 pages** excluding the front title page) in PDF format called "BMKG - Assignment 2 - Group *your number here*.pdf" describing:
    a. for Task 1, your thought processes and motivation for crafting each rule,
    b. for Task 2, your motivations for selecting the methods you did for learning the new rules, your hypotheses about how well they will perform, and your motivations about why you have these hypotheses,
    c. for Task 3, explain why applying the rules to predict the relations is necessary and discuss the performance of your predictions, what are the limitations? How could they be improved?
    d. for Task 4, clearly state in your report what method you used to train your KG embeddings, how you constructed them, and how you applied them to predict the missing relations. Again, discuss the performance of your predictions, what are the limitations? How could they be improved?

4. **One person** from your group will submit a **SINGLE .zip archive** to Student Portal, containing the following three files:
    a. “BMKG_Assignment_2_handcrafted_rules.txt”
    b. “BMKG_Assignment_2_predictions.xlsx”
    c. “BMKG - Assignment 2 - Group *your number here*.pdf”
5. **Very important:**
    a. Please be very careful to correctly paste your SPARQL queries in the answer sheet for Tasks 1 and 2 (“BMKG_Assignment_2_handcrafted_rules.txt”). We will copy-paste these to test your rules, so if there are any typos or mistakes in the syntax or spelling - we will not be able to test them and you will not receive the marks for that query.
    b. Similarly, please carefully paste the correct URIs for the subjects and objects in the answer sheet for Tasks 3 and 4 (“BMKG_Assignment_2_predictions.xlsx”), otherwise we cannot assess your work and you cannot receive the marks.
    c. In your report, it is not sufficient to just state the steps you took in the assignment. It is also important to emphasise in your own words **why** each step is necessary or beneficial for the final solution. Please try to state each choice and motivation very clearly throughout the report. Please also state clearly your **group number** on the front page of your report.

Tips and potentially useful resources:

1. A list of properties (relations) that can be attributed to subjects of type <https://www.wikidata.org/wiki/Q5> (Human) in the Wikidata ontology: https://www.wikidata.org/wiki/Wikidata:List_of_properties/human
2. **Tip:** it can be helpful to **explore** the example graph to see what kind of information is inside (both properties and entities) to aid you in deciding what rules to craft in Task 1. You can explore the graph in whichever you choose. One way is to import it into an RDF visualisation tool of your choice (e.g. GraphDB’s visualisation features). Another way is to use SPARQL queries to list the unique properties and entities in the graph.
3. Lecture and lab slides for “Knowledge Graph Completion”. These materials will show how to construct handcrafted rules, learn new rules automatically, briefly how to apply KG embeddings, and how to calculate confidence scores.
4. Survey research paper published in the Semantic Web journal, which indexes and discusses prominent and useful information about knowledge graph completion methodologies: http://www.semantic-web-journal.net/system/files/swj1167.pdf and https://persagen.com/files/misc/Wang2017Knowledge.pdf

Contact:

Prof. Michel Dumontier: michel.dumontier@maastrichtuniversity.nl

Dr. Kody Moodley: kody.moodley@maastrichtuniversity.nl

Dr. Remzi Celebi: remzi.celebi@maastrichtuniversity.nl