DSCI 558: Building Knowledge Graphs

Homework 3: Entity Resolution & Knowledge Representation

Released: Feb 7th, 2021

Due: Feb 14th, 2021 @ 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

In this homework, you will perform record linkage between two different datasets, implement and test two blocking techniques, and then represent some of your data using RDF. The record linkage and blocking tasks will be done using The Record Linkage ToolKit (RLTK), an open-source record linkage platform. You will use RDFLib, a Python library for working with RDF, for the task of knowledge representation. We provide a python notebook (ER_KR.ipynb), which contains instructions, code and descriptions on how to use the tools we mentioned.

Task 1: ER (7 points)

In this task, you are given a pair of datasets in the "The 784 Data Sets for EM" from Magellan project (https://sites.google.com/site/anhaidgroup/useful-stuff/data). Your dataset assignment is listed in the "Dataset_Assignment" file. Your goal is to match records from these 2 datasets using record linkage methods. This means you need to figure out which pairs of entities in the two datasets are referring to the same entity.

Your given pair of datasets contain several attributes. For the task of linking, we are interested in the fields that are present in both datasets.

For the ER task, you will need to download 3 different files:

- Input Table A: contains the first dataset
- Input Table B: contains the second dataset
- Labeled Data L: contains a set of labeled data (usually a portion of the Cartesian product A x B). The final column in L indicates if a pair of entities refer to the same entity (1 is True and 0 is False)

Before moving to this section's subtask, familiarize yourself with RLTK by following the provided example (example.zip) and the online **examples**:

- https://rltk.readthedocs.io/en/latest/step_by_step.html
- https://rltk.readthedocs.io/en/latest/real_world_example.html

Task 1.1 (2 points)

Select at least 3 attributes that you think are the most useful for record linkage task. Then:

- Construct RLTK datasets with the selected fields.
    - Note: You can transform the data if needed (For example, if you have phone numbers, you can preprocess the phone numbers, extract the area code and then use it.)
    - See https://rltk.readthedocs.io/en/latest/step_by_step.html#Construct-RLTK-datasets for details on how to create RLTK datasets with data transformations.
- Explain your choices in the report.

Task 1.2 (1 points)

In this task, you will use RLTK to implement a blocking technique and evaluate their effectiveness.
(https://rltk.readthedocs.io/en/latest/step_by_step.html#Blocking)

- Design a blocking technique to reduce the number of pairs used in your record linkage task. Explain your design in the report.
- Evaluate the performance of your blocking technique using *reduction ratio* and *pairs completeness*. As most entity pairs are unlabeled, you only need calculate the *pairs completeness* within the Labeled Data L.
- Output your blocked data to a csv file with no header (Firstname_Lastname_hw03_blocked.csv). Each row should represent two values: id of entity in Table A and id of entity in Table B. (The id columns are the first columns in the datasets)

Note:

- The output blocked data should contain all pairs of entities that you still need to compare after the blocking.

Task 1.3 (4 points)

In this task, you will use RLTK to implement a record linkage method for your blocked pairs. For each selected field:

- Analyze the given data and choose the string similarities that you think are appropriate.
- implement a method that computes the field similarity between the records for the 2 datasets. Explain your choices in the report.
- Design a scoring function to combine your field similarities. Explain your choices of weights in the scoring function in the report.
- Implement a method for record linkage using your scoring function.
- Run your record linkage method on your blocked data (Task 1.2). Export your prediction to an csv file (Firstname_Lastname_hw03_el.csv) with no header.
- Run your record linkage method on the pairs in the Labeled Data L. Export your prediction to an csv file (Firstname_Lastname_hw03_el_labeled.csv).
- Evaluate your method on the Labeled Data L using RLTK. Report the precision, recall and F1-score in your report. If your performance is low, explain the reasons in your report.

Note:

- Each row in your two output csv files should contain three values: id of entity in Table A, id of entity in Table B and the prediction.
- For prediction, 1 indicates that the two entities are the same and 0 denotes that they are different.
- The number of rows in these output files should be the same with your output blocked data

Task 2: KR (3 points)

In this task, you will represent the data (after linking) using RDF. The ontology (vocabulary/schema) you will use is **schema.org**. As this ontology may not include all necessary classes and properties to model your data, you will need to extend the ontology with classes that you define on your own.

Task 2.1 (1 point)

Describe in the report the model you will use to generate the RDF data to describe the merged entry with **all available attributes** (not only your three selected attributes) from the two sources you have matched.
Use the appropriate classes and properties from **schema.org**. Define your own if you could not locate any suitable one in **schema.org**. Using the example in model.ttl to create your own model file (Firstname_Lastname_hw03_model.ttl).

Task 2.2 (1 point)

Implement a python program that uses the data from the 2 datasets (from Task 1) and your output file (Firstname_Lastname_hw03_el.csv). The program should convert the combined data to RDF triples (in turtle format, ttl) using the model you defined in Task 3.1, the generated file should be named Firstname_Lastname_hw03_triples.ttl.

Notes:

- If there are mismatches between field values of the two datasets, you can choose values from either dataset to create your triples.
- See the attached notebook for an example of how to create and generate RDF graph (triples) in ttl format (syntax).

Task 2.3 (1 point)

Choose an entity in your combined dataset and visualize the triple data in your report. Use the online tool at http://www.ldf.fi/service/rdf-grapher to visualize the triples in a graph.

Submission Instructions

You must submit (via Blackboard) the following files/folders in a single .zip archive named Firstname_Lastname_hw03.zip:

- Firstname_Lastname_hw03_report.pdf: pdf file with your answers to Tasks 1 and 2
- Firstname_Lastname_hw03_blocked.csv: as described in Task 1.2

- Firstname_Lastname_hw03_el.csv: as described in Task 1.3
- Firstname_Lastname_hw03_el_labeled.csv: as described in Task 1.3
- Firstname_Lastname_hw03_model.ttl: as described in Task 2.1
- Firstname_Lastname_hw03_triples.ttl: as described in Task 2.2
- source: This folder includes all the additional code you wrote to accomplish the tasks