AnHai's Group

Search this site

Home
People
Research
Publications

Current Projects
Data Cleaning & Integration
Data Science
Magellan
py_entitymatching
py_stringsimjoin
py_stringmatching
activelearn
py_labeler

Useful Stuff
Data Repository
How to Create Packages

Past Projects
Knowledge Bases/Graphs
Crowdsourcing
Schema/Ontology Matching
Silicon Valley Time Off

Courses
CS 638 DS
Fall 2016
CS 838 DS
Spring 2018
CS 564

Outreach & Funding
Walmart Labs
NSF IIS-Medium 2016

Miscellaneous
Wisconsin DB Group

Courses > CS 838 Spring 2018 (Also Known as CS 839) > Project Description for CS 839 Spring 2018 >

CS 839 Spring 2018, Project Stage 4

In this project stage you will combine the two tables (and optionally add any other table/data that you want). Then you will do some data analysis on the integrated table, to infer insights. This analysis is something of your own choosing. But it must involve one of the key techniques that we will cover in the class: classification, clustering, correlation discovery, anomaly detection, or OLAP-style exploration.

What to do?

1) Recall that in the previous project stage you have performed entity matching between two tables A and B. Specifically, you have trained a matcher M. If you haven't already, now you need to apply matcher M to the set of candidate tuple pairs obtained after blocking on A and B, to obtain the set of all matches between A and B.

2) Next, you should create the schema of a table E, which is the target table into which you will merge the two tables A and B. If tables A and B have identical schema, then table E has the same schema. Otherwise, the schema of table E will be the union of the schemas of tables A and B.

3) Next, you will write a Python script to combine the tuples of tables A and B to create the tuples for table E. Note that we have discussed such a step several times in the class. This Python script will use the matches that you have created between the entities in Tables A and B. It will also have to decide on how to merge data between the two tables (using for example a set of merging rules).

4) Next, you execute this Python script to populate the table E. Note: In this step, if you want to add more data, such as combining a table D with tables A and B to create table E, that is fine too.

5) Finally, you perform some data analysis on Table E to infer insights. We have discussed this step in the class.

What to submit?

Submit the following on your group's website:

- a directory that contains
  - a CSV file storing Table E

    - the set of matches between Tables A and B (these matches should be stored in a single file)
    - the Python script that you used to merge the two tables A and B
- a pdf file that discusses the following:
    - how did you combine the two tables A and B to obtain E? Did you add any other table? When you did the combination, did you run into any issues? Discuss the combination process in detail, e.g., when you merge tuples, what are the merging functions (such as to merge two age values, always select the age value from the tuple from Table A, unless this value is missing in which case we select the value from the tuple in Table B).
    - statistics on Table E: specifically, what is the schema of Table E, how many tuples are in Table E? Give at least four sample tuples from Table E.
    - append the code of the Python script (that merges the tables) to the end of this pdf file.
    -
    - What was the data analysis task that you wanted to do? (Example: we wanted to know if we can use the rest of the attributes to accurately predict the value of the attribute loan_repaid.) For that task, describe in detail the data analysis process that you went through.
    - Give any accuracy numbers that you have obtained (such as precision and recall for your classification scheme).
    - What did you learn/conclude from your data analysis? Were there any problems with the analysis process and with the data?
    - If you have more time, what would you propose you can do next?

Comments

You do not have permission to add comments.