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 2

In this project stage your team will select two Web data sources, then extract data from the sources. Subsequent stages will use the output of this stage.

Requirements

- You must select two Web data sources from which structured data can be extracted by using the rule-based wrapper construction method discussed in the class. These two data sources must contain information about a set of overlapping entities, such as books, movies, cars, etc. This is because later we will have to perform entity matching as a class project stage, and we need the two sources to have overlapping entities, so that we can match between the two sources, to find data that refer to the same real-world entities.
- Each of the above two sources should contain a reasonable amount of data, and the two sources should have a reasonable amount of overlapping entities. For example, suppose we extract a relational table A from the first source where each tuple describes a person, and suppose we extract a similar table B from the second source. Then each table should have at least 3000 tuples, and they should share at least 100 persons (you can only eyeball the data for this latter requirement, and that is sufficient).
- Then extract data from these two sources to form two tables A and B (one from each source). The two tables should have the same schema, and each tuple in each table must describe a single entity (all of the same type). For example, if the entity type is person, then each tuple describes a person, and a possible table schema can be A(name, city, state, zip, phone) (and the same schema for Table B).
  Each table must have at least 3000 tuples and be in CSV format.

What to Submit?

On your team's project page, by the deadline:

- **Provide a link to a DATA directory** that stores both tables A and B. They should be stored in two files, each file storing a table in CSV format. We should be able to browse these files to examine the tables. There should be a README file in the same directory that explains the tables (e.g., the meaning of the attributes) and lists the number of tuples per table.

- **Provide a link to a CODE directory** that stores all of your code (this directory must also be browsable).
- **Provide a link to a pdf file** that describes the following:
    - a description of the two Web data sources that you have selected. Recall that you are supposed to select two Web data sources from which you can extract structured data.
    - a description of how you have extracted structured data from the two data sources.
    - describe the type of entity you extract, briefly describe the two tables, list the number of tuples per table.
    - the names of open-source tools you have used in this project stage and a brief description of what they do.

Comments

Commenting disabled due to a network error. Please reload the page.

You do not have permission to add comments.

Sign in | Recent Site Activity | Report Abuse | Print Page | Powered By Google Sites