CLASS PROJECT, CS 784, FALL 2015

Introduction

The goal of this project is to get your hand "dirty" as a data scientist (and to practice certain materials taught in the class). After finishing the project, you will gain a much better appreciation for working with "data in the wild", a better understanding of what it means to work as a data scientist, a deeper understanding of the class materials, and a glimpse into some research efforts in data science (and be involved in such research efforts).

Specifically, in this project, you will select two Web sources, crawl to retrieve HTML data, perform information extraction to convert the HTML data into two relational tables (that describe entities such as persons, products, books, movies, papers, etc.). Next, you will use Magellan, a data matching system developed here at Wisconsin, to match the two tables (i.e., find tuples that refer to the same real-world entity). Finally, you will propose and implement an extension to Magellan (this last part is conditional, depending on whether we will still have enough time at the end of the semester).

This project helps you gain more skills in crawling Web pages, performing information extraction to convert Web pages into structured data, and using a data matching system. You will also gain an understanding for the entire data matching process, for data integration work performed by analysts and developers, and more generally for solving semantics intensive data management tasks.

Prerequisites

You should have some understanding of data extraction and matching, such as gained in CS 784, and have a basic working knowledge of Python, or be willing to learn quickly.

Project Steps

1) Form teams of 1-2 persons (due: Wed Sept 23)

Deliverables for this step:

+ You will enter information about your team into a page. More details later.

2) Obtain two relational tables A and B (two weeks, due: Wed Oct 7)

+ select two Web sites that list data that you can convert later into two relational tables. Examples of such data include products, employees, researchers, papers, movies, music albums, etc. Examples of such Web sites include amazon.com, walmart.com, DBLP, Google Scholar, IMDB, etc.
+ write two crawlers to crawl the Web sites to obtain HTML data (this data will be made public later, so do not select Web sites with sensitive data). From each Web site you will obtain a set of HTML pages.
+ decide which attributes you will extract from HTML pages, then write scripts to extract those attributes. At the end of this step, you should have converted HTML data from each Web site into a relational table. We will refer to these two tables as A and B. Each table must have at least 3,000 tuples. The tables must be in the format listed at the end of this page.
+ when deciding which attributes to extract, make sure to extract enough attributes so that later you can reliably match the tuples across tables A and B.
+ some Web sites will allow you to retrieve data in XML format. You are supposed to crawl, retrieve HTML data, and then write scripts to extract structured data from it. So pls ignore XML data if any is available.

Deliverables for this step:

+ your team will set up a Web page that provides links to HTML data and the two tables.
+ HTML data of each Web site must be listed within a directory for that web site.
+ the two tables are stored in two files named tableA.csv and tableB.csv
+ you will send me the URL pointing to your team's Web page.

3) Perform blocking on the two tables A and B (two weeks, due: Wed Oct 21)

+ write code to do blocking; this takes the two tables A and B and outputs a table C of tuple pairs judged likely to match. Table C must be in the format described at the end of this page.
+ the above code will most likely use a set of rules. We can talk more about this in the class.

Deliverables for this step:

+ code to do blocking
+ a file called blocking-explanation.txt that explains in plain English the blocking rules or methods that you use
+ table C

4) Create golden data and match tables A and B using Magellan (three weeks, due Wed Nov 11)
+ randomly sample from table C a table G of at least 300 tuples
+ manually label the tuples in G
+ table G must be in a format understandable to Magellan, as specified at the end of this page
+ match tables A and B using Magellan, using table G to evaluate the matching results
+ when matching your goal is to keep precision (P) at least 90% while maximizing the recall (R)

Deliverables for this step:
+ table G (with manually created labels)
+ the set of matching rules that you have created
+ precision and recall that you have managed to achieve on the golden data
+ a log of actions you have taken
+ estimated number of hours spent reaching that precision and recall
+ description of your matching experience (using Magellan) and what you think is good/bad and can be improved for Magellan (you will write and submit a project report that contains this).

5) Select one thing that you want to work on to extend EMS (4.5 weeks, due Fri Dec 11)
Deliverables for this step:
+ a proposal on what you want to work on next
+ code for the revised system and other misc stuff (to be decided)

What We Will Provide
+ We will supply two sample tables A and B, a set of tuple pairs judged likely to match C, golden data, our rules, and precision and recall that we have reached on the above two tables.
+ We will also supply the code of Magellan in Python.

Required Format for Tables A and B in Step 2
To be described later.

Required Format for Table C in Step 3
To be described later.

Required Format for Table G in Step 4
To be described later.
