Syllabus

General Info

Course CS229 Data Wrangling and Management
Instructor Charilaos Skiadas (skiadas at hanover dot edu)
Term Fall 2016-2017
Office SCH 121C / LYN 108
Office Hours MWF 10am-11am, and by appointment
Book online ACM resources[1]
Websites for notes[2]
Class times MWF 12:00pm-1:10pm in LYN120A

Course Description

Data Wrangling and Management is a very broad subject, but at its heart the course aims to prepare students for the modern demands on the processing and handling of data. We will cover a variety of topics, for example:

1. How is data transmitted between the requester and the provider? This includes questions of transmission protocols, authorization, as well as file format and structure.
2. How is data to be stored? This includes the study of databases as well as distributed storage techniques like Hadoop.
3. What are typical processing steps that this data undergoes? This would include use of Python scripts as well as web scrapers, database queries, and other tools.
4. How do we turn that processed data into a deliverable “data product” that a client can use? This would include for example creating a web server and documenting the API to use.

We will touch on all of these topics, and you will get considerable practice in these areas by implementing a data product as a culminating course project.

There are many important topics that are related but that we will not cover. For instance, there is a lot of work in data mining and machine learning, to develop algorithms for extracting information from data. Also the question of visualizing information is a whole topic on its own. These aspects all deserve their own courses.

[1] http://learning.acm.org
[2] https://skiadas.github.io/DataWranglingCourse/site/

Course Components

Reading Assignments

In the class schedule page[3] you will find, for each class day, a list of links to reading assignments. Your homework will require you to have a solid understanding of the material covered there, so I strongly encourage you not to get behind.

Class Attendance

You are expected to attend every class meeting. You are only allowed to miss 3 classes without excuse. From that point on, every unexcused absence will result in a reduction of your final score by one percentage point, up to a total of 5 points. Excused absences should be arranged in advance, and backed by appropriate documentation. Emergencies will be dealt with on an individual basis. There are very few reasons that would qualify as an excuse for an absence.

Lab Assignments

There will be lab assignments roughly once or twice each week. You are expected to work on these assignments on your own, but you are welcome to ask me questions, and you are welcome to discuss general topics related to the assignment with your classmates. We will typically start these assignments in class, but you will be expected to complete them outside of class.

Exams

There will be one midterm, on Friday, October 14th, and a final/2nd midterm during finals week. **You have to be here for the exams**. If you have conflicts with these days, let me know as soon as possible. Do not plan your vacation before you are aware of the finals schedule. In terms of your final grade, the exams you did better on will weigh more.

Project

For a large part of the course you will be engaged in a collaborative project with a classmate (groups of 2 only please). In that project you will create a “data product”. While the project can vary a lot and is up to you to decide, it would need to cover the following:

1. It would collect data from one or more sources.
2. It would process that data in some way to produce new data.

[3] skiadas.github.io/DataWranglingCourse/site/schedule.html

3. It would provide access to that data via an appropriate interface, typically a web API.

The deliverables for the project would be:

1. A GitHub repository of the project that would include the code that delivers the data product.
2. Documentation of the data product and its API, via the automatically-created GitHub web pages.
3. A running version of the data product on vault.

This may sound overwhelming at first glance, but along the term we will develop the necessary tools to do this.

Getting Help

- You should never hesitate to ask me questions. I will never think any less of anyone for asking a question. Stop by my office hours or just email me your question, which has the great benefit of forcing you to write it down in clear terms, which often helps you understand it better.
- You are allowed, and in fact encouraged, to work together and help each other regarding the class material, as well as the topics related to the lab assignments. However, you may NOT directly help each other with your lab assignments.

Grading

Your final grade depends on class attendance, homework, midterms and the final, as follows:

Component	Percent
Attendance	5%
Assignments	35%
Project	30%
Worst Exam	10%
Best Exam	20%

This gives a number up to 100, which is then converted to a letter grade based roughly on the following correspondence:

Letter grade	Percentage Range
A, A-	90%-100%
B+, B, B-	80%-90%
C+, C, C-	70%-80%

Letter grade	Percentage Range
D+, D, D-	60%-70%
F	0%-60%