Syllabus

General Info

Course CS229 Data Wrangling and Management and CS328 Data Management and Web Services
Instructor Charilaos Skiadas (skiadas at hanover dot edu)
Term Fall 2018-2019
Office SCH 111 / LYN 108
Office Hours MWF 9:30am-10:30am in SCH 111, and by appointment
Book online ACM resources[1]
Websites for notes[2]
Class times MWF 10:40am-11:50am in LYN120A

Course Description

This course spans a broad variety of topics that all have one thing in common: The need to work with data. The course is meant as an introduction to the challenges and techniques used to process and manage information. The course is offered at two levels:

- CS 229: Data Wrangling and Management
- CS 328: Data Management and Web Services

The majority of the content is common to both levels. The programming assignments differ considerably. Students in the 3xx course will focus on creating a web service that would *provide* data services to consumers, while students in the 2xx will focus more on obtaining and processing data from various sources.

Some of the topics we will consider are:

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 and the issues solved by or caused by distributed storage.
3. How is data collected? We will discuss database queries, web scraping as well as using web services.
4. Students in the 3xx course will also examine in more detail how to provide a "data product" clients can use, in the form of a web service and API.

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

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.

Textbook

A large part of the material will be covered in my course notes on the website, but there are many resources linked from the notes. We will be covering a variety of topics that no single textbook could hope to cover. We will therefore be using a variety of linked resources, some freely available (e.g. the Python documentation) and some not.

I am asking all students to join the ACM (Association for Computing Machinery) instead of buying a textbook. The ACM delivers resources that advance computing as a science and a profession, and this includes the ACM Learning Center[3] which provides you access to hundreds of online books and videos related to computing. The annual student fee for the ACM is around $20. I encourage you to continue your membership every year and to take advantage of the opportunities and resources it offers.

Course Components

Reading Assignments

In the class schedule page[4] 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 Participation

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.

There will also be numerous in-class group activities that you will be expected to participate in.

Lab Assignments

There will be lab assignments roughly once a week. You are expected to work on these assignments on your own, but you are welcome to ask me questions, and you are

[3] https://learning.acm.org/
[4] skiadas.github.io/DataWranglingCourse/site/schedule.html

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). The project differs depending on the class. In both instances you are expected to maintain you code in a GitHub repository owned by one of the people in the group.

CS 229 Project The goal of the project is to demonstrate the ability to collect data from varying sources and formats, and successfully merge them together. In this project you will need to produce a script or set of scripts that:

1. Collects data from at least two different sources, in real time.
2. Combines this data in a suitable way based on common elements.
3. Processes this data in some way to produce new information.
4. Exports this new information in some way (e.g. a JSON file or a spreadsheet).

CS 328 The goal of the project is to demonstrate the ability to serve data to clients and response to client requests in various forms. In this project you will need to produce a web service that:

1. Offers a Web API that clients can use to request data.
2. Either collects data from some source as needed, or uses locally stored data.
3. Includes at least one form of handling dynamically generated queries.
4. Includes documentation about how to use the service.

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:

CS 229

Component	Percent
Participation	10%
Assignments	30%
Project	20%
Worst Exam	15%
Best Exam	25%

CS 328

Component	Percent
Participation	10%
Assignments	25%
Project	30%
Worst Exam	15%
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%
D+, D, D-	60%-70%
F	0%-60%