Data Analysis and Visualization Gr.1 and Gr.2
Wednesday 9:30-12:00 @ KMB304(Group1)
Friday 14:00-17:00 @ KMB204(Group2)
Instructor: Assist.Prof.Dr. Alper Yilmaz
Email: alyilmaz@yildiz.edu.tr
Office Hours: by appointment
http://yarbis.yildiz.edu.tr/alyilmaz/course/viewCourse/id/7095

Course Goals

In this course, you will:

1. Be familiar with Linux commandline and use terminal to execute tasks.
2. Learn and practice the Unix philosophy of chaining small tools via "pipe" in order to accomplish larger and complex tasks.
3. Be introduced to concept of databases and have chance to use light version of SQL and learn modern types of databases.
4. Be able to draw graphs by scripting via gnuplot software.

Course Materials

There's a companion textbook written by the instructor and its PDF version will be provided during the course. The instructor will go through the topics, examples and questions from the book. You are expected to listen to course and take notes during lectures. Also, you are strongly encouraged to study the examples and questions on your own. For programming lectures, reading and studying the material is not enough, the only way to learn and grasp the concepts is **TO PRACTICE IN TERMINAL**. Please go over the examples in book and you're more than welcome to search online for more examples or exercises.

Important announcements and necessary lecture materials will be provided at YARBIS page of this lecture.

Grading

Your grade will come from the following sources:

- Midterm: 35%
- Final: 40%
- Quiz: 10%
- Code Academy: 5%
- Course Material Improvement: 5%
- Attendance: 5%
- Bonus: Visualization challenge (5 points)

Midterm and *final* are performed in computer lab and students are asked to figure out correct commands in terminal to achieve desired output.

Although the instructor had best intentions to help the students, last year the students abused these best intentions and attempted mass cheating. Thus, the format, number and style of *quizzes* is not determined yet. It will be determined and announced when number of students is finalized.

As mentioned above, best way to learn terminal is to practice. There's a website where students can follow guided practices. Thus students are expected to sign up at Code Academy website and complete Learn the Command Line and Learn SQL courses online. You need to provide your username to instructor so that he can form an online classroom to track your progress. There are paid lessons, projects and quizzes in this website but we'll be going thru only **FREE** lessons.

Course material improvement will be accomplished via Github. The students are expected to sign up for an account and then edit course book in a collaborative manner. Improvements such as spelling error corrections are welcome but it will get you low scores. Higher scores can be obtained if you can add explanations, samples, questions that can help others understand the topic better.

As for *attendance*, if you attend all lectures or miss only one lecture then you'll get 5 points for attendance. For every 1-2 lectures missed you'll lose 1 point.

Visualization challenge is optional and it requires submission of a drawing which will summarize a command or chain of commands **visually**.

Communication

I'm trying to respond emails as quickly as possible. If you don't get a response within 1-2 days please don't hesitate to send a reminder email.

The changes pertaining to exam date, time and assignment due dates should be decided in class after discussing with everybody. Please don't ask for changes individually, otherwise notification of whole class becomes a hassle.

Schedule

Below is the tentative schedule for the course. Depending on the speed we go through topics there might be shifts in the schedule. ***Some weeks you might be asked to study following week's lecture so that we can go through it faster in class***. For each week, first date is for Group 1 (Wednesdays) and second date is for Group 2 (Fridays).

September 21/23. Introduction

The instructor will demonstrate the power of Linux command line and point out fundamental differences between Linux OS and Windows OS. Bear with the instructor if he complains about Windows a lot ;)

September 28/30. Thinks to consider and commands to navigate and manipulate file system

You'll learn about important concepts that will prevent you from being stuck while learning basics of command line. Also, you'll learn commands to navigate (change directories) and manipulate (create, copy, move and delete file/folders) the file system.

Assignment for next week: Let's familiarize ourselves with GitHub platform. Please make *useful* edits via GitHub for the book chapter 1 - Linux Operating System due **NEXT WEEK**

October 5/7. Viewing files

Let's see what are the contents of a file. Very large file? No, problem. We can view top or bottom rows of a file instantly no matter how big it is. Viewing small portion of a file quickly will help us having an idea about its content.

Assignment: Code Academy - Learn Command Line - Unit 1 is **DUE** this week

Assignment: GitHub pull requests for Chapter 1 are **DUE** this week

***October 12/14.* Generating, modifying, sorting and counting content**

Generating sequence of numbers even in shuffled form is possible with *seq* and *shuf* commands. *tr* command is a tiny but very significant command capable of replacing characters. *sort* and *uniq* commands, when combined, can perform task of counting even for very long lists.

Assignment: Code Academy - Learn Command Line - Unit 2 is **DUE** this week

Quiz 1 - Topics To Be Determined

***October 19/21.* Working with two inputs, whether it's a list or a file**

comm can find intersection (or other set operations) between to sets (aka lists) and *join* can merge to different files by using a common column. *paste* performs simple merge (side by side)

Assignment: Code Academy - Learn Command Line - Unit 3 is **DUE** this week

***October 26/28.* Find the line that contains a pattern**

grep is an essential tool to reveal lines that contain desired pattern. The pattern can be a simple text or a regular expression

Study for next week: Please study *sed* and *awk* sections from our book for next week. You need to have an idea about these commands in order to follow next week's lecture.

***November 2/4.* sed and awk**

sed and *awk* are very important two commands (actually awk is a simple programming language) to manipulate text and filter text, respectively. *sed* is capable performing edits within a stream of data. *awk* can be used to view rows that meet certain criteria. Also, *awk* allows user to have manipulate the output.

(November 9/11). Midterm

***November 16/18.* Loops**

Terminal can actually be programmed. It's possible to use *for* or *while* loops that are common in most programming languages in bash environment as well. After learning loops, it's trivial to process thousands of files with single command.

Assignment: GitHub pull requests for Chapter 2 are **DUE** this week

***November 23/25.* Databases and introduction to SQL**

You'll get an idea about the meaning of databases. You'll be explained the need for normalization of complex data into separate tables. Finally, you'll be introduced to structured query language *SQL* and softwares that use SQL to query their data. *MySQL* is more common but difficult to setup thus you'll be learning *Sqlite3*.

***November 30/December 2.* Create table, insert data into tables**

Let's understand the concept of *schema* and see that we have to strictly follow the schema when inserting data into table.

Quiz 2 - Topics To Be Determined

Assignment: GitHub pull requests for Chapter 3 are **DUE** this week

***December 7/9.* Querying tables**

Creating a table and inserting data into a table was cumbersome, after that it's payback time. SQL is robust in querying tables. Filtering and sorting in SQL are like a breeze. *Aggregate* functions adds even more power into queries

Assignment: Code Academy - Learn SQL - Unit 1 is **DUE** this week

***December 14/16.* Querying multiple tables and beyond SQL (NoSQL)**

Joining tables is much easier than its terminal counterpart *join* command. The *schema* was such a pain when dealing with tables. Thus, there's new trend of database technologies, loosely named NoSQL, which lack a schema. An example from graph database will blow your mind.

Quiz 3 - Topics To Be Determined

***December 21/23.* Gnuplot - part 1**

If you had taken *MATLAB* courses, you'd be familiar with the concept of using scripts to generate graphs. *Gnuplot* is very efficient in doing so. Just by couple of lines of code, data from a large file can be converted into a graph.

Assignment: Code Academy - Learn SQL - Units 2 **AND** 3 are **DUE** this week

***December 28/30.* Gnuplot - part 2**

More examples regarding *Gnuplot*

Assignment: GitHub pull requests for Chapters 3,4 and 5 are **DUE** this week

Final lecture: A make up quiz will take place in order to replace lowest scoring quiz score

Acknowledgments

This syllabus was adapted from Benjamin Schmidt and Andrew Goldstone.