Syllabus
Dealing with Data
TECH-GB 2346 30, Spring 2019

2/13/2020 thru 5/7/2020

Course Information

- When: Thursdays, 6:00pm - 9:00pm
- Where: KMEC Room 4-80

Professor information

- Prof. Alex Siegman
- Email: alex.siegman@nyu.edu
- Website: https://www.siegmanAI.com
- Office: KMC 8-171, Desk F
- Office Hours: By Request

Teaching Assistant

- TF: Karan Gupta
- Email: karan.gupta@nyu.edu

Important Information

This course is a hands-on, interactive lab environment to learn two data-processing focused programming languages (Python, SQL). Students must bring a laptop to every class.

Course Description

A key business differentiator today appears to be implementing machine learning and data science to deliver business value. However, these applications are meaningless if the underlying data is not there to be applied.

Imagine having the world's greatest vocalist without a single musical note to sing; or, the world's most relaxing bathing experience, but no water. These scenarios illustrate machine learning without data to power it.

This course aims to provide interested students with knowledge of the "plumbing" that provides data applications such as machine learning algorithms with its data "water".

We will explore the technologies used to construct data pipelines that Deal With Data by collecting raw data, transforming it into usable data, storing and making it usable for downstream business applications such as analytical reports, visual KPI dashboards and machine learning applications.

This course's exploration will include hands-on programming with Python and SQL in an interactive lab environment during class. We will also explore database technologies including relational databases, NoSQL databases, and other big data technologies. And, we will explore the data applications possible through the capabilities of Python's pandas library.

This course's exploration will NOT include assessing different data science approaches, explaining machine learning algorithms, designing data visualization with tools such as Tableau, the statistics behind data analytics, business intelligence reporting design, and programming in R. While these are all fantastic areas of study – they are covered in-depth elsewhere.

This course will equip you with three primary toolsets. First, you will have a hands-on understanding of how data pipelines transform raw data into usable data for advanced applications, such as KPI dashboards and machine learning, to better understand tech solution designs and delivery speed. Second, you will have a hands-on understanding of the Python programming language to again better understand technical colleagues, but even more – perhaps automate some of your routine tasks or analyze data. Third, you will have a hands-on understanding of SQL and working with databases to take advantage of democratized data when available at your firm

Prerequisites

This course will be most useful to students who have no formal programming training but a healthy passion for Excel and Excel formulas.

This course is NOT useful to Computer Science majors and/or those who already have experience with programming in VBA, Python, JavaScript, Java, C, C#, C++, SQL, or any other programming language.

While there are no formal prerequisites for this course, it is expected that students have first-hand experiences with digital consumer software such as mobile apps, websites, and multi-user collaboration tools such as Google Docs.

We also expect participants to have a basic understanding of fundamental technologies such as the internet, social networks, and databases and have experience with data applications of any kind, whether sophisticated statistics or basic math using spreadsheet tools such as Excel and Excel Pivot Tables.

Attendance

We expect students to attend all classes. If you plan on missing more than 2 class meetings during the semester, please consider taking the class at some other time.

Grading

Course participants will be evaluated based on class participation, a group project, team member ratings for your group project, and individual homework assignments.

All students are expected to follow the NYU Stern Student Code of Conduct and will be held to the commitments you made to that code upon enrolling at NYU Stern:

- 10% = In-class participation
- 10% = team member ratings for your group project
- 30% = Group Project (end of course deliverable)
- 50% = Individual homework assignments

Individual Homework Assignments

During the course, students will be asked to complete 7 short assignments to practice what we explored together in class. Some assignments will ask you to write code whereas others will ask you to explain technical concepts and their business applications.

Your two worst performing assignments will be dropped from consideration so your grade will be based on your 5 best assignments. Late submissions are accepted up to 1 week after the due date with grade penalties. Incomplete assignments will reduce your final grade.

Group Project

Construct a Python+SQL Data Pipeline to Power a Data Application.

During the course, students will learn what a data pipeline is, how to construct one using Python and SQL, and, using them to power simple data applications (e.g. data reporting tables, simple histograms).

Students will then apply what they have learned to produce their own data pipeline and simple business application in groups of 3 – 4 students.

The final deliverable will be composed of two parts: (1) a presentation to the class in our final session of your group’s data pipeline and business application; and, (2) the working code that is your data pipeline. Your group’s presentation of the data pipeline and business application, your peers’ evaluation of your group’s presentation, and the working code will inform your group project grade.

EVERY member of the group is required to present a meaningful portion (2 minutes or more) of the 10 minute presentation during the allotted time.

Each student will also be given a survey to evaluate their other group members' contributions to the project. What you learn in class is a foundation for learning more advanced techniques on your own. You are encouraged to explore different and new techniques using Python and SQL as part of your group project.

Class Topic Schedule (subject to change):

Class 1:

- Course Introduction
- Python, Part 1 – your first program, data types, and variables preview

Class 2:

- Python, Part 2 – variables, working with strings and booleans, and IF, ELIF, ELSE statements

Class 3:

- Python, Part 3 – sets, dictionaries, lists, and nested data structures
- Introduction to group projects

Class 4:

- Python, Part 4 – program flow control, while and for loops, and functions

Class 5:

- Python Part 5 – getting data into / out of text files, inspecting / cleaning data with Python

  Recommended Group Milestone #1 – select your public dataset(s) and define the data app using the recommended method

Class 6:

- Python, Part 6 – pattern matching and regular expressions, using Python with relational databases

  Recommend Group Milestone #2 – kickoff design and coding for your data app's get, clean, and transform stages

Class 7:

- Python Part 7 – using Python's Pandas

  Recommended Group Milestone #3 – kickoff design and coding for your data app's active stage

Class 8:

- Python Part 8 – using Python for data visualization

Recommended Group Milestone #4 – finish coding for your data pipeline and app

Class 9:

- SQL, Part 1 – relational database design, cont’d, loading data into SQLite Studio, and SELECT statements

Class 10:

- SQL Part 3 – sub-queries, with queries, case, group by aggregations, and joins

Recommended Group Milestone #5 – test run your data app and its pipeline

Class 11:

- Big Data technologies

Class 12:

- Group Presentations