CMU 95-865 UNSTRUCTURED DATA ANALYTICS (SPRING 2018 MINI-3 SECTION A3, 6 UNITS)

Instructor: George H. Chen (georgechen [at symbol] cmu.edu)

Time and location: (lectures) Mondays and Wednesdays, 10:30am-11:50am HBH 1202, (recitations) Fridays 1:30pm-2:20pm location TBA

TAs: Emaad Manzoor (emaad [at symbol] cmu.edu), Mallory Nobles (mnobles [at symbol] andrew.cmu.edu)

Office hours: (TAs) Mondays 2pm-4pm HBH 2007C, (George) Thursdays 2:30pm-3:30pm and by appointment HBH 2216

Course description: Companies, governments, and other organizations now collect massive amounts of data such as text, images, audio, and video. How do we turn this heterogeneous mess of data into actionable insights? A common problem is that we often do not know what structure underlies the data ahead of time, hence the data often being referred to as "unstructured". This course takes a practical approach to unstructured data analysis via a two-step approach:

(1) We first examine how to identify possible structure present in the data via visualization and other exploratory methods.
(2) Once we have clues for what structure is present in the data, we turn toward exploiting this structure to make predictions.

Many examples are given for how these methods help solve real problems faced by organizations. Along the way, we encounter many of the most popular methods in analyzing unstructured data, from modern classics in manifold learning, clustering, and topic modeling to some of the latest developments in deep neural networks for analyzing text, images, and time series. We will write lots of Python code and also work with Amazon Web Services (AWS) for cloud computing.

Learning objectives: By the end of the course, students are expected to have developed the following skills:

- Recall and discuss common methods for exploratory and predictive analysis of unstructured data
- Write Python code for exploratory and predictive data analysis that handles large datasets
- Work with cloud computing using Amazon Web Services (AWS)
- Apply unstructured data analysis techniques discussed in class to solve problems faced by governments and companies

Skills are assessed by homework assignments, a mid-mini quiz, and a final exam.

Prerequisites: Python coding experience. To assess Python proficiency, there is an initial homework assignment (HW0) that will be released on the first day of class and due at the beginning of the second class. This assignment is meant to be diagnostic: if you find it challenging, then you will very likely struggle to complete the other assignments in the course in a reasonable amount of time. Please talk to the instructor if you're not sure about whether you have the appropriate coding background for the course.

Instructional materials: There is no official textbook for the course. We will provide reading material as needed.

Homework: After the initial basic Python assignment (HW0), there are 3 homework assignments (HW1, HW2, HW3) that give hands-on experience with techniques discussed in class. All assignments involve coding in Python and working with sizable datasets (often large enough that for debugging purposes, you should subsample the data). We will be making use of standard Python machine learning libraries such as SCIKIT-LEARN and KERAS. Despite the four homework assignments being of varying difficulty, they are equally weighted. Assignments are submitted in Canvas.

Grading: Grades will be determined using the following weights:

Assignment	Percentage of grade
Homework	25%
Mid-mini quiz	35%
Final exam	40%

Letter grades are assigned using a curve.

Cheating and plagiarism: We encourage you to discuss homework problems with classmates. However, you must write up solutions to homework assignments on your own. At no time during the course should you have access to anyone else's code to any of the assignments including shared via instant messaging, email, Box, Dropbox, GitHub, Bitbucket, Amazon Web Services, etc. Do not use solutions from previous versions of the course. If part of your code or solutions uses an existing result (e.g., from a book, online resources), please cite your sources. For the mid-mini quiz and final exam, your answers must reflect your work alone. Penalties for cheating range from receiving a 0 on an assignment to failing the course. In extreme circumstances, the instructor may file a case against you recommending the termination of your CMU enrollment.

Additional course policies:

Late homework: You are allotted a total of two late days that you may use however you wish for the homework assignments. By using a late day, you get a 24-hour extension without penalty. For example:

- You could use the two late days on two different assignments to turn them in each 1 day (24 hours) late without penalty.
- You could use both late days on a single homework assignment to turn it in 2 days (48 hours) late without penalty.

Note that you do *not* get fractional late days, e.g., you cannot use 1/2 of a late day to get a 12-hour extension. We will keep track of how many late days you have left based on the submission times of your homework assignments on Canvas. Once you have exhausted your late days, work you submit late will not be accepted. This policy only applies to homework; the mid-mini quiz and final exam must be submitted on time to receive any credit.

Re-grade policy: If you want an assignment regraded, please write up a note detailing your request and submit it to the instructor. Note that the entire assignment will be regraded and it is possible that your score may be lowered. The course staff will make it clear by what date re-grades for a particular assignment are accepted until. Re-grade requests submitted late will not be processed.

Mobile phones/laptops: Please do not use phones and laptops in class.

Course schedule (subject to revision):

The course is roughly split into two halves. The first half is on exploratory data analysis in which given a dataset, we compute and visualize various aspects of it to try to understand its structure. The second half of the course turns toward making predictions once we have some idea of what structure underlies the data.

Date	Topic
Wed Jan 17	Course overview Exploratory analysis: Frequency analysis **HW0 out**
Mon Jan 22	Exploratory analysis: Co-occurrence analysis, scatter plots, correlation, causation **HW0 due; HW1 out**
Wed Jan 24	Exploratory analysis: Visualization of high-dimensional data, intro to clustering
Mon Jan 29	Exploratory analysis: Clustering
Wed Jan 31	Exploratory analysis: More clustering **HW1 due; HW2 out**
Mon Feb 5	Exploratory analysis: Topic modeling
Wed Feb 7	**Mid-mini quiz**
Mon Feb 12	Predictive analysis: Intro to categorization/classification
Wed Feb 14	Predictive analysis: Adaptive nearest neighbor methods **HW2 due; HW3 out**
Mon Feb 19	Predictive analysis: Intro to neural nets and deep learning
Wed Feb 21	Predictive analysis: Image analysis with convolutional neural nets
Mon Feb 26	Predictive analysis: Sentiment analysis with recurrent neural nets
Wed Feb 28	TBD **HW3 due**
Mar 5-9	**Final exam week**