CMU 94-775 UNSTRUCTURED DATA ANALYTICS FOR POLICY
(SPRING 2018 MINI-3 SECTION A4, 6 UNITS)

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

Time and location: (lectures) Tuesdays and Thursdays, 3pm-4:20pm HBH 1204, (recitations) Fridays 4:30pm-5:20pm HBH 1204

TAs: Dylan Fitzpatrick (djfitzpa [at symbol] cmu.edu), Runshan Fu (runshanf [at symbol] andrew.cmu.edu)

Office hours: TBA

Course webpage: www.andrew.cmu.edu/user/georgech/94-775/

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 provides a practical introduction to unstructured data analysis and is composed of three parts:

(1) Basic Python programming especially as it pertains to working with data
(2) Exploratory data analysis: identifying possible structure present in the data via visualization and other exploratory methods
(3) Predictive data analysis: 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. There is a final project in this course which must address a policy question.

How this course differs from 95-865 "Unstructured Data Analysis": 95-865 has Python programming as a prerequisite, emphasizes more of the technical skill development (assessed through two in-class exams involving coding), and does not have any sort of policy focus. On the other hand, 94-775 does not assume any Python experience and has a policy-focused final project instead of a final exam. 94-775 does not require cloud computing (part of 95-865 requires the use of Amazon Web Services). Despite these differences, there is heavy material overlap between 94-775 and 95-865.

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
- Apply unstructured data analysis techniques discussed in class to solve problems faced by governments and companies

Skills are assessed by homework assignments, a quiz, and a final project. The homework and quiz are meant to be done individually whereas the final project is done in groups.

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

Homework: There are 3 homework assignments that give hands-on experience with techniques discussed in class. Assignments involve coding in Python and are submitted via Canvas.

Grading: Grades will be determined using the following weights:

Assignment	Percentage of grade
HW1	8%
HW2	8%
HW3 (shorter than HW1 and HW2)	4%
Final project proposal	10%
Quiz	35%
Final project	35%

Letter grades are assigned using a curve. Note that HW3 is designed to be shorter than HW1 and HW2 so that you have more time to work on the final project.

Cheating and plagiarism: In short, the only part of this course that is meant to be a group effort is the final project. While you are welcome to discuss homework problems with classmates, 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 homework assignments including shared via instant messaging, email, Box, Dropbox, GitHub, Bitbucket, Amazon Web Services, etc. If part of your homework code or solutions uses an existing result (e.g., from a book, online resources), please cite your sources. For the quiz, 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 quiz and final project 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; ; see course webpage for most up-to-date calendar): The course is roughly split into three parts: (I) a crash course on Python especially as it pertains to data analysis, (II) basic exploratory analysis for identifying what structure might be in a dataset, and (III) basic predictive analysis for making predictions once we have some idea of what structure underlies data.

Date	Topic
Part I: Python for Data Analysis	
Tuesday Mar 20	Course introduction, Python crash course
Thursday Mar 22	Python crash course, continued **HW1 released**
Tuesday Mar 27	Python crash course, continued
Part II: Exploratory Data Analysis	
Thursday Mar 29	Basic text analysis **HW1 due, HW2 released**
Tuesday Apr 3	Basic text analysis, continued
Thursday Apr 5	Clustering
Tuesday Apr 10	Clustering, continued **HW2 due, final project proposals due**
Thursday Apr 12	**Quiz**
Part III: Predictive Data Analysis	
Tuesday Apr 17	Introduction to classification **HW3 released**
Thursday Apr 19	Classical classification methods
Tuesday Apr 24	Neural nets and deep learning **HW3 due**
Thursday Apr 26	Neural nets and deep learning, continued
Tuesday May 1	**Final project presentations**
Thursday May 3	**Final project presentations, continued** **Final project report due**