CMU 94-775 UNSTRUCTURED DATA ANALYTICS FOR POLICY
(SPRING 2022 MINI 4 SECTION A4, 6 UNITS; CROSSLISTED ALSO AS 94-475)

Instructor: George H. Chen (email: georgechen ♣ cmu.edu) — replace "♣" with an "at" symbol

Time and location: (lectures) Tuesdays and Thursdays, 1:25pm-2:45pm HBH 1202, (recitations) Fridays 4:40pm-6:00pm HBH 1202

TAs:

- Abhirup Panja (apanja ♣ andrew.cmu.edu)
- Yashasvi Madhukumar (ymadhuku ♣ andrew.cmu.edu)

Office hours: TBA (see course webpage/Canvas)

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 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. 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 emphasizes more of the technical skill development (assessed through two in-class exams involving coding), and does not have any sort of policy focus. 94-775 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 Google Colab). 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.

Prerequisites: If you are a Heinz student, then you must have completed 95-791 "Data Mining" *and* one of either 90-819 "Intermediate Programming with Python" or 95-888 "Data-Focused Python". If you are not a Heinz student and would like to take the course, please contact the instructor and clearly state what Python courses you have taken/what Python experience you have.

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	12%
HW2	12%
HW3 (shorter than HW1 and HW2)	6%
Final project proposal	10%
Quiz	30%*
Final project	30%

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.

**Students with the most instructor-endorsed posts on Piazza will get a bonus of up to 10 points on the quiz (so that it is possible to get 110 out of 100 points).*

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.

Course schedule (subject to revision; see course webpage for most up-to-date calendar): The course is roughly split into two parts. The first part 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 part of the course turns toward making predictions once we have some idea of what structure underlies the data.

- Part I: Exploratory data analysis
  - Week 1:
    - Lecture 1 (Tue Mar 15): Course overview, analyzing text using frequencies
    - Lecture 2 (Thur Mar 17): Basic text analysis demo, co-occurrence analysis
    - Recitation (Fri Mar 18): Python review
  - Week 2:
    - Lecture 3 (Tue Mar 22): Co-occurrence analysis (cont'd), visualizing high-dimensional data with PCA
    - Lecture 4 (Thur Mar 24): PCA (cont'd), manifold learning
    - **HW1 due Thur Mar 24, 11:59pm**
    - Recitation (Fri Mar 25): More on PCA, argsort
  - Week 3:
    - Lecture 5 (Tue Mar 29): Manifold learning (cont'd), clustering
    - Lecture 6 (Thur Mar 31): Clustering (cont'd) — k-means, GMMs

        - Recitation slot (Fri Apr 1): Lecture 7 – Clustering (cont'd) — interpreting GMMs, automatically selecting the number of clusters
        - **Required: sign-up to meet with TAs to discuss final project proposals during TA office hours**
    - Week 4:
        - Lecture 8 (Tue Apr 5): Topic modeling
        - **Final project proposal due Tue Apr 5, 11:59pm by email (1 email per group)**
        - No class on Thur Apr 7 & Fri Apr 8 (CMU Spring Carnival)
    - Week 5:
        - **HW2 due Mon Apr 11, 11:59pm** (note that we release HW solutions 2 days after the due date in case anyone is using 2 late days, so this due date is set up so that you get solutions before the quiz)
        - Lecture slot (Tue Apr 12): Quiz review session
        - Lecture slot (Thur Apr 14): **Quiz**
        - **No recitation (instead sign up to meet with TAs to discuss final project status)**
- Part II: Predictive data analysis:
    - Week 6:
        - Lecture 9 (Tue Apr 19): Introduction to predictive data analytics
        - Lecture 10 (Thur Apr 21): Introduction to neural nets and deep learning
        - Recitation (Fri Apr 22): More on prediction
    - Week 7:
        - Lecture 11 (Tue Apr 26): Image analysis with convolutional neural nets
        - **HW3 due Tue Apr 26, 11:59pm**
        - Lecture 12 (Thur Apr 28): Additional deep learning topics; course wrap-up
        - Recitation (Fri Apr 29): **Final project presentations**
        - **Final project slide decks + Jupyter notebooks due Mon May 2, 11:59pm by email (1 email per group)**