CMU 95-865 UNSTRUCTURED DATA ANALYTICS
(FALL 2020 MINI-2 SECTIONS A2/B2, 6 UNITS)

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

Lectures, time and location:

Currently, the plan is for lectures prior to Thanksgiving break to be in-person and live at the same time (i.e., I teach in a classroom and start a Zoom session at the start of class). After Thanksgiving, all instruction will be purely remote.

- Section A2: Tuesdays and Thursdays 5:10pm-6:30pm, HBH 1204 (until Thanksgiving) & live over Zoom
- Section B2: Mondays and Wednesdays 1:30pm-2:50pm, HBH 1204 (until Thanksgiving) & live over Zoom

Recitations: Fridays 1:30pm-2:50pm, remote (Zoom)

TAs:

- Xinyu Yao (xinyuyao ♣ andrew.cmu.edu)
- Xuejian Wang (xuejianw ♣ andrew.cmu.edu)

Office hours: TBD

Course webpage: www.andrew.cmu.edu/user/georgech/95-865/

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 cloud computing (Google Colab).

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

Skills are assessed by homework assignments and two exams.

Prerequisites: If you are a Heinz student, then you must have either (1) passed the Heinz Python exemption exam, or (2) taken 95-888 "Data-Focused Python" or 90-819 "Intermediate Programming with 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. 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 use standard Python machine learning libraries such as SCIKIT-LEARN and PYTORCH. Despite the three homework assignments being of varying difficulty, they are equally weighted. Homework assignments are submitted in Canvas.

Exams: There will be two quizzes of equal weight. These will require Python programming and submitting a completed Jupyter notebook. Example past exams will be provided. For Fall 2020, exams will be conducted as follows. On the day of the exam, any time during the day (Pittsburgh time, starting right after midnight), you can open the exam, but as soon as you open it, your time limit will start for when you must submit the exam by. You will have 80 minutes for each quiz. **Thus, do not open the exam until you are ready to take it.** This particular policy is in place to allow for students in different time zones to take the exam at a time that is more convenient for them, but keep in mind that at 11:59pm Pittsburgh time, the exam will no longer be available. Also, very importantly, we do not accept late exams whatsoever.

Grading: Grades will be determined using the following weights:

Assignment	Percentage of grade
Homework	20%
Quiz 1	40%
Quiz 2	40%*

Letter grades are assigned on a curve.

*We will have a Piazza discussion forum. Students with the most instructor-endorsed answers will receive a slight bonus at the end of the mini, which will be added directly to their quiz 2 score (a maximum of 5 bonus points; quiz 2 is out of 100 points prior to any bonus points being added).

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 exams, 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 (i.e., you do not have to tell us that you are using a late day as we will automatically figure this out). *Once you have exhausted your late days, work you submit late will not be accepted.* This policy only applies to homework; the exams 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 outline (subject to revision; see course webpage for most up-to-date calendar): 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.

- Week 1 (Oct 26–30):
  - Lecture 1: Course overview, basic text processing, frequency analysis
  - Lecture 2: Basic text analysis demo, co-occurrence analysis
  - Recitation: Python review
- Week 2 (Nov 2–6):
  - Lecture 3: Finding possibly related entities
  - Lecture 4: Wrap up finding possibly related entities; intro to dimensionality reduction with PCA
  - Recitation: Dimensionality reduction
  - **HW2 due Friday 11:59pm**
- Week 3 (Nov 9–13):
  - Lecture 5: Manifold learning
  - Lecture 6: Intro to clustering
  - Recitation: Quiz 1 review
- Week 4 (Nov 16–20):
  - Lecture 7: More on clustering
  - Lecture 8: Wrap up clustering; topic modeling with latent Dirichlet allocation
  - **Friday has no recitation and will be for Quiz 1 instead:** upon opening the quiz, you have 80 minutes to complete it
- Week 5 (Nov 23–27):
  - **HW2 due Monday 11:59pm**
  - Lecture 9: Introduction to predictive data analytics
  - No class on Wednesday/Thursday/Friday due to Thanksgiving
- Week 6 (Nov 30–Dec 4):
  - Lecture 10: Introduction to neural nets and deep learning
  - Lecture 11: Image analysis with convolutional neural nets
  - Recitation: Neural nets
- Week 7 (Dec 7–Dec 11):
  - Lecture 12: Time series analysis with recurrent neural nets
  - Lecture 13: Other deep learning topics, wrap-up
  - Recitation: Quiz 2 review
  - **HW3 due Friday 11:59pm**
- Final exam period Dec 14–20: **Quiz 2** (exact day TBD)