CMU 95-865 UNSTRUCTURED DATA ANALYTICS
(SPRING 2021 MINI-4 SECTIONS A4/B4/K4/Z4, 6 UNITS)

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

**Lectures, time and location:**

- Section A4: Mondays and Wednesdays 4:50pm-6:10pm Pittsburgh time, HBH 1202 & live over Zoom
- Section B4: Mondays and Wednesdays 3:10pm-4:30pm Pittsburgh time, HBH 1202 & live over Zoom
- Section K4: Wednesdays and Fridays 10:30am-11:50am Adelaide time, live over Zoom
- Section Z4: asynchronous (watch Zoom recording)

**Pittsburgh recitations:** Fridays 3:10pm-4:30pm HBH A301 & live over Zoom

**Adelaide recitations:** TBD

**TAs for Pittsburgh:** Jingbo Jiang (jingboj ♣ andrew.cmu.edu), Xuejian Wang (xuejianw ♣ andrew.cmu.edu)

**TA for Adelaide:** Erick Rodriguez (erickger ♣ 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 and that are both 80 minutes long. These will require Python programming and submitting a completed Jupyter notebook. Example past exams will be provided. The first quiz is held during a recitation slot. The second quiz is held during the final exam period.

**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.

**Course outline (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.

- Lecture 1 (Mar 22 for PIT, Mar 24 for ADL): Course overview
- Part I: Exploratory data analysis
  - Lecture 2 (Mar 24 for PIT, Mar 26 for ADL): Basic text analysis demo, co-occurrence analysis
  - Lecture 3 (Mar 29 for PIT, Mar 31 for ADL): Finding possibly related entities
  - Lecture 4 (Mar 31 for PIT, Apr 2 for ADL): Visualizing high-dimensional data with PCA
  - Lecture 5 (Apr 2 for PIT, Apr 2 for ADL): Manifold learning
  - Lecture 6 (Apr 7 for PIT, Apr 7 for ADL): Clustering I
  - **HW1 due Apr 7, 11:59pm Pittsburgh time**
  - Lecture 7 (Apr 12 for PIT, Apr 9 for ADL): Clustering II

  - Lecture 8 (Apr 14 for PIT, Apr 14 for ADL): Clustering III
  - Lecture 9 (Apr 19 for PIT, Apr 16 for ADL): Topic modeling with latent Dirichlet allocation
  - **HW2 due Apr 19, 11:59pm Pittsburgh time**
  - **Quiz 1 is roughly at the end of Part I of the course (Apr 16 recitation slot for Adelaide, Apr 23 recitation slot for Pittsburgh)**
- Part II: Predictive data analysis:
  - Lecture 10 (Apr 21 for PIT, Apr 21 for ADL): Introduction to predictive data analytics
  - Lecture 11 (Apr 26 for PIT, Apr 23 for ADL): Introduction to neural nets and deep learning
  - Lecture 12 (Apr 28 for PIT, Apr 23 for ADL): Image analysis with convolutional neural nets
  - Lecture 13 (May 3 for PIT, Apr 28 for ADL): Time series analysis with recurrent neural nets
  - Lecture 14 (May 5 for PIT, Apr 30 for ADL): Course wrap-up
  - **HW3 is due during the final exam period (week of May 3-7 for Adelaide – exact day/time TBA; May 10 at 11:59pm for Pittsburgh)**
  - **Quiz 2 is during the final exam period (week of May 3-7 for Adelaide – exact day/time TBA; May 13 at 1pm-2:20pm for Pittsburgh)**