CMU 95-865 UNSTRUCTURED DATA ANALYTICS
(FALL 2022 MINI 2 SECTIONS A2/B2/C2, 6 UNITS)

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

- Section A2: Tuesdays and Thursdays 4:40pm-6pm, HBH 1002
- Section B2: Tuesdays and Thursdays 1:25pm-2:45pm, HBH 1004
- Section C2: Tuesdays and Thursdays 3:05pm-4:25pm, HBH 1002

Recitations: Fridays 1:25pm-2:45pm, HBH A301
TAs (listed here alphabetically by last name):

- Yuxin (Isabella) Hu (email: yuxinhu ♣ andrew.cmu.edu)
- Shahriar Noroozizadeh (email: snoroozi ♣ andrew.cmu.edu)
- Sumedh Shah (email: sumedhns ♣ andrew.cmu.edu)
- Hoi Ying (Lisa) Yeung (email: hyeung ♣ andrew.cmu.edu)

Office hours: TBD (check the course webpage for updates)
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 be coding lots of Python and dabble a bit with GPU 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 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 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 each 80 minutes long. These will require Python programming and submitting a completed Jupyter notebook. Example past exams will be provided.

Grading: Grades will be determined using the following weights:

Assignment	**Percentage of grade**
Homework	30%
Quiz 1	35%
Quiz 2	35%*

Letter grades are assigned on a curve.

*We will have a Piazza discussion forum. Students with the most instructor-endorsed posts on Piazza will receive a slight bonus at the end of the mini, which will be added directly to their Quiz 2 score (a maximum of 10 bonus points, so that it is possible to get 110 out of 100 points on Quiz 2).

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 source(s). For 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 (denoted below in red) 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 (denoted below in blue) of 95-865 turns toward making predictions once we have some idea of what structure underlies the data.

- Week 1:
  - Lecture 1 (Tue Oct 25): Course overview, analyzing text using frequencies
  - Lecture 2 (Thur Oct 27): Basic text analysis demo, co-occurrence analysis
  - There will be an optional Python review to be scheduled outside of class time
- Week 2:
  - Lecture 3 (Tue Nov 1): Co-occurrence analysis (cont'd), visualizing high-dimensional data with PCA
  - Lecture 4 (Thur Nov 3): PCA (cont'd), manifold learning
  - Recitation slot (Fri Nov 4): Lecture 5 – Manifold learning (cont'd), clustering
- Week 3:
  - Lecture 6 (Tue Nov 8): Clustering (cont'd) — k-means, GMMs
  - **HW1 due Tue Nov 8, 11:59pm**
  - Lecture 7 (Thur Nov 10): Clustering (cont'd) — interpreting GMMs, automatically selecting the number of clusters
  - Recitation slot (Fri Nov 11): Clustering on images
- Week 4:
  - Note: There will be a Quiz 1 review session scheduled outside of class time
  - Lecture 8 (Tue Nov 15): Topic modeling
  - Lecture 9 (Thur Nov 17): Intro to predictive data analysis
  - Recitation slot (Fri Nov 18): **Quiz 1 (80-minute exam)**
    - Quiz 1's coverage: up to and including the end of week 3's content
- Week 5:
  - Lecture 10 (Tue Nov 22): Hyperparameter tuning, decision trees & forests, classifier evaluation
  - **HW2 due Tuesday Nov 22, 11:59pm**
  - **No class on Thursday Nov 24 and Friday Nov 25 (Thanksgiving)**
- Week 6:
  - Lecture 11 (Tue Nov 29): Intro to neural nets and deep learning
  - Lecture 12 (Thur Dec 1): Image analysis with convolutional neural nets
  - Recitation slot (Fri Dec 2): More on classifier evaluation
- Week 7:
  - Lecture 13 (Tue Dec 6): Time series analysis with recurrent neural nets
  - Lecture 14 (Thu Dec 8): Additional deep learning topics and course wrap-up
  - Recitation slot (Fri Dec 9): Quiz 2 review
  - **HW3 due Friday Dec 9, 11:59pm**
- **Quiz 2** (80-minute exam) will be during the final exam week (exact date/time/location TBA)
  - Quiz 2 focuses on topic modeling and all the prediction topics taught in the course; i.e., Quiz 2 emphasizes material after Quiz 1's coverage (note that by how the course is set up, material from weeks 4–7 naturally at times relates to material from weeks 1–3, so some ideas in these earlier weeks could still possibly show up on Quiz 2, but if they show up, it will be in the context of topic modeling or prediction — please focus your studying on material from weeks 4–7)