CMU 95-865 UNSTRUCTURED DATA ANALYTICS
(FALL 2024 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 5pm-6:20pm, HBH 1002
- Section B2: Mondays and Wednesdays 5pm-6:20pm, HBH 2008
- Section C2: Tuesdays and Thursdays 3:30pm-4:50pm, HBH 1002

Recitations: Fridays 2pm-3:20pm, HBH A301

TAs (sorted alphabetically by last name):

- Ryan Chen (email: shuyic ♣ andrew.cmu.edu)
- Zekai Fan (email: zekaifan ♣ andrew.cmu.edu)
- Yubo Li (email: yubol ♣ andrew.cmu.edu)
- Tanyue Yao (email: tanyuey ♣ 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, including going over basics of large language models. We will be coding lots of Python and dabble a bit with GPU computing (Google Colab).

Note regarding GenAI and foundation models (such as large language models): As likely all of you are aware, there are now technologies like (Chat)GPT, Gemini, Claude, Llama, etc which will all be getting better over time. If you use any of these in your homework, please cite them. For the purposes of the class, I will view these as external collaborators/resources. For exams, I want to make sure that you actually understand the material and are not just telling me what someone else or GPT/Gemini/etc knows. This is important so that in the future, if you use AI technologies to assist you in your data analysis, you have enough background knowledge to check for yourself whether you think the AI is giving you a solution that is correct or not. For this reason, exams in this class will explicitly not allow electronics.

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 sklearn and PyTorch. Despite the three homework assignments being of varying difficulty, they are equally weighted. Homework assignments are submitted in Gradescope.

Exams: There will be two in-person in-class exams that are each 80 minutes long. These exams are "paper and pencil" exams. You are not allowed to use any electronics to complete the exam (no phones, no calculators, no tablets, no computers, etc). Note that this is the third semester in which exams in 95-865 are paper and pencil exams. *Even though we will be providing past exams, past exams from Spring 2023 and earlier are very different because they used to be done on a computer. Exams from Fall 2023 onwards are paper and pencil exams.*

You may bring as many sheets of notes as you would like but electronic devices will strictly be prohibited.

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 such as ChatGPT, stackoverflow, etc), please cite your source(s) *(note: you are not required to cite lecture slides or demos from 95-865)*. For exams, your answers must reflect your work alone (and not that of anyone else or of any AI technology). 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 Gradescope (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 use the Gradescope regrade feature. The course staff will make it clear by what date and time 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 (Mon Oct 21/Tue Oct 22): Course overview, analyzing text using frequencies
  - Lecture 2 (Wed Oct 23/Thur Oct 24): Basic text analysis demo, co-occurrence analysis
  - Recitation slot (Fri Oct 25): Python review
- Week 2:
  - Lecture 3 (Mon Oct 28/Tue Oct 29): Wrap up basic text analytics, co-occurrence analysis
  - Lecture 4 (Wed Oct 30/Thur Oct 31): Co-occurrence analysis (cont'd), visualizing high-dimensional data with PCA
  - Recitation slot (Fri Nov 1): *Lecture* 5 — PCA (cont'd), manifold learning (Isomap, MDS)
- Week 3:
  - **No class Mon Nov 4/Tue Nov 5** (there's Democracy day prior to 5pm on Tue Nov 5 and for simplicity to keep the different sections synced, I'm just cancelling class both days; the lecture that was supposed to be here happens earlier and is on Friday Nov 1 of week 2)
  - **HW1 due Mon Nov 4, 11:59pm**
  - Lecture 6 (Wed Nov 6/Thur Nov 7): Manifold learning (cont'd)
  - Recitation slot (Fri Nov 8): More on PCA, argsort
- Week 4:
  - Note: We will be scheduling a Quiz 1 review session outside of class time
  - Lecture 7 (Mon Nov 11/Tue Nov 12): Clustering
  - Lecture 8 (Wed Nov 13/Thur Nov 14): Clustering (cont'd)
  - Recitation slot (Fri Nov 15): **Quiz 1 (80-minute exam)**
    - Quiz 1's coverage: up to and including the end of week 3's content
- Week 5:
  - Lecture 9 (Mon Nov 18/Tue Nov 19): Topic modeling

  - Lecture 10 (Wed Nov 20/Thur Nov 21): Wrap up topic modeling, intro to predictive data analysis
  - Recitation slot (Fri Nov 22): *Lecture* 11 — Intro to neural nets and deep learning
- Week 6:
  - Lecture 12 (Mon Nov 25/Tue Nov 26): Image analysis with convolutional neural nets
  - **HW2 due Mon Nov 25, 11:59pm**
  - **No class Wed Nov 27–Fri Nov 29: Thanksgiving holiday**
- Week 7:
  - Lecture 13 (Mon Dec 2/Tue Dec 3): Time series analysis with recurrent neural nets
  - Lecture 14 (Wed Dec 4/Thur Dec 5): Text generation with RNNs and generative pre-trained transformers (GPTs); course wrap-up
  - Recitation slot (Fri Dec 6): TBD
  - Note: We will be scheduling a Quiz 2 review session outside of class time
- **HW3 due Mon Dec 9, 11:59pm**
- **Quiz 2** (80-minute exam): time/location TBD
  - Quiz 2 focuses on material from weeks 4–7 (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— please focus your studying on material from weeks 4–7)