CMU 95-865 UNSTRUCTURED DATA ANALYTICS
(FALL 2019 MINI-2 SECTIONS A2/B2/K2, 6 UNITS)

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

Lectures, time and location:

- Section A2: Tuesdays and Thursdays 4:30pm-5:50pm Eastern Time, HBH 1002
- Section B2: Mondays and Wednesdays 1:30pm-2:50pm Eastern Time, HBH 1202
- Section K2: Tuesdays 9am-11:50am Australian Central Time, Room 5

Recitations for Pittsburgh sections A2 & B2: Fridays 1:30pm-2:50pm Eastern Time, HBH A301

Recitations for Adelaide: TBD

TAs for Pittsburgh: Daniel Chen (dpchen ♣ andrew.cmu.edu), Emaad Manzoor (emaad ♣ 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 Amazon Web Services (AWS) for cloud computing.

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 using Amazon Web Services (AWS)
- Apply unstructured data analysis techniques discussed in class to solve problems faced by governments and companies

Skills are assessed by homework assignments and two quizzes.

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 16-791 "Applied Data Science". 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 KERAS. Despite the three homework assignments being of varying difficulty, they are equally weighted. Assignments are submitted in Canvas.

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.

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 quizzes, 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 quizzes 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 21–Oct 25):
  - Lecture 1: Course overview, basic text processing, frequency analysis; **HW1 released**
  - Lecture 2: Basic text analysis demo, co-occurrence analysis
  - Recitation: Python review
  - **For Pittsburgh sections, no recitation on Oct 25 due to CMU's Tartan Community Day**
- Week 2 (Oct 28–Nov 1):
  - Lecture 3: Finding possibly related entities, PCA, Isomap
  - Lecture 4: t-SNE
  - Recitation: t-SNE
  - **HW1 due Thursday 11:59pm Eastern Time; HW2 released**
- Week 3 (Nov 4–Nov 8):
  - Lecture 5: Introduction to clustering, k-means, Gaussian mixture models
  - Lecture 6: Clustering and clustering interpretation demo, automatic selection of k with CH index
  - Recitation: Quiz 1 review
- Week 4 (Nov 11–Nov 15):
  - Lecture 7: Hierarchical clustering, topic modeling

  - Lecture 8: Introduction to predictive analytics
  - **Quiz 1** (during recitation slot for Pittsburgh, to be scheduled for Adelaide)
  - **HW2 due Thursday 11:59pm Eastern Time; HW3 released**
- Week 5 (Nov 18–Nov 22):
  - Lecture 9: Model validation, decision trees/forests
  - Lecture 10: Introduction to neural nets and deep learning
  - Recitation: SVM, ROC curves
- Week 6 (Nov 25–Nov 29):
  - Lecture 11: Image analysis with convolutional neural nets
  - Lecture 12 (Australia only): Time series analysis with recurrent neural nets, other deep learning topics, wrap-up
  - Recitation: TBD
- Week 7 (Dec 2–Dec 6):
  - Lecture 12 (Pittsburgh only): Time series analysis with recurrent neural nets, other deep learning topics, wrap-up
  - Lecture 13 (Pittsburgh only): Recitation material from week 6 that's covered in Adelaide
  - **Quiz 2** (during recitation slot for Pittsburgh, to be scheduled for Adelaide)
  - **HW3 due Thursday 11:59pm Eastern Time**

Important: After the course is over, be sure that you terminate your AWS instances and delete your data volumes so you don't get charged for unused compute resources!