Carnegie Mellon University

Heinz College of Information Systems and Public Policy

94-775 & 94-475
Practical Unstructured Data Analytics

Spring 2024

Instructor:	Woody Shixiang Zhu	**Time:**	TR 3:30-4:50 (B3)
E-mail:	shixianz@andrew.cmu.edu	**Room:**	HBH 1204

Course description Organizations like companies, governments, and others are currently gathering a huge amount of data that is composed of various forms such as text, images, audio, and video. The question is how to convert this diverse and disorganized data into useful information. One common issue is that the underlying structure of the data is not always known before analyzing it, which is why it is called "unstructured." This course aims to provide a hands-on approach to analyzing unstructured data. We first investigate how to recognize any potential structure that may be present in the data through utilizing visual representation and other techniques for investigating the data.

Once we have indications of what structure may be present in the data, we can use it to make predictions. Throughout the course, we will come across several widely used techniques for analyzing unstructured data. This includes both established methods such as manifold learning, clustering, and topic modeling, as well as newer approaches like deep neural networks for analyzing text, images, and time series. Programming in Python using tools like Jupyter Notebook or Colab will be a significant component of the course. Additionally, the use of ChatGPT is also encouraged throughout.

Prerequisites If you are a Heinz student, then you must have already completed 95-791 "Data Mining" and also one of either 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. It would be better to have working knowledge of undergraduate level probability, linear algebra, and statistics.

Course Materials There is no official textbook for the course. I will post all the lecture notes and related readings on Canvas. You can also find a list of recommended reading below:

- ☐ *Mining of Massive Datasets* by Jure Leskovec, Anand Rajaraman, and Jeffery D. Ullman;
- ☐ *Foundations of Machine Learning* by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar;
- ☐ *Machine learning: A probabilistic perspective* by Kevin P. Murphy;
- ☐ *Applied Statistics and Probability for Engineers* by Douglas C. Montgomery and George C. Runger;
- ☐ *Introduction to Time Series and Forecasting* by Peter J. Brockwell and Richard A. Davis.

Instructor Office hours TBD

Teaching Assistants

- Zekai Fan (zekai@cmu.edu)

Grading policy Your grade will be evaluated based on 3 *homework assignments* and 2 *quizzes*. The grade composition consists of

- Homework (10% + 15% + 15%)
- Quiz 1 (30%)
- Quiz 2 (30%)

Letter grades are determined based on a curve.

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 closed-book quizzes of equal weight and that are each 80 minutes long. You are allowed to take one A4-size paper of cheat sheet and expected to work on the exam independently. There are no make-ups; if there is any conflict, please let me and TAs know ASAP and you may take the exam before the assigned period.

Class attendance and participation The learning process of this class is based on in-class discussion and participation. Attendance and careful preparation of the course material is therefore highly recommended.

Late submission policy We have the following accommodation policies to help with emergent situations: 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). This policy only applies to homework; the exams must be submitted on time to receive any credit. For example: 1. You could use the two late days on two different assignments to turn them in each 1 day (24 hours) late without penalty; 2. 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. Once you have exhausted your late days, work you submit late will not be accepted. This policy only applies to homework; the exams (quizzes) must be submitted on time to receive any credit.

Communications All communication from your instructor will take place in Canvas. You are expected to check Canvas every day for important course-related information. However, by following the course instructions, you can also ensure that you do not miss important instructions, announcements, etc. by adjusting your account settings to receive important information directly to your email account.

For all your administrative requests, such as homework regarding, please email your TA (do not leave message on Canvas or Piazza). To request a regrade for an assignment, submit a written explanation to your TA and copy the instructor. Keep in mind that the entire assignment will be reviewed and your grade may decrease.

For content questions and help, because questions can often be addressed for the good of the group, please do not email your questions directly to the instructor. Instead, course and content questions will be addressed on Piazza (Signup at Link with access code "2023"). Feel free to set your post to private to ask questions about your grade or other issues unique to you. Please be courteous when posting on Piazza and treat fellow students, TA, and instructor with respect. In the public post, please do not show any of your answers related to the homework problems, such as code snippets. If you would like to show the plots (which does not disclose the explicit answer to the questions) from your implementation in the discussion, please either make them private post (only share with teaching staffs) and/or add watermarks to those images/results. Please be specific when raising the question. In principle, instructors are not responsible for the program debugging and will not comment on the pure coding problem. For example, please do not send the code file to TA or posting a question showing a section of code and asking such as "why it doesn't work".

Plagiarism Plagiarism is considered a serious offense. You are not allowed to copy and paste or submit materials created or published by others, as if you created the materials. All materials submitted and posted must be your own original work.

Academic integrity All students are expected to comply with CMU's policy on academic integrity. Please read the policy and make sure you have a complete understanding of it.

Tentative Course schedule:

The course is divided into three main sections. The first focuses on understanding the nature of unstructured data and learning to uncover its patterns and structure through computational methods. The second section builds on this understanding by using it to make predictions about the data. The last section introduces advanced modeling techniques and cutting-edge machine learning methods in unstructured data analytics.

Please refer to Carnegie Mellon 2023-2024 Academic Calendar for more information about course schedule.

Week 0: Preparation

Before starting with the course, please get some first insights into Python and Google Colab so that we can depart from a similar level.

❗ Please read the Python Cheatsheet and take a Python Quiz before the course if you are not familiar with programming in Python.

Week 1: Introduction & Unstructured Data Modeling

💡 Topics

* Course overview;
* Introduction to unstructured data;
* Traditional image and graph modeling;
* Text modeling and co-occurrence analysis.

📝 Homework 1 released (Jan 18, 2024).

📂 Recitation

* Reading / writing data and package installation on colab.
* spaCy and sklearn tutorial.

Week 2: Dimensionality Reduction

💡 Topics

* Principal Component Analysis (PCA);
* Manifold learning;
* t-distributed stochastic neighbor embedding (t-SNE).

📂 Recitation

* Demo: Text modeling and analysis.

Week 3 & 4: Clustering and Topic Modeling

Topics

* *k*-Means;
* Gaussian Mixture Model (GMM);
* Hierarchical Clustering and Density-Based Clustering;
* Latent Dirichlet Allocation (LDA).

Homework 1 due and Homework 2 released in Week 3 (Feb 1, 2024).

Recitation

* Case study: Police 911 calls-for-service text analysis (Week 3).
* Quiz 1 (Week 4).

Week 5: Predictive Data Analysis

Topics

* Hyperparameter tuning;
* Linear regression;
* Decision trees & forests;
* Classifier / predictor evaluation.

Homework 2 due and Homework 3 released (Feb 15, 2024).

Recitation

* Case study: COVID-19 cases and deaths analysis.

Week 6 - 7: Advanced Unstructured Data Modeling

Topics

* Spatio-temporal data and auto regressive model;
* Recurrent neural network (RNN);
* Attention mechanism and Transformer;
* Convoluntional neural network;
* Graph neural network.

Recitation

* Review session: HW Q&A and Quiz practice (Week 6).

Week 7: Generative Modeling and Analytics

Topics

* Variational autoencoder;
* Diffusion model;
* Large language model (LLM).

Homework 3 due in Week 7 (Feb 29, 2024).

Recitation

* Quiz 2.