



Heinz College of
Information Systems
and Public Policy

94-775 & 94-475 Practical Unstructured Data Analytics

Spring 2023

Instructor:	Woody Shixiang Zhu	Time:	TR 2:00-3:20 (A3) 3:30-4:50 (B3)
E-mail:	shixianz@andrew.cmu.edu	Room:	HBH 1204 (Lab: HBH A301)

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. The course will involve a lot of coding using Python and we will also explore using GPU computing through Google Colab.

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.

Instructor Office hours Tue 5:00-6:00 PM

Teaching Assistants

- Yingxue Li (yingxuel@andrew.cmu.edu)
- Annabel Qihui Hu (qihuih@andrew.cmu.edu)
- Xixiang Hu (xiyanghu@andrew.cmu.edu)
- Lisa Hoi Ying Yeung (hyeung@andrew.cmu.edu)

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

- Homework (30%)
- Quiz 1 (35%)
- Quiz 2 (35%)

The grading scale in Table 1 will be used in the course. Scores below 65% equate to a failing grade (R)

Table 1: Grading Scale

A+	97% – 100%	B+	85% – 88.99%	C+	73% – 76.99%
A	93% – 96.99%	B	81% – 84.99%	C	69% – 72.99%
A–	89% – 92.99%	B–	77% – 80.99%	C–	65% – 68.99%

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.

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.

If you have already used the above 3 days of homework extension, and if you submit the homework late: one day late the grade will be discount to 75% of your total, two days late the grade will discount to 50% of your total, three days late the grade will discount to 25% of your total. Past three days, your homework will not be accepted.

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. 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 two main sections. The first focuses on analyzing a dataset to uncover its patterns and structure through computation and visualization. The second section builds on this understanding by using it to make predictions about the data.

Please refer to Carnegie Mellon 2022-2023 [Academic Calendar](#) for more information about course schedule.

Week 0: Preparation

Before starting with the course, please get some first insights into Python 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 & Text Modeling

💡 Topics

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

📝 Homework 1 released

Week 2: Dimensionality Reduction

💡 Topics

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

Week 3: Clustering

💡 Topics


- * k -Means;
- * Gaussian Mixture Model (GMM).

 Homework 1 due and Homework 2 released

Week 4: Topic Modeling

 Topics

- * Latent Dirichlet Allocation (LDA).

 Quiz 1

Week 5: Predictive Data Analysis

 Topics

- * Hyperparameter tuning;
- * decision trees & forests;
- * classifier evaluation.

 Homework 2 due and Homework 3 released

Week 6: Neural Networks and Deep Learning


 Topics

- * Introduction to Neural Networks;
- * Convolutional neural network (CNN).

Week 7: Time Series Analysis

 Topics

- * Auto Regressive Model;
- * Recurrent neural network (RNN) and long-short term memory (LSTM).

 Homework 3 due and Quiz 2