Unstructured Data Analytics
ITAO70250

Office: 337 Mendoza

Email: seth.berry@nd.edu

Office Hours

M – 11:00 to 1:00

W – 11:00 to 1:00

F – 10:00 to 12:00

We can also always find mutual times that will work.

Class Days and Time

Section 1: TR, 8:00 to 9:50

Section 2: TR, 10:00 to 11:50

Location – Stayer B003

Course Description

Huge amounts of the world's data is unstructured. Developing competency in how to harness this type of data in order to develop critical insights has significant value for today's business. This course introduces the fundamental concepts of unstructured data analytics, from data acquisition and preparation to applying supervised and unsupervised machine learning approaches such as text analysis, dimension reduction, object recognition/detection, and recommender systems. In the context of unstructured data analytics, students will also be introduced to the principles of deep neural network and transformers.

Learning Goals

By successfully completing this course, you will fulfill the following objectives:

- Gain a foundational understanding of both supervised and unsupervised machine learning approaches to unstructured data.
- Develop an applied knowledge of some of the common unstructured data acquisition, exploration, and preparation approaches using R and Python.
- Understand the theoretical concepts behind text analysis, deep neural networks, and recommender systems.
- Develop an applied knowledge of how to implement the approaches discussed in the course using R and Python.

Readings

There is no official textbook for this course, but here are some good resources:

Text Mining with R

R for Data Science

Creating Functions

Additional resources will be linked within course notes and on Canvas.

Homework

During the course of the mod, we will have 3 homework assignments (worth 60, 60, and 80 points). All homework assignments must be submitted in a compiled file (knitted from R Markdown or a Python-flavored notebook of your preference) – no other file types will be accepted and reminders won't be given. Homework is to be completed on your own. While you are welcome to think through problems together, all code and words should be your own. Homework is due within 9 days of it being assigned.

Presentations

As opposed to a final exam, we will be having presentations on our last day of class. These are to be completed individually and presentations will have a 3 minute time limit. The goal is to answer a question or solve a problem. Presentation guidelines will follow, but general creativity and appropriate technique use will figure heavily into your grade. This is a chance for you to find interesting data, not just go with what might be easy on Kaggle.

Engagement

Engagement is not just coming to class, but being an active participant. Throughout class, you will be given the opportunity to practice content. At the end of each class, you need to

turn in your code (it does not need to be pretty and can just be any text-based file). Each submission is worth 10 points for up to a maximum of 100 points.

Grade Breakdown

Engagement – 100 points (25%)

Homework – 200 points (50%)

Presentation – 100 points (25%)

Total – 400 points

Schedule

Week	Date	Topic	Method	Homework
1	01/11 (T)	Introduction and Modern I/O		
	01/13 (R)	Regular Expressions		
2	01/18 (T)	APIs		#1 Assigned
	01/20 (R)	Scraping		
3	01/25 (T)	Text Processing	tf-idf	
	01/27 (R)	Sentiment Analysis	word & sentence level sentiment	#2 Assigned #1 Due
4	02/01 (T)	Matrix Algebra & Dimension Reduction	PCA	
	02/03 (R)	Topic Models	LDA & Transformers	
5	02/08 (T)	Text Classification	Deep Neural Networks	#3 Assigned #2 Due
	02/10 (R)	Object Classification	Convolutional Neural Networks	
6	02/15 (T)	Object Detection	YOLO and friends	
	02/17 (R)	Recommender Systems	Hybrid systems	
7	02/22 (T)	Presentations		