Unstructured Data Analytics
ITAO70250

Office: 337 Mendoza

Email: seth.berry@nd.edu

Office Hours

MWF – 10:00 to 12:00 (Open Time)

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 and summarization, text recognition and classification, sentiment analysis, topic modeling, and image classification. In the context of unstructured data analytics, students will also be introduced to the principles behind such classic machine learning algorithms such as naive bayes, support vector machines, and neural networks.

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 summarization, sentiment analysis, topic modeling, naive bayes, neural networks, and support vector machines.
- 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

The apply family

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

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 an 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 will be composed of three distinct parts – initial submission, individual feedback session, second submission – and each will count towards 1/3 of the points. The initial submission must be completed before any feedback can be offered.

Presentations

As opposed to a final exam, we will be having presentations on our last day of class. You can work individually or as duos. These presentations are not to exceed 4 minutes and will be on a course topic of your choosing. 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	Assignments
1	02/02 (T)	Introduction & Programming	
	02/04 (R)	Regular Expressions	
2	02/09 (T)	Data Collection (1a)	
	02/11 (R)	Data Collection (1b)	Homework #1
3	02/16 (T)	Text Analysis (2)	
	02/18 (R)	Sentiment Analysis	
4	02/23 (T)	Topic Modeling (4)	Homework #2
	02/25 (R)	Lab 1 (3)	
5	03/02 (T)	Text Classification (5)	
	03/04 (R)	Lab 2 (6)	Homework #3
6	03/09 (T)	Optical Character Recognition (7)	
	03/11 (R)	Lab 3 (8)	
7	02/16 (T)	Image Classification (9)	
	02/18 (R)	Presentations	

1a. Using APIs

1b. Web scraping

2. Term frequency, inverse document frequency, part of speech tagging, and relationships
3. Practicum on text collection, exploration, and preparation
4. Latent Semantic Analysis, Latent Dirichlet Allocation, and NNMF
5. Naive Bayes for document classification
6. Practicum on text analysis
7. Support Vector Machines and their application to identifying text (OCR)
8. Practicum on supervised text analysis
9. Artificial Neural Networks and image classification.