Unstructured Data Analytics
ITAO70250

Office: 337 Mendoza

Email: seth.berry@nd.edu

Office Hours

Tuesdays & Thursdays – 9:00 to 12:00

These are the *official* office hours. If you find my door ajar to any degree (it will typically be less than 10° and the office will be dark), then you are more than welcome to drop in chat with me about anything (stats, programming, career, etc.).

Class Days and Time

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

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

Location – L004

Course Description

The vast majority 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 artificial 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.
- 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.

Attendance

While I will not be taking attendance in a strict sense, we will have in-class exercises for our 7 main topics (each worth 10 points). This will serve two purposes: 1) it will be good practice and 2) it will be a participation grade.

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

Homework

During the course of the mod, we will have 4 homework assignments (worth 20, 40, 60, and 80 points). All homework assignments must be submitted in an html file. You will have 2 weeks to complete each assignment.

Presentations

As opposed to a final exam, we will be having presentations on our last day of class. These presentations are not to exceed 3 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.

Grade Breakdown

In-class exercises – 70 points (11%)

Homework – 200 points (31%)

Presentation – 50 points (17%)

Participation – 30 points

Total – 350 points

Schedule

Week	Date	Topic	Assignments
1	01/13 (M)	Introduction	
	01/15 (W)	Data Collection and Preparation (1)	CSS Diner and Regex Tester
2	01/20 (M)	Text Analysis (2)	
	01/22 (W)	Sentiment Analysis	
3	01/27 (M)	Topic Modeling (4)	Homework #1
	01/29 (W)	Lab 1 (3)	
4	02/03 (M)	Text Classification (5)	
	02/05 (W)	Lab 2 (6)	Homework #2
5	02/10 (M)	Optical Character Recognition (7)	
	02/12 (W)	Lab 3 (8)	
6	02/17 (M)	Image Classification (9)	Homework #3
	02/19 (W)	Deep Networks with Python	
7	02/24 (M)	Emerging Issues	
	02/27 (Tr)	Presentations	

1. Web data in JSON, HTML and/or XML formats, API data, text, images
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.