Global Health Survey Design
GH 60585

Office: 337 Mendoza

Email: seth.berry@nd.edu

Office Hours

Tuesdays, Thursdays, & Fridays – 12:00 to 2:00

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

Friday – 2:00 to 5:00

Location – Jordan 310

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 textual 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 every week. These 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:

Question Understanding Aid (QUAID)

Survey Research Methods

Journal of Survey Statistics and Methodology

Social Science Computer Review

lavaan

psych

Homework

We will have 3 homework assignments. These assignments will largely be based upon your own project work.

Presentations

As opposed to a final exam, we will be having presentations on our last day of class.

Grade Breakdown

In-class exercises – 40 points (11%)

Homework – 90 points (31%)

Presentation – 50 points (17%)

Participation – 30 points

Total – 350 points

A = 333+ points

A- = 315-332 points

B+ = 305-314 points

B = 294-304 points

B- = 280-293 points

C+ = 270-279 points

Schedule

Week	Date	Topic	Assignments
1	01/25 (F)	Item and Survey Design (1)	
2	02/08 (F)	Survey Programming (2)	Homework #1
3	03/15 (F)	Analyses (3)	Homework #2
4	04/22 (F)	Visualization & Presentations (4)	Homework #3

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