Contact and Office Hours

- **Instructor for weeks 1 and 3**: Matthew Denny, email: matthewjdenny@gmail.com (pronouns: he/him). He goes by Matt, or Dr. Denny if you prefer.
- **Instructor for weeks 2 and 4**: Kelsey Shoub, email: kelsey.shoub@gmail.com (pronouns: she/her). She goes by Kelsey, or Dr. Shoub, if you prefer.
- **Course Teaching Assistant**: Helge Marahrens, email: hmarahre@iu.edu
- **Office Hours**: TBD, please check course webpage.

Course Overview

The social sciences have experienced an explosion of interest in data science, with a special emphasis on the quantitative analysis of textual data, over the past decade. This interest has not been misplaced, as data science techniques and text-as-data studies have revealed important social phenomena across a wide range of disciplines, and at a previously inconceivable scale. Yet, as with any emerging fields, there are many poorly documented pitfalls, as well as significant technical and theoretical challenges. This course seeks to arm its participants with the theoretical background, practical experience, and technical capacity to pursue cutting edge social science research using text data. This course is designed to cover the key technical aspects of conducting research with text data with a special emphasis on techniques drawn from the emerging field of data science: from data collection and preprocessing, through to description and inferential analysis. Students are expected to have some experience with R programming, and some background in statistical analysis for the social sciences.

Goals

By the end of the course, participants should possess the basic skills necessary to perform most data collection and management tasks they are likely to encounter over the course of a research project intended for publication as a scholarly journal article. They should also have the skills and experience to collect and preprocess digital text data, and apply a number of standard text analysis techniques to their data.

Prerequisites

Some background in statistical inference, regression analysis, etc. with some basic R programming experience a plus.
• This course is intended for graduate students in the social sciences with some background in statistical analysis for the social sciences. We expect that you will have some experience with hypothesis testing, regression analysis and some basic background in statistics, as we do not intend to cover this material during the course. It is important that you have this background so that you will be able to follow along with the statistical models for text we cover in the later part of the course.

• We do not require any programming experience, although having some experience with R will certainly make the early part of the course less challenging. The first week of the course will be focused on getting everyone up to speed on R programming, including those with no background, so if you do not have experience using R or with programming, this course is still for you. For those with some experience using R, you may find the first week to be more review, before we dive into text analysis.

**Homework**

Students will have a weekly homework assignment (for a total of four assignments), typically assigned on Wednesday and due on Monday. A writeup with instructions for each assignment will be posted on the course website. Each student will submit their assignment on Monday and we will leave time for discussion of the assignments. More details will be provided in the assignment, but the writeup will typically be 2-5 pages, single spaced (including figures). Below is a brief overview of the four assignments for the class:

• **Programming Assignment**: Your first assignment will involve several coding exercises. You will submit an R script with your solutions, and comment in that R script as to how you came up with the solution.

• **Data Collection and Description**: Your second assignment will involve collecting a corpus of documents that you will analyze over the rest of the course. We will go over a variety of techniques for collecting these data in class. You will be expected to provide a writeup on the dataset you selected (including some basic descriptive statistics), and to provide a copy of your dataset in a format we will describe in the homework assignment. You may collect any corpus you like, and we will provide examples for those who do not have one in mind. Your corpus should satisfy the following criteria:
  - Your dataset should be 100+ documents.
  - Your dataset should be 100+ pages (30,000+ words) in total. This is important because some of the algorithms we use in this class require a reasonably large amount of text to produce actually interpretable results.
  - Your dataset should have one categorical variable and one continuous variable per document for at least 100 documents. You can hand-code these for a subset of documents if you are working with a larger dataset and there is not an easy way to gather these covariates automatically/programmatically.

• **Text Preprocessing and Exploration**: In this assignment, you will apply some of the preprocessing and description techniques we go over in the class to the corpus you collected in your previous homework assignment.

• **Text Analysis**: In the final assignment, you will apply some of the inferential methods we discuss in the final week of the class to your corpus.
Grades

Grades will be assigned based on performance on the homework assignments. Each assignment will be graded (0 - no submission, 1 - does not demonstrate comprehension of the methodology, 2 - adequate comprehension demonstrated, 3 - excellent comprehension demonstrated). Grade Scale (based on sum of assignment grades):

- A+ (12)
- A (10-11)
- A- (8-9)
- B+ (6-7)
- B- (<6)

Schedule

More details, along with links to assigned readings and class lab materials will be provided on the course Canvas page. Each class will typically start with a lecture component, followed by a lab. We will post the R scripts for each lab on the course Canvas page so that participants may follow along and adapt the provided code to help with their assignments.

Week 1: Programming and Data Management

This week will be taught by Dr. Denny.

- **Monday**: No class.
- **Tuesday**: Course Introduction, Setting up R, and R Basics.
- **Wednesday**: Looping and Conditionals, Data I/O and R Packages
- **Thursday**: Data Management
- **Friday**: Data Management Part 2

Week 2: Web Scraping, String Manipulation and Corpus Creation

This week will be taught by Dr. Shoub.

- **Monday**: Manipulating Strings
- **Tuesday**: Web Scraping
- **Wednesday**: Web Scraping (Part 2)
- **Thursday**: Loading, Representing and Describing Text Data
- **Friday**: Dictionaries
Week 3: Text Preprocessing and Exploration

This week will be taught by Dr. Denny.

- **Monday**: Text Preprocessing
- **Tuesday**: Basic NLP: Parts of Speech and Named Entity Recognition
- **Wednesday**: Term Category Associations
- **Thursday**: Text Reuse
- **Friday**: Word Embeddings

Week 4: Text Analysis

This week will be taught by Dr. Shoub.

- **Monday**: Topic Models
- **Tuesday**: Topic Models (Part 2)
- **Wednesday**: Supervised Learning
- **Thursday**: Supervised Learning (Part 2)