Collecting and Analyzing Big Data

Neal Caren
University of North Carolina, Chapel Hill

This course is an introduction to collecting and analyzing "big data" for social scientists. Over the last decade, the variety and types of data available to researchers have exploded. This includes not only contemporary data, such as from websites and social media platforms, but also historical data, from digitized interviews to 19th century newspapers. At the same time, analytic techniques from computer science are increasingly being used to solve social science problems.

One week is not enough time to master the techniques for collecting and analyzing big data. You will, however, be able to establish the foundation for developing these skills. The course is designed as a practical overview. The emphasis each class will be on applying the specific techniques rather than on their mathematical basis. The course will provide an overview in that each lesson will introduce a new method in order to demonstrate the range methods. Combined, students will have the skills and resources to apply these methods to theoretically-relevant problems in the social sciences.

Learning objectives:

By the end of the course, I expected that students will be able to:

- Collect data from the internet using web scraping and APIs.
- Read and write digital text files.
- Analyze data using supervised learning technique such as random forest models.
- Analyze data using unsupervised learning techniques such as topic models.
- Understand and apply current methods for analyzing texts.
- Link machine learning methods to relevant social science questions.
- Program in Python

Course Credit

Students have the option of submitting a research paper in order to receive ECTS credits. These research papers (6000 to 8,000 words) should apply one or more of the techniques used in the course to a theoretically interesting research question. Papers should generally follow the format of a research article in the student's discipline, although the literature review may be concise than normal. Additionally, students must provide code, and where feasible, data, to replicate the analysis. This is to be completed within 8 weeks after the course.

Requirements

Students should have a Python distribution appropriate for data science. The recommended way to do this is to install Continuum's Anaconda Python distribution. It is free and available for all operating systems. Students are *not* expected to have any knowledge of Python.

Course Outline

1. **Big data, machine learning and the social sciences**
   This lecture will unpack some of the major findings from the intersection of social

science and big data. The focus will be on the specific tools and methods that were used. We will also review the major sources of data and tools currently available for data science.
Reading: Müller and Guido, Chapter 1.

2. **Getting Started with Python**
   This lecture will focus on getting students up and running with Python for social science applications. This includes both an overview of the elements of the Python data science stack (e.g. IPython/Jupyter, pandas, matplotlib, scikit-learn) but a more detailed introduction to working with Python.
3. **Harvesting data from the web: APIs**
   Collecting big data is often done through web application programming interfaces, or APIs. This is a way for developers, or researchers, to access data stored on governmental or corporate servers. For example, Twitter, Facebook, and Yelp, all make some of their data available through APIs. This lecture will introduce the basics of collecting data from an API in Python.
   Reading: Mitchell, Ryan. 2015. *Web Scraping with Python: Collecting Data from the Modern Web* O'Reilly Media, Inc. Chapters 1,2.
   "Chronicling America API." http://chroniclingamerica.loc.gov/about/api/
4. **Harvesting data from the web: Web scraping**.
   A second major source of big data is collecting it directly from websites. Web scraping involves visiting one more pages and collecting and storing the relevant information in an automated fashion. This lecture will introduce the basics of web scraping in Python.
   Reading: Mitchell, Ryan. 2015. *Web Scraping with Python: Collecting Data from the Modern Web* O'Reilly Media, Inc. Chapter 4.
5. **Manipulating Big Data**
   By most estimates, 80% of data analysis is cleaning and merging the data. This lecture introduces best practices for preparing your data in Python.
   Reading: Vanderplas, Jake. 2017. *Python Data Science Handbook.*O'Reilly Media, Inc. Chapter 3.
6. **Supervised Learning I**
   You are likely familiar with supervised learning, but you probably don't call it that. Supervised learning in the machine language term for when you are modeling one variable as a function of another set of variables, such as linear or logistic regression. This lecture reviews common methods for regression and classifications such as linear regression and introduces more complex algorithms.
   Reading: Müller and Guido, Chapter 2.
7. **Model Evaluation**
   Keeping the data used to evaluate your model separate from the data used to develop your model is critical to the machine learning workflow. This is of particular concern when there are concerns about overfitting. This lecture introduces the idea of cross validation and reviews methods for evaluating model fit.
   Reading: Müller and Guido, Chapter 5.
8. **Supervised Learning** II
   This lecture extends on focus on supervised learning to techniques to include decision

trees and random forest models.
Reading: Müller and Guido, Chapter 2.

9. **Working with Text Data**
   This lecture introduces the basics of manipulating and analyzing text data, including counting and analyzing term frequencies for text categorization.
   Reading: Müller, Andreas C. and Sarah Guido, Chapter 7.
10. **Unsupervised Learning with Text Data**
    This lecture will introduce methods for analyzing themes in text data. The focus will be on topic modeling which involves assigning each document to one or multiple topics.
    Reading: Müller, Andreas C. and Sarah Guido, Chapter 7.

Reading list

Müller, Andreas C. and Sarah Guido. 2017. *Introduction to Machine Learning with Python: A Guide for Data Scientists* O'Reilly Media, Inc. 392 pages.

Additional readings will be made available.

The lecturer

Neal Caren is an Associate Professor of Sociology at the University of North Carolina, Chapel Hill. His research interests center on the quantitative analysis of protest and social movements. His work has been published in the *American Sociological Review*, *Social Forces*, *Social Problems*, and the *Annual Review of Sociology*. The data in many of his publications has been either scraped from the web, downloaded using APIs, or otherwise involved collected and analyzing texts. He is the author of a well-used publicly available script for converting Lexis-Nexis article downloads into a CSV file. For several years, he has run a graduate workshop on computational social science and digital data collection, has given external workshops on the topic, and has many several tutorials available online. He is also the editor of the social movements journal *Mobilization*.