LSE-PKU Summer School
Beijing, China

LPS-MY202 | Data Collection and Management with Python

COURSE OUTLINE

INSTRUCTOR

Dr Blake Miller is an Assistant Professor of Computational Social Science in the Methodology Department at the London School of Economics and Political Science. He received his PhD in Political Science and Scientific Computing from the University of Michigan in 2018 where he was also a graduate research affiliate in the Lieberthal-Rogel Center for Chinese Studies. Before coming to LSE, he was a Post-Doctoral Fellow at the Dartmouth College Program in Quantitative Social Science. For more information, please visit: www.blakeapm.com

COURSE OVERVIEW

The massive amount of data available online continues to increase the bounds of scientific inquiry. Researchers in both academia and the private sector can gain a greater understanding of human behaviour by analysing the abundant social data stored online. To make use of these data, one must first master technical skills necessary to gather and process these data, which can be quite challenging to do properly.

The main goal of this course is to provide students with the necessary tools for the construction, pre-processing, and cleaning of data found online. After taking this course, students will have mastered the requisite tools needed to construct datasets out of unstructured, semi-structured, and structured online data.

In this course students will learn the following tools:

- Pandas and JSON: How to store and access data in different formats.
- Parsing HTML: How to use basic HTML and CSS to extract information automatically from websites.
- How to write a basic scraper: write a program to dynamically crawl a website
  and gather relevant data.
- APIs: How to incorporate social network and geolocation data (e.g. from Twitter, Weibo, Google, Baidu, etc.) in one's data.

SQL: Basics of relational databases and how to access them via Python.

PREREQUISITES

- Essesstial: Knowledge of computer programming basics, (at least at the level of an undergraduate introductory programming course).
- Essesstial: At least basic knowledge of Python or R (or another programming language). **The course will be taught using Python.**
- Helpful: Knowledge of HTML and CSS.

ASSESSMENT

Assessment will be based on coursework (worth 50% of the final mark) and a final exam (worth 50% of the final mark).

READINGS

A full reading list and electronic course pack will be provided to registered students approximately six weeks before the beginning of the programme.
• Shaw, Zed A. Learn Python 3 the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code. Addison-Wesley Professional, 2017.