Introduction to Python

Social Science Methods Workshops 2022, Lund University

Instructor: Nils Holmberg, Strategic Communication, Lund University

Contact: nils.holmberg@isk.lu.se

Dates: October 17 – 21, at 13-16 pm (schedule)

Format: Campus-based (location)

1 Introduction

The Python programming language (check out Figure 1) has gained popularity in all types of data science in recent years. On this introductory workshop, we are aiming at getting acquainted with the basic syntax of Python, as well as learning how to extend this basic functionality by calling powerful modules built by members of the large Python user community. Thus, we take a “standing on the shoulders of giants” approach to Python!

Figure 1: Official logo of the Python programming language

Another principle that has guided the design of this course is that participants should gain an understanding of how Python can be used to accomplish fairly common research tasks within the social sciences more effectively and more transparently. Such applications of Python include collecting and analyzing survey data, performing content analyses on textual and visual material, and using charts to visualize results.

Table 1: Overview of schedule and course contents

Course date	Course topic	Python packages
2022-10-17	the anaconda3 environment, basic python syntax, process table data	jupyter, scipy, pandas
2022-10-18	data visualization, reproducible data analysis, sharing and collaborating	seaborn, altair, plotly
2022-10-19	text analysis, manifest content, data cleaning, tokenization, copora	nltk, pandas, plotly
2022-10-20	text analysis, latent content features, sentiment analysis, topic modelling	scikit-learn, spacy
2022-10-21	image analysis, latent content features, object recognition, captioning	opencv, pytorch, tf

2 Course materials

Course participants are encouraged to install a Python environment beforehand. The environment used on this course will be Anaconda3 and it consists of several modules, including support for interacting with Jupyter notebooks, which will be the main format of instruction on the course. If you plan to use a managed laptop from Lund University, please check with IT support how to install Anaconda3 via Software Center.[1]

3 Course contents

In this introduction to Python for social scientists you will learn how to use the Python programming language as a *unified platform* for handling multiple tasks and workflows connected to running research studies, collecting data, and analyzing numeric and textual data. In Table 1 we present an overview of the course topics covered each day, and further down we expand on each module in more detail.

First off, we will familiarize ourselves with the Anaconda3 environment. This software consists of several modules, including the Python command prompt, a text editor for writing scripts, and an interactive development environment (IDE). We will practice writing basic flow control and calling functions, then we will read, analyze and process some generic tabular datasets (McKinney, 2012; VanderPlas, 2016).

On day 2, we will continue performing basic data summarization using descriptive statistics, but now we will put more emphasis on how to use quantitative data to produce compelling and intuitive visualizations. Using the matplotlib and seaborn packages, we will learn how to aggregate survey data and present them as scatter plots, bar charts, and time series

[1] Please note that the behavior of the Anaconda3 python environment will be different depending on if it is installed as administrator or user.

(Embarak et al., 2018). If there is time, we will also use the scikit-learn package to perform multiple regression analyses in Python.

Next up, we will focus on analyzing and processing textual data. In order to do so, we will need to extend the basic functionality of Python by importing packages such as Natural Language Toolkit and spaCy, which will allow us to leverage natural language processing using machine learning models. We will practice on text preprocessing, cleaning, tokenization, and classifying movie reviews and/or news articles (Kedia & Rasu, 2020; Sarkar, 2016).

We will finish off the course by introducing some approaches to computational image analysis. Using frameworks such as PyTorch and TensorFlow, we will try to perform automatic image analyses at scale. Starting with manifest image features such as pixel dimensions and color histograms, the course proceeds to dealing more latent features such as image classification, and object recognition and localization (Géron, 2019; Szeliski, 2010).

References

Embarak, D. O., Embarak, & Karkal. (2018). *Data analysis and visualization using python.* Springer.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* O'Reilly Media.

Kedia, A., & Rasu, M. (2020). *Hands-on python natural language processing: Explore tools and techniques to analyze and process text with a view to building real-world NLP applications.* Packt Publishing Ltd.

McKinney, W. (2012). *Python for data analysis: Data wrangling with pandas, NumPy, and IPython.* " O'Reilly Media, Inc.".

Sarkar, D. (2016). *Text analytics with python.* Springer.

Szeliski, R. (2010). *Computer vision: Algorithms and applications.* Springer Science & Business Media.

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data.* " O'Reilly Media, Inc.".