PhD Python Programming Course

Limperg - 2022

Instructor:	Ties de Kok | University of Washington	**Date:**	13-6 to 17-6
Email:	tdekok@uw.edu	**Place:**	Tilburg - Cube 223

Workshop Page:

All course-specific materials are made available through Github.
Github Repository: *Limperg Python Github Repository*

Main Resources:

This course uses the following two resources as core foundation:

- Ties de Kok, *Learn Python for Research*, GitHub
- Ties de Kok, *Python Natural Language Processing (NLP) Tutorial*, GitHub

Objectives:

This programming course is designed to introduce the participants to the basic principles needed to use Python for Accounting research. We will discuss the following core elements: an efficient Python workflow, the Python programming language, Python for data-handling, Python for gathering data from the web, Python for natural language processing (NLP), and various miscellaneous topics.

At the end of the programming course, an active participant should be comfortable to:

- set up a workflow to efficiently incorporate Python into their projects,
- comprehend and implement basic Python programming operations,
- use `Pandas` and `Numpy` for basic data handling tasks,
- execute basic web scraping tasks using `Requests` and `Requests-HTML`,
- process and analyze text documents using common Python NLP packages,
- perform basic analyses on disclosure documents such as EDGAR fillings,
- incorporate version control into their Python workflow using Git and Github.

Deliverables and grading:

The course consists of 4 deliverables, one per modules 1 to 4. These deliverables are required to be handed in order to complete the course and obtain credit. The deliverables are due on **July 3th 2022** (2 weeks after our last class). You can hand them in by emailing them to *tdekok@uw.edu*.

Prerequisites:

Prior knowledge of the Python programming language is not required to participate in this course.

☞ You will need to bring your own laptop and follow the `Setup Instructions` before the first class.

Module descriptions:

There are 5 modules. Each module consists of (1) a lecture recording, (2) an in class demonstration, and (3) a problem set. You will watch the recording asynchronously before class and work on the demonstration and problem set during class time. The daily end time is 4pm, however, I will stick around until around 5pm to help those that want to work longer.

Module 1: Python introduction
Class: Monday - June 13th - 10:30am to 4pm/5pm
Lecture recording - part 1: `Part 1: Python Basics`
Lecture recording - part 2: `How to use Jupyter Lab / Jupyter Notebooks`

- Structure of the programming course
- Python Programming Language
- Python eco-system
- Using Python
- Jupyter Notebook
- Python syntax

Module 2: Data handling using Pandas
Class: Tuesday - June 14th - 10:30am to 4pm/5pm
Lecture recording: `Part 2: Handling Data with Pandas`

- Introduction to Pandas
- Opening / Closing various file types
- Basic Pandas operations
- Basic visualizations

Module 3: Gathering data from the web
Class: Wednesday - June 15th - 10:30am to 4pm/5pm
Lecture recording: `Part 3: Gathering data from the web`

- Terminology / Ethics / Tools
- Interacting with an API
- Web scraping a page
- Reverse-engineer HTTP requests
- Browser automation with Selenium

Module 4: Natural Language Processing
Class: Thursday - June 16th - 10:30am to 4pm/5pm
Lecture recording: `Part 4: Natural Language Processing`

- What is NLP / Textual Analysis
- Terminology / Tools

- Processing and Cleaning text
- Direct feature extraction (Regular expressions / dictionary counting)
- Representing text numerically
- Machine learning

Module 5: Tools for Reproducible Research
Class: Friday - June 17th - 10:30am to 12:45pm + We can chat about your projects in the early afternoon.
Lecture recording: `Part 5: Best Practices`

- Version control with GitHub
- Best practices when programming
- Using Jupyter with Stata and/or R
- Speed up code with multi-processing
- Running code remotely on a server