PhD Python Programming Course

Limperg - 2021

Instructor:	Ties de Kok | University of Washington	**Date:**	21-6 to 25-6
Email:	tdekok@uw.edu	**Place:**	Online

Workshop Page:

All course-specific materials are made available through Github and Discord.
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 11th 2021** (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.

Module descriptions:

The course consists of 5 modules, below is a short overview for each module.

Each module consists of (1) a lecture recording, (2) a demonstration recording, and (3) a problem set. You will watch the recordings asynchronously and you will work on the problem set on the respective day via Discord.

Module 1: Python introduction
Live session: Monday - June 21st - 6pm to 9pm

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

Module 2: Data handling using Pandas
Live session: Tuesday - June 22 - 6pm to 9pm

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

Module 3: Gathering data from the web
Live session: Wednesday - June 23 - 6pm to 9pm

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

Module 4: Natural Language Processing
Live session: Thursday - June 24 - 6pm to 9pm

- 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
Live session: Friday - June 25 - 6pm + Happy Hour

- 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