Limperg Python Programming Course

June 2020

Instructor:	Ties de Kok | University of Washington	**Date:**	16 to 20 June 2020
Email:	tdekok@uw.edu	**Place:**	Tilburg University

Workshop Page:

All course-specific materials are made available through a companion repository hosted on GitHub.

This repository is located here: *Limperg Python Programming Course repository*

Main Resources:

This course uses the following two resources as core foundation:

- Ties de Kok, *Learn Python for Research*, GitHub, 2018.
- Ties de Kok, *Python Natural Language Processing (NLP) Tutorial*, GitHub, 2018.

Additional Resources:

- Al Sweigart, *Automate the boring stuff with Python* (Free HTML version), No Starch Press, 2015.

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. Each element will be introduced by a lecture and demonstration in the morning followed by a hands-on session in the afternoon where the participants will work on a mini-task relating to the materials introduced in the morning.

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.

Prerequisites:

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

☞ It is required to bring your own laptop, check the end of this syllabus!

Session descriptions:

Below a short overview of the content that we will discuss during each of the sessions.

Each session will encompass a whole day, on Saturday we will end a bit earlier. In the morning I will give an introductory lecture and a demonstration, in the afternoon you will get hands-on experience based on the material introduced in the morning. All slides and materials will be made available on GitHub.

Day 1 (Tuesday, 16-5-2020): Python introduction

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

Day 2 (Wednesday, 17-5-2020): Data handling using Pandas

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

Day 3 (Thursday, 18-5-2020): Gathering data from the web

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

Day 4 (Friday, 19-5-2020): 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

Day 5 (Saturday, 20-5-2020): Tools for Reproducible Research

- 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