Limperg Python Programming Course

March 2019

Instructor:	Ties de Kok | Tilburg University	**Date:**	18 to 22 March 2019
Email:	t.c.j.dekok@uvt.nl	**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 Friday 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 (Monday, 18-3-2019): Python introduction

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

Day 2 (Tuesday, 19-3-2019): Data handling using Pandas

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

Day 3 (Wednesday, 20-3-2019): Gathering data from the web

- Terminology / Ethics / Tools
- Interacting with an API
- Web scraping a page
- Reverse-engineer HTTP requests
- Dealing with Javascript elements

Day 4 (Thursday, 21-3-2019): 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 (Friday, 22-3-2019): 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

Schedule overview

	Mo - Thu (18-3 to 21-3)	Friday (22-3)
09:00 - 10:00	09:30 – 12:00 Lecture (75 min) Demonstration (60 min) TBD	
10:00 - 11:00		10:00 – 12:00 Lecture (60 min) Demonstration (45 min) TBD
11:00 - 12:00		
12:00 - 13:00		
13:00 - 14:00	13:00 – 16:00 Mini tasks (180 min) TBD	13:00 – 14:30 Instructions assignment (30 min) QA (45 min) TBD
14:00 - 15:00		
15:00 - 16:00		
16:00 - 17:00		

Preparation | hardware:

Large parts of the course involve so-called “mini tasks”, these hands-on parts require a personal computer. For the instructions I will assume that you are using the Windows operating system, however, it should be no problem to participate with a computer running Mac OS or any of the Linux distributions.

Preparation | software:

We will be using the Python 3.7 version of the Anaconda Distribution as a starting point. The `Anaconda Distribution` is the most convenient way to get started with Python for data science purposes as it makes it easy to install, run, and upgrade a comprehensive Python environment.

☞ We will be using Python 3 exclusively, however, I will include a note whenever an important difference between Python 3 and Python 2 comes up.

Step 1: Install Anaconda on Windows/macOS/Linux:

Please make sure that you have the Python 3.7 Anaconda Distribution installed on your computer (3.5 or 3.6 is also fine). Downloads are available here: Anaconda Distribution

☞ Not all Python packages/libraries that we will be using come pre-installed with Anaconda. Please follow step 2 to install all the necessary packages.

Step 2: Install additional requirements:

Installing each package manually is tedious and prone to errors, a better approach is to create a new `Conda` environment with the provided `environment.yml` file.

Please follow these steps:

1. Download the `environment.yml` file to your system: `download environment.yml`
2. Open a command prompt / shell and `cd` (change dir) to the folder containing the `environment.yml`
3. Run the following command: `conda env create -f environment.yml`

   ☞ Installing everything will take a while.
4. Activate the `limperg-python` environment by typing:
   - `activate limperg-python` on Windows
   - `source activate limperg-python` on Mac OS or Linux.

Note, if you want to use Spacy, NLTK, and/or Textblob then it is important to also download the corresponding language models. Without the language model these packages will not be very useful.

Install them as follows:

☞ I can help you during the first day to get everything setup if you run into problems.

- NLTK (Link to docs)

  In a Jupyter Notebook run:

```python
import nltk
nltk.download()
```

- TextBlob (Link to docs)

  In the command line / terminal run:

  ```
  python -m textblob.download_corpora
  ```

- Spacy (Link to docs)

  ☞ If you installed using `requirements.yml` you can skip this step as the Spacy models are included.

  In the command line / terminal run:

  ```
  python -m spacy download en
  ```

Text editor: We will primarily be using the `Jupyter Notebook` as our Python interface, this only requires a browser. However, it would be convenient to also have a basic text editor installed. For Windows I recommend installing `Notepad++` as a good first basic editor.

Complete overview of all additional packages:

☞ You don't need to run the commands below if you followed the steps above!

```
$ conda install spacy
$ conda install textblob
$ conda install nltk
$ conda install tqdm
$ conda install deepdish
$ conda install xlrd
$ conda install openpyxl
$ conda install pytables
$ conda install qgrid
$ pip install pyldavis
$ pip install fuzzywuzzy
$ pip install requests-html
$ pip install https://github.com/explosion/spacy-models/releases/download/
   en_core_web_sm-2.0.0/en_core_web_sm-2.0.0.tar.gz#en_core_web_sm
```