# UW Python Programming Camp

Summer 2018

| | | | |
|---|---|---|---|
| **Instructor:** | Ties de Kok \| Tilburg University | **Date:** | June, July 2018 |
| **Email:** | t.c.j.dekok@uvt.nl | **Place:** | University of Washington |

### Workshop Page:

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

This repository is located here: *UW Python Programming Camp repository*

### Main Resources:

This workshop 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.
- Brandon Rhodes, *PyCon Pandas Tutorial* (GitHub page, Video), 2015.

### Objectives:

This programming camp is primarily designed to introduce the participants to the basic principles needed to use Python for Accounting and Finance research. We will discuss the following core elements: an efficient Python workflow, Python for data-handling, Python for gathering data from the web, using Python for natural language processing (NLP), handling SEC EDGAR filings with Python, and various miscellaneous topics. Each element will be introduced by a brief lecture and demonstration followed by a hands-on session where the participants will work on a mini-task relating to that element.

At the end of the programming camp, 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.

### Prerequisites:

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

☞ Make sure to prepare your laptop accordingly, check the end of this syllabus!

# Programming camp, week 1 (25-26 June 2018)

| | Monday (6-25-18) | Tuesday (6-26-18) |
|---|---|---|
| 09:00 - 10:00 | | |
| 10:00 - 11:00 | | |
| 11:00 - 12:00 | | |
| 12:00 - 13:00 | | |
| 13:00 - 14:00 | | |
| 14:00 - 15:00 | 14:00 – 16:30<br>Python introduction (45 min)<br>Demonstration (30 min)<br>Mini task (75 min)<br>PACCAR 490 | 14:00 – 16:30<br>Pandas introduction (30 min)<br>Demonstration (15 min)<br>Mini task (105 min)<br>PACCAR 490 |
| 15:00 - 16:00 | | |
| 16:00 - 17:00 | | |

# Programming camp, week 2 (2-3 July 2018)

| | Monday (7-2-18) | Tuesday (7-3-18) |
|---|---|---|
| 09:00 - 10:00 | | |
| 10:00 - 11:00 | | |
| 11:00 - 12:00 | | |
| 12:00 - 13:00 | | |
| 13:00 - 14:00 | | |
| 14:00 - 15:00 | 14:00 – 16:30<br>Web scraping introduction (45 min)<br>Demonstration (15 min)<br>Mini task (90 min)<br>PACCAR 490 | 14:00 – 16:30<br>NLP introduction (60 min)<br>Demonstration (15 min)<br>Mini task (75 min)<br>PACCAR 490 |
| 15:00 - 16:00 | | |
| 16:00 - 17:00 | | |

# Programming camp, week 3 (9-10 July 2018)

| | Monday (7-9-18) | Tuesday (7-10-18) |
|---|---|---|
| 09:00 - 10:00 | | |
| 10:00 - 11:00 | | |
| 11:00 - 12:00 | | |
| 12:00 - 13:00 | | |
| 13:00 - 14:00 | | |
| 14:00 - 15:00 | 14:00 – 16:30<br>EDGAR walk-through (60 min)<br>Hands-on (90 min)<br>PACCAR 490 | 14:00 – 16:30<br>Version Control (GitHub)<br>Jupyter with Stata or R<br>Running code on server<br>Amazon Mechanical Turk → Python + Javascript<br>PACCAR 490 |
| 15:00 - 16:00 | | |
| 16:00 - 17:00 | | |

## Session descriptions:

Below a short overview of the content that I will discuss during each of the 6 sessions. Most of the sessions are standalone as long as you have a basic understanding of Python and the Jupyter Notebook (either from prior knowledge or by having attended the first session).

Each session will consist of roughly 45 minutes of introduction, a brief demonstration, and a mini-task to get hands-on experience. The slides used for each introduction will be made available on GitHub.

### Session 1 (6-25-2018): Python introduction

☞ Recommended prior sessions: None

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

### Session 2 (6-26-2018): Data handling using Pandas

☞ Recommended prior sessions: 1

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

### Session 3 (7-2-2018): Gathering data from the web

☞ Recommended prior sessions: 1

- How does the web work?
- Terminology / Ethics / Tools
- Interacting with an API
- Web scraping a page
- Reverse-engineer HTTP requests
- Dealing with Javascript elements

## Session 4 (7-3-2018): Natural Language Processing

☞ Recommended prior sessions: 1

- What is NLP / Textual Analysis
- Terminology / Tools
- Processing and Cleaning text
- Direct feature extraction (Regular expressions / dictionary counting)
- Representing text numerically
- Machine learning

## Session 5 (7-9-2018): EDGAR walk through

☞ Recommended prior sessions: 1, 2, 3, 4

- Obtain an index of EDGAR filings
- Download EDGAR filing
- Process HTML of EDGAR filing
- Extract section from EDGAR filing
- Calculate metrics based on text of filing

## Session 6 (7-10-2018): Miscellaneous topics

☞ Prior sessions that are required: None

- 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
- Amazon Mechanical Turk using Python and Javascript

## Preparation | hardware:

Large parts of the workshop 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 3.6 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 3.6 Anaconda Distribution installed on your computer. 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 `uw-python-camp` environment by typing:
   - `activate uw-python-camp` on Windows
   - `source activate uw-python-camp` 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 camp 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 textacy
$ 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
```