P C Q R

Potsdam Center for Quantitative Research

http://www.uni-potsdam.de/pcqr

Workshop

Python for text mining

LECTURER
Dr. Nikos Askitas
(IZA Bonn)

DATE AND TIME
June 14, 2018 – 9 am to 5 pm
June 15, 2018 – 9 am to 3 pm

VENUE
University of Potsdam
Griebnitzsee Campus
House 7, room 241

About the Workshop

Python, originally a language for the web, is now a prime statistical language, sporting a rich collection of diverse modules that include regressions, machine learning, all kinds of stats, supreme graphing, agent based simulations etc. According to the TIOBE Index (https://www.tiobe.com/tiobe-index), as of February 18, Python is the fourth most popular programming language, being first among scripting languages. In comparison Stata ranks somewhere between 50 and 100. According to the World Economic Forum, Python ranks among the top skills that the world tech giants require by both engineers and data scientists.
As more and more markets (marriage market, transport market, labor market, etc) move online or are born exclusively online, our ability to study markets and understand socioeconomic phenomena will depend on being able to leverage the internet as a data source. This means data and text mining will be an important skill for social scientists. In recognition of this fact the European parliament is working on excluding data and text mining from future digital copyright legislation. The course covers the basics of Python selectively, depending on which language elements are necessary for the examples. The core aim is to study:

- Hit the limits working with Stata's built in rudimentary web browser and regular expressions.
- The basics of how to install and manage a python installation and its modules.
- How to construct and brand a web browser in Python.
- How to use it to download pages from the web and store them.
- How to use regular expressions (module: re) to harvest data out of html documents.
- The data types Python provides for storing data (module: pandas).
- Some graphing, basic regressions with Python etc.
- Integration of python with Stata.

Prerequisites

Some familiarity with programming concepts (e.g. Stata) is assumed. To access course material, a NextCloud Client is necessary (installation through https://nextcloud.com/install/ and log-in via https://cloud.iza.org). The Anaconda-Navigator we will use as well as Python itself is available through https://www.anaconda.com/download.

Registration

Please register for participation by writing an E-Mail to pohle@empwifo.uni-potsdam.de by June 7, 12 am at the latest. The maximal number of participants will be 20. Students and researchers from all Universities in Brandenburg and Berlin are welcome, but PCQR members and their staff take precedence.

About The Lecturer

Nikos Askitas joined IZA in 2000 and he is the Institute's Coordinator of Data and Technology. As head of its Research Data Center (IDSC) and Head of its ICT unit he is responsible for all data and technology issues and is hence a member of both the research and the service units of IZA. On the one hand, working as a mathematician, data scientist and economist, he explores new avenues of research and policy in labor economics but also in social science more generally while on the other hand he oversees the Institute's ICT operations.
His mathematics research was published in such journals as *Manuscripta Mathematica*, *Mathematische Zeitschrift*, *Knot Theory and Its Ramifications*, *The Kobe Journal of Japan*, *Topology and Its Applications*. His economics research, theoretical and empirical, spans a wide range of topics and is published among others in the *International Journal of Manpower*, *Wirtschaftsdienst*, *Journal of Forecasting*, *Cityscape - Journal of Policy Development and Research* of the Office of Policy Development and Research at the US Dept. of Housing and Urban development, *AStA Wirtschafts - und Sozialstatistisches Archiv*, *PLoS ONE* etc.
His current research interests include web and big data, time series, forecasting and nowcasting, technology and labor, game theory, statistics, adaptive systems etc.