MQE/PhD Guest Course:

“Big Data and Machine Learning with Python”
Mon., 11 through to Fri., 15 March, 2019

Michael E. Rose, PhD Researcher,
Max-Planck Institute for Innovation and Competition, Munich

Course Outline

This course introduces Python to young empirical researchers. Applications cover data access and manipulation, plotting, machine learning and natural language processing. Participants furthermore work collaboratively with version control (git).

Prerequisites

Precourses: (Workload 5 to 10 hours)

- Self-paced, interactive online courses for Python
- At least one of the two:
  1. datacamp.com (“Intro to Python for Data Science”, Sessions 1, 2, 3)
  2. codeacademy.com (“Learn Python 2”, Sessions 1, 2, 3, 4, 5, 7, 8)

Teaching sessions

First session: Input/Output in Python and APIs

- Basic concepts in Python (recap from pre-course)
- Script design and coding workflow (script mode and terminal vs. notebook vs. IDE)
- Coding principles (scaffolding, debugging)
- Version control, introduce Git hands-on
- Styleguide PEP8

Second session:

- Pandas as key module for data wrangling in Python
- Data formats: csv, json, xml, pdf, html, plain text
- API with examples: google maps/geopy, Wikipedia, scopus
- Exercise: Obtaining data from Google Map API (provided) and output as csv

Third session: Plotting

- matplotlib
- gmplot
- pandas
- seaborn
- [plotly]

Excersises:

- Generating different plots from provided and self-collected data (citation time series, distributions of patents, …)
- Creating interactive plots on plotly

Fourth session: Supervised ML

- Difference between supervised and unsupervised learning
- Workflow: Splitting labelled data randomly, train, evaluate on test sample, change parameters
- k-nearest neighbors, classifier and regression version
- Linear models and regularization: ridge (L1), LASSO (L1)
- Decision Tree and Random Forest
- [Neural Networks]
- [dedupe]

Exercises:

- Various prediction tasks

Fifth session: Unsupervised ML

- Dimensionality reduction and clustering
- PCA and Scaling
- NMF
- k-means clustering and plots
- agglomerative clustering and dendrograms
- DBSCAN

Exercises:

- Clustering of house prices, sport outcomes
- Plotting of cluster algorithms
- PCA of financial time series

Sixth session: NLP

- Text extraction from pdfs (not for Windows)
- Word clouds
- Noun phrases and sentiment analysis with TextBlob
- Text cleaning: Stopwords, stemming, tokenization, n-grams
- Text to data: Vectorization
- Mathematics of LDA
- Evaluations of topic selection

Exercises:

- Sentiment analysis of speeches of central bankers before, during and after the crisis
- Text similarity of scientific abstracts by bag of words and plotting on heat map
- Small-scale LDA of patents and plotting distribution of topics by document

Optional session: Networks

- Network analysis with networkx
- Exercise: Creating a network from patents’ fields (patents are linked when they list the same fields) to derive combinatorial newness with graph theory

Time and Place

The course will take place in Kaulbachstr. 45, Room 006.
The lecture starts on Mon, **March 11 at 9:10am**.
Timetable: 9:10am- 4:50pm; breaks every 50 min for 10 min, lunch break: 12am-2pm.

Literature (reference, not prerequisite)

- Downey, Allen B.: "How to think like a Computer Scientist"
- Shapiro, J. and M. Gentzkow: "Code and Data for the Social Sciences: A Practitioners Guide"
- Mueller, A. and Sarah Guido: "Introduction to Machine Learning with Python"
- Gentzkow, M., B. Kelly and M. Taddy: "Text as Data"

Evaluation

The course is not graded: pass y/o on the basis of the exercises.
A certificate of attendance will be issued.

Registration

Please register for the course by sending an email to mgse-master@econ.lmu.de
by **Mon., Jan. 14, 2019.**
Doctoral students interested in participating are encouraged to attach a short motivation letter (max. 1 page) detailing the importance of the course for their research.
We will inform you about your participation in the course by **Mon., Jan. 28, 2019.**

Additional Information:

The class work is interactive, please bring your laptop.
For further questions please contact: Regine Reichenbach, Graduate Master Office,
mgse-master@econ.lmu.de, Tel.: (089) 2180-6951.

About the Lecturer:

Michael E. Rose, PhD Researcher at MPI for Innovation and Competition (Harhoff group)

- Experience in teaching ML and Big Data in Python to PhD students, Computational Mathematics in Matlab to Master's students, Time Series Econometrics in Python to Master's students, SQL, Excel and VBA to Master's students
- Daily usage of Python, ML in own research
- Lead development of two python packages (scopus and sosia)