Data Science Institute – Day 1

Introduction to Python
Data Analytics

June 5\textsuperscript{th}, 2017
Kang P. Lee
ITS-RS / UI3
Course Outline

• Morning
  
  • **Session 1: Basic Concepts** (60 minutes, 09:00-10:00am)
  
  • Break (15 minutes, 10:00-10:15am)
  
  • Session 2: Demonstration of Python Data Analytics (105 minutes, 10:15am-12:00pm)

• Lunch Break (90 minutes, 12:00-1:30pm)

• Afternoon
  
  • Session 3: Hands-On Practice (120 minutes, 1:30-3:30pm)
Goal & Scope of This Course

Learn how to use Python for data analysis, so that you can learn in depth for yourself after completing this course.
Goal & Scope of This Course

We’re going to cover only the key concepts in Python data analytics

• Python basics for data analytics
• Python data analytics libraries
• Jupyter Notebook
Quick Survey on Prior Experience

- **Python**
  - I have experience with Python
  - I have experience with programming, but not with Python
  - I have no experience with programming

- **Data analytics**
  - I have experience with data analytics
  - I have no experience with data analytics
Introduction to the Instructor

• Name: Kang Pyo Lee

• Motto: “Learn from data!”

• Education
  • Ph.D. in Computer Science from Seoul National University
    “A Keyword-Based Exploratory Search Framework Exploiting User Tagging”
  • M.S. in Computer Science from Seoul National University
    “RDF-Based Integration of Heterogeneous Biological Databases”

• Interests: big data, data science, social media analytics, market research, etc.

• Hobby: organizing things - everything from physical objects to digital files
Introduction to the Instructor

• Previous Work

  • Data scientist at Samsung Big Data Center
  
  • Developing business-critical insight by analyzing big data
    
    • Leading big data based market research projects (with 4 analysts) for gaining insight into customers and markets
    
    • Leading a project (with 10 engineers and analysts) that designed and developed the social media analytics service for Samsung employees
    
    • Instructing employees who consider adopting social media analytics to their work
Introduction to the Instructor

• Current Work

• Staff data scientist at ITS-RS and UI3

• Supporting various research on campus on big data and social media analytics

  • Taking the role of a data analytics consultant with a primary focus on social media analytics

  • Working as a facilitator for the Social Media Analytics Research Topic Group

  • Leading the Social Media Analytics Platform (SMAP) project
Data analytics is the process and methodology of analyzing data to draw meaningful insight from the data.
Why Is It So Popular?

We now see the limitless potential for gaining critical insight by applying data analytics
Typical Process of Data Analytics

1. Requirement Understanding
2. Data Understanding
3. Data Preparation
4. Data Exploration

The most time-consuming part
The most exciting part

Decision Making
Problem

Insight Development
Decision Making

Modeling & Evaluation
Deployment

Modeling Problem
Types of Data Analytics

Data Analytics

**Descriptive Analytics**
What has happened or is happening?
“How has the population been changing?”

**Predictive Analytics**
What could happen in the future?
“How will the population change over the next ten years?”

**Prescriptive Analytics**
What should we do to make that happen or not happen?
“What actions should be taken in order to avoid the demographic cliff?”
Confusion – Data Analysis vs. Data Analytics

• They’re often used interchangeably, but technically speaking...

Data Analysis

Vs.

Data Analytics

Individual acts of analyzing data

Process & methodology for analyzing data
Confusion – Big Data vs. Data Analytics

• What they have in common is that both refer to data, but technically speaking...

Big Data

Vs.

Data Analytics

Focuses on issues with handling non-traditional “big” data

Focuses on gaining meaningful insight regardless of the size of the data
Confusion – Machine Learning vs. Data Analytics

AI

Machine Learning Vs. Data Analytics

Deep Learning

Data analytics depends heavily on machine learning
Confusion – AI vs. Data Analytics

AI Vs. Data Analytics

Machine Learning

Deep Learning

The goals are different!
- AI: intelligence
- Data analytics: insight
Python is a general-purposed high-level programming language

- Web development
- Networking
- Scientific computing
- Data analytics
- ...

Python as a Programming Language
The nature of Python makes it a perfect-fit for data analytics

- Easy to learn
- Readable
- Scalable
- Extensive set of libraries
- Easy integration with other apps
- Active community & ecosystem
## Popular Python Data Analytics Libraries

<table>
<thead>
<tr>
<th>Library</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>numpy, scipy</td>
<td>Scientific &amp; technical computing</td>
</tr>
<tr>
<td>pandas</td>
<td>Data manipulation &amp; aggregation</td>
</tr>
<tr>
<td>mlpy, scikit-learn</td>
<td>Machine learning</td>
</tr>
<tr>
<td>theano, tensorflow, keras</td>
<td>Deep learning</td>
</tr>
<tr>
<td>statsmodels</td>
<td>Statistical analysis</td>
</tr>
<tr>
<td>nltk, gensim</td>
<td>Text processing</td>
</tr>
<tr>
<td>networkx</td>
<td>Network analysis &amp; visualization</td>
</tr>
<tr>
<td>bokeh, matplotlib, seaborn, plotly</td>
<td>Visualization</td>
</tr>
<tr>
<td>beautifulsoup, scrapy</td>
<td>Web scraping</td>
</tr>
</tbody>
</table>
iPython is a Python command shell for interactive computing

Jupyter Notebook (the former iPython Notebook) is a web-based interactive data analysis environment that supports iPython
Comparison – R vs. Python

• Comparison between R and Python has been absolutely one of the hottest topics in data science communities

R Vs. Python

R came from the statisticians community, whereas Python came from the computer scientists community

Python is said to be a challenger against R, but in general it’s a tie

It’s up to you to choose the one that best fits your needs

For detailed comparison, refer to https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis
Comparison – Other Data Analytics Tools vs. Python

IBM SPSS, SAS, RapidMiner, KNIME... Vs. Python

• Commercial
• Graphical user interface (GUI)
• Easy & convenient

• Open source
• Command line interface (CLI)
• Inconvenient, but powerful
# Data Analytics Settings for This Course

<table>
<thead>
<tr>
<th>Component</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Python Version</td>
<td>Python 3 (vs. Python 2)</td>
</tr>
<tr>
<td>Data Analytics Environment</td>
<td>Jupyter Notebook</td>
</tr>
<tr>
<td>Data Analytics Software Toolkit</td>
<td>Anaconda (vs. Enthought Canopy)</td>
</tr>
<tr>
<td>Data Analytics Libraries</td>
<td>NumPy &amp; Pandas for data analysis</td>
</tr>
<tr>
<td></td>
<td>Plotly for visualization</td>
</tr>
</tbody>
</table>
Course Outline

• Morning
  • Session 1: Basic Concepts (60 minutes, 09:00-10:00am)
  • Break (15 minutes, 10:00-10:15am)
  • Session 2: Demonstration of Python Data Analytics (105 minutes, 10:15am-12:00pm)

• Lunch Break (90 minutes, 12:00-1:30pm)

• Afternoon
  • Session 3: Hands-On Practice (120 minutes, 1:30-3:30pm)
Summary

• Typical Python data analytics process for beginners
  1. Identify the dataset of interest from a file/database/web
  2. Load the dataset into a Pandas dataframe
  3. Check the column names and see the first few rows
  4. Derive additional columns if needed and handle missing data
  5. Do analysis with visualization or apply advanced data analytics techniques
What We’re Going to Do at Hands-On Sessions

• You’re going to analyze your own dataset yourself following the same process you’ve learned in the morning sessions
  → Don’t forget to bring a CSV file

• Make sure your Jupyter Notebook is working well

• Download all the materials for this course from the Data Science Institute website (https://uiowa.edu/datascience/resources)

• Mr. Mingrui Liu will be serving as a TA
Course Outline

• Morning
  • Session 1: Basic Concepts (60 minutes, 09:00-10:00am)
  • Break (15 minutes, 10:00-10:15am)
  • Session 2: Demonstration of Python Data Analytics (105 minutes, 10:15am-12:00pm)

• Lunch Break (90 minutes, 12:00-1:30pm)

• Afternoon
  • Session 3: Hands-On Practice (120 minutes, 1:30-3:30pm)