Data Wrangling with Python

Creating actionable data from raw sources

Elevator Pitch

Data is the new oil but it comes as crude, just like oil. To do anything meaningful - modeling, visualization, machine learning, for predictive analysis – you first need to wrestle and wrangle with data. This course teaches the essential basics of data wrangling using Python.

Key Features

- Focuses on essential basics of wrangling to get you up and running with analysis in no time
- Teaches the tricks and know-how of “how to solve data wrangling problems”
- Added bonus topics – random data generation, data integrity checks

Short Description

Data science is the ‘coolest job of 21st century’ (till the time Skynet takes over the world). But for all the emphasis on ‘Data’, it is the ‘Science’ that makes you - the practitioner - valuable. To practice high-quality science with data, first you need to make sure it is properly sourced, cleaned, formatted, and pre-processed. This course teaches you the most essential basics of this invaluable component of the data science pipeline – data wrangling.

Extended Description

“Data is the new Oil” and it is ruling the modern way of life through incredibly smart tools and transformative technologies. But oil does not come out in its final form from the rig. It has to be refined through a complex processing network. Similarly, data needs to be curated, massaged and refined to be used in intelligent algorithms and consumer products. This is called “wrangling” and (according to Forbes) all the good data scientists spend almost 60-80% of their time on this, each day, every project. It involves scraping the raw data from multiple sources (including web and database tables), imputing, formatting, transforming – basically making it ready, to be used flawlessly in the modeling process.

This course aims to teach you all the core ideas behind this process and to equip you with the knowledge of the most popular tools and techniques in the domain. As the programming framework, we have chosen Python, the most widely used language for data science. We work through real-life examples, not toy datasets. At the end of this course, you will be confident to handle a myriad array of sources to extract, clean, transform, and format your data for the great machine learning app you are thinking of building. Hop on and be the part of this exciting journey.

Approach

This course takes a practical approach to equip beginners with the most essential tools in the shortest possible time. The lessons start with absolute basics of Python focusing mainly on data structures, and then quickly jump into NumPy and Panda libraries as fundamental tools of data wrangling. It emphasizes why you should

stay away from traditional way of data cleaning, as done in other languages, and take advantage of specialized pre-built routines in Python. Thereafter, it covers how using the same Python backend, one can extract and transform data from diverse array of sources - internet, large database vaults, or Excel financial tables. Further lessons teach how to handle missing or wrong data, and reformat based on the requirement from the downstream analytics tool. The course emphasizes learning by real example and showcases the power of inquisitive mentality and imaginative mind for success.

Learning Objectives

- Able to manipulate complex and simple data structure using Python and it's built-in functions
- Use the fundamental and advanced level of Pandas DataFrames and numpy.array. Manipulate them at run time.
- Extract and format data from various formats (textual) – normal text file, SQL, CSV, Excel, JSON, and XML
- Perform web scraping using Python libraries such as BeautifulSoup4 and html5lib
- Perform advanced string search and manipulation using Python and RegEX
- Handle outliers, apply advanced programming tricks, and perform data imputation using Pandas
- Basic descriptive statistics and plotting techniques in Python for quick examination of data
- Practice data wrangling and modeling using the random data generation techniques - Bonus Topic

Audience

Software professionals, web developers, database engineers, and business analysts who want to move towards a career of full-fledged data scientist/analytics expert or whoever wants to use data analytics/machine learning to enrich their current personal or professional projects.

Prior experience with Python is not an absolute requirement, however the knowledge of at least one object-oriented programming language (e.g. C/C++/Java/JavaScript), and high school level math is highly preferred. It is a bonus if you have rudimentary idea about relational database and SQL.

Even seasoned Python app/web developers can benefit from this course as it focuses on data engineering aspects, which are not generally used web-based programming.

Author Biography

Dr. Tirthajyoti Sarkar works as a Sr. Principal Engineer in semiconductor technology domain where he applies cutting-edge data science/machine learning techniques for design automation and predictive analytics. He writes regularly about Python programming and data science topics. He holds a Ph.D. from the University of Illinois and certifications in Artificial Intelligence and Machine learning from Stanford and MIT.

Shubhadeep Roychowdhury works as a Sr. Software Engineer at a Paris based Cyber Security startup where he is applying the state-of-the-art Computer Vision and Data Engineering algorithms and tools to develop cutting edge product. He often writes about Algorithm implementation in Python and similar topics. He holds a Master Degree in Computer Science from West Bengal University Of Technology and certifications in Machine Learning from Stanford. He lives in Paris with his wife and kid.

Outline

Day One

We will start by discussing the background of data wrangling and the importance of it in data science, which will serve as the main motivation behind this course. Afterwards, the students will be introduced to the tools and libraries that we will be using throughout the course, for example, Jupyter/IPython notebooks, virtualenv in Python, installing and testing numpy, pandas, beautifulsoup4, and matplotlib in a virtualenv. Once they are familiar with the tools of the trade, students will spend the rest of the day getting familiar with basic Python and data structures using Python. We will end the day by introducing elementary operations using Pandas and NumPy array, which includes descriptive statistics (central tendency, SD, variance, and so on).

At the end of this day, students should be able to:

- Write basic programs in Python involving fundamental data structures and having idea about Pandas and Numpy
- Advanced operations with Python Data Structure.

Lesson One: Introduction to Data Structure using Python

This is the starting point of our course. We will learn about the necessity of data wrangling and set up our environment for the rest of the course.

- Topic 1: An Introduction to Data Wrangling (10 min)
- Topic 2: Lists, Sets, Strings, Tuples, and Dicts (20 min lecture + 10 minutes exercise)
- **Activity 1:** Create a small script to take inputs from a user, apply some basic rules we learned, and output them on the screen (30 minutes)

Lesson Two: Advanced Operations on Built-In Data Structure

In this lesson, the students will learn about the basics of coding in Python and end up experimenting with the basic data structures and how to manipulate them.

- Topic 1: Advanced Operation on Lists, Sets, Dicts, Tuples - List Comprehension, Map, Filter, Iterators (20 minutes + 10 minutes exercise)
- Topic 2: Opening files from Python - ‘r’ and ‘rb’, Reading and Writing of data, Basic OS funcs (10 minutes + 10 minutes exercise, play around with different file formats)
- **Activity 1:** Write user inputs to a file line by line at the same time applying the operations you learned on them (30 minutes)

Lesson Three: Introduction to Numpy, Pandas, and Matplotlib

In this lesson, the students will learn the fundamentals of three of the four necessary libraries we will need in the course.

- Topic 1: np.array() - What is it and how to use it. Creating and using different tensors (arrays with various dimensions) (10 minutes + 10 minutes exercise)
- Topic 2: pd.DataFrame. What is a DataFrame and how does it work. Creating a DataFrame (20 minutes + 10 minutes)

- Topic 3: Refresher of basic descriptive statistics. (10 minutes)
- Topic 4: Using NumPy and Pandas to calculate basic descriptive statistics on the DataFrame. (15 minutes)
- **Activity 1:** Extracting data from CSV, creating a DataFrame, and extracting some descriptive statistics. (30 minutes)

Day Two

We continue our journey into data wrangling in Day Two. At the end of this day, students will be confident about handling different types of data sources. They will also have a much deeper understanding of the working of pandas DataFrame and how to use it effectively to do finer grained manipulation on data. They will learn to handle null or absent data, indexing, filtering, bucketing, grouping, and several other kind of operations that spans the breadth of the subject. Finally, they will be introduced to some tricks of data cleaning and formatting. We assume a basic familiarity in programming, statistics, necessary libraries, and tools for the Day Two course.

At the end of this day, students should be able to:

- Confidently handle DataFrames and perform several different operations on it
- Read and handle data from different sources
- Perform some well-known programming tricks on data cleaning

Lesson Four: Deep Dive into Data Wrangling with Python

We build on top of where we left off on Day One. Here, the student will learn about several different operations involving pandas DataFrame and NumPy array.

- Topic 1: Slicing, Filtering, Grouping, Bucketing in Pandas. (20 + 10)
- Topic 2: Boolean Indexing and Subsets. Why do we use them and not the pure Python ‘and’ or ‘or’ operators? (10 + 10)
- Topic 3: Handle Missing Data: dropna, imputations (20 + 10)
- **Activity 1:** Use the Adult Income Dataset from UCI to practice various Pandas operation (45)
- **Activity 2:** Use Pima Indian Dataset to examine and perform data imputations and cleaning (45)

Lesson Five: Get Comfortable with Different Kind of Data Sources

Students learn how to handle different kind of data sources. In real life you always need this knowledge. This involves text based and web based.

- Topic 1: How to read data from different text based (and non-text based) sources (15 mins)
    - How to read and parse data from CSV, JSON. (15 mins + 5 mins)
    - Can you find out a way to read tabular data from PDF? Which library do you use? How do you install it? (15 mins + 5 mins)
- Topic 2: Reading data from the web. What is the structure of a web page? How to use BeautifulSoup to read data from a webpage (20 mins + 10 mins)
- **Activity 1:** Reading tabular data from multiple web pages on the disk, that is, save them as Excel files. (20 mins)
- **Activity 2:** Reading data from a PDF and exporting it as HTML (20 mins)

Lesson Six: Learning the Hidden Secrets of Data Wrangling

Students learn some standard tricks to handle different type of data. And also to format them better.

- Topic 1: Advanced list comprehensions and zip. (20 mins + 10 mins)
- Topic 2: Using “format” statement to format data (10 mins + 5 mins)
- Topic 3: Identify and clean outliers (30 mins)
- **Activity 1:** Given a messy data frame, use the knowledge you acquired on it and come up with a nicely formatted data frame. What kind of outliers are there? What are their effect on the data? Can you clean them? (45 mins)

Day Three

On the final day of our study we will learn some advanced topics related to data gathering and wrangling and finally we will apply our knowledge to some practical, real life data sets (some famous ones!) and present our findings as form of reports. Students will also have a concluding lecture which gives them a clear picture of what comes next. At the end of the day, they will leave the training with all the tools and tricks in their arsenal to start working on messy and real life data and make it a clean and nice thing to work with.

Lesson Seven: Advanced Web Scraping and Data Gathering

Students learn advanced techniques and methods to get and clean data from various kinds of web pages, XML and APIs.

- Topic 1: Advanced web scraping and dealing with APIs: Use “requests” and “Beautifulsoup” to read various different web pages and gather data from them. What do you learn about xpath? (20 mins + 15 mins)
- Topic 2: Read Data from XML (10 mins + 15 mins)
- Topic 3: Read data from an API. Can you make different kind of requests like get, post etc.? What about custom headers? (30 mins + 10 mins)
- Topic 4: Introduction to regex: Use regex in tricky string matching which would have been very clumsy to write otherwise. (15 mins + 15 mins)
- **Activity 1:** A full-scale web scraping utilizing the knowledge from above including regex to extract key information. Create key plots from this data. (45 mins)
- **Activity 2:** Build your own mini IMDB like database of movies and actors by reading JSON data stream using an API. (60 mins)

Lesson Eight: RDBMS and SQL

Students learn the basics of RDBMS and SQL and get some data from there

- Topic 1: Refresher of RDBMS and SQL (10 mins)
- Topic 2: Use a RDBMS (MySQL/PostgreSQL/SQLite) and write some SQL (10 mins+ 15 mins)
    - Create and Populate a DB
    - Create Python code to connect with it.
    - Write some SQL to read data from there.
- **Activity 1:** Process a MySQL database (on your disk) from your notebook. Read, filter, analyze, and visualize data sets. Join/merge/create new tables and write back to the database. (45 mins)

Lesson Nine: Application in real life and Conclusion of course

Students will apply the gathered knowledge in a real life data set and investigate various aspects of it. They will also have a concluding lecture to tell them what comes next

- Topic 1: Apply your knowledge
    - **Activity 1:** Obtain a data set. Create a Data Frame and display some basic stats on it. Use different techniques on the DataFrame to wrangle with it. Tell a story. Visualization is not needed but will be nice addition to your story. Write a report (textual format) about the data wrangling aspects. (1h 30 min)
- Topic 2: Concluding Lecture - What comes next? (20 mins -30 mins)

Hardware and Software Requirements

Hardware Requirements

For an optimal student experience, we recommend the following hardware configuration:

- OS: Windows 7 SP1 64-bit, Windows 8.1 64-bit or Windows 10 64-bit, Ubuntu Linux, or the latest version of OS X
- Processor: Intel Core i5 or equivalent
- Memory: 4GB RAM (8 GB Preferred)
- Storage: 35 GB available space

Software Requirements

You'll also need the following software installed in advance:

- Browser: Google Chrome/Mozilla Firefox Latest Version
- Notepad++/Sublime Text as IDE (Optional, as you can practice everything using Jupyter notebook on your browser)
- Python 3.4+ (latest is Python 3.7) installed (from https://python.org)
- Python libraries as needed (Jupyter, Numpy, Pandas, Matplotlib, BeautifulSoup4, and so)

Access to installation instructions can be provided separately to course material for large training centers and organizations. All source code is publicly available on GitHub and fully referenced within the training material.

Primary Keywords

- Python
- Data wrangling
- Data analytics
- Data science
- Web scraping
- Pandas
- NumPy
- Matplotlib
- BeautifulSoup4