1 Course Description

This course introduces Python as a tool to collect, process and analyze large data sets from a variety of sources to create information that guides businesses decision-making. Students will get familiarized with Python as a language and as a platform to integrate different technologies and techniques for data analytics, including: (1) collection of online information; (2) tools and strategies for data storage; and (3) data analysis methods. Each class will start with the introduction of a concept or tool and end with in-class hands-on exercises using example datasets. Throughout the course students will apply these techniques to do their Homework and their Term Project.

2 Learning Objectives

Upon completion of this course, the student will be able to:

1. Use Python as a general-purpose programming language

2. Collect data available online in an automated fashion
   
   (a) Method 1: Scraping web sites
   
   (b) Method 2: Using sites’ APIs

3. Process and store data in the appropriate format for future analysis
   
   (a) Method 1: Files (e.g., CSV files)
   
   (b) Method 2: SQL databases (e.g., MySQL)
   
   (c) (optional) Method 3: NoSQL databases (e.g., MogoDB)

4. Apply data analytics tools to extract relevant information
   
   (a) Descriptive Statistics
   
   (b) Graphical Analysis
   
   (c) Linear Regression
   
   (d) (optional) Classification Trees
3 Source Materials

1. Textbooks:
   (a) Main: McKinney (2012), Python for Data Analysis, O’Reilly
   (b) Other: Russel (2011), Mining the Social Web, O’Reilly

2. Online references
   (a) Python 2 Documentation: https://docs.python.org/2/
   (b) pandas online reference: http://pandas.pydata.org/pandas-docs/stable/

(The instructor may assign additional readings from public sources / websites.)

4 Evaluation Method and Grading Scale

- Individual Assignments: 40%
- Term Project: 30%
- Final Exam: 30%

5 Individual Assignments

Assignments will be done by individual students and posted on Blackboard. Specific assignments will appear approx. 1 week prior to due date. Please monitor Blackboard for postings.

6 Term Project

The term-project will be done in 2 or 3 person teams and will involve the application of the methods mentioned in the class. Students will identify a question they would like to answer using publicly available data, gather the data from an online source, store it and analyze it using some of the methods shown in class. Students will produce a 5-10 page report identifying the questions, the gathered data and the results of the analyses and conclusions. Teams will be self-selected, and all team members will receive the same grade.

7 BlackBoard

Blackboard will be used for posting all lecture materials and assignments. You will also be asked to submit your assignments through Blackboard. Please check it regularly and familiarize yourself with navigation, downloading, uploading, etc.

8 Academic Conduct / Cheating, Plagiarism

We encourage discussion about individual assignments and the term project. However, the project that is submitted for grading must be the work of the 2-3 person team alone. Similarly, completed homework assignments must be your work alone. Specifically, discussion of results that are identical or nearly identical across projects will be regarded as cheating. Also, your answers on the final exam
must reflect your work alone. Sanctions for cheating include lowering your grade and/or failing the course. In egregious instances, the instructor may recommend the termination of your enrollment at CMU.

9 Late Work

You are expected to turn in all work on time. Because we understand that sometimes you need some more time to finish up your work and because exceptional circumstances may arise, we allow the submission of late assignments, subject to the following penalty rule:

If a work is delivered \( t \) seconds late, its score is adjusted by multiplying it by

\[
1 - \left( \frac{t}{24 \cdot 5 \cdot 60 \cdot 60} \right)^4
\]

which is the same as \( 1 - (t/5)^4 \) if \( t \) is measured in days. This penalty is applied automatically and there will be no exceptions to it. Note that the penalty for delivering an assignment 2 days late will be less than 3%, which in practice corresponds to having a 2-day grace period for all assignments. No resubmissions are allowed after the due date. Below is a graph depicting the maximum attainable grade as a function of the number of late days.

If interested on the motivations for this grading rule, you can check Daniel Reeves’ post on [http://messymatters.com/deadlines](http://messymatters.com/deadlines).

Note that we strongly recommend that you plan to deliver the assignments on time and use the extra time only if you really need it.

10 Teaching Assistants

This course may have several teaching assistants who will be participating in grading as well. Their names and contact information will be circulated at a later time.
11 Topic Outline & Important Dates

Week 1: 〈2015-03-17 Tue〉

Lecture 1: The Python Environment and Language I

• Basic concepts and environment
  – Installing Python
  – IPython
• Basic constructs and rules
• Documentation

Lecture 2: The Python Environment and Language II

• Collections
• Dictionaries
• List Comprehensions

Week 2: 〈2015-03-24 Tue〉

Homework 1: Available on 〈2015-03-24 Tue〉; Due on 〈2015-03-31 Tue〉.

Lecture 1: Essential Python Libraries I

• NumPy
  – ndarray
• pandas
  – Series
  – DataFrame

Lecture 2: Essential Python Libraries II

• pandas
  – DataFrame

Week 3: 〈2015-03-31 Tue〉

Homework 2: Available on 〈2015-03-31 Tue〉; Due on 〈2015-04-07 Tue〉.
Lecture 1: Data Loading and Storage

- Text Format
  - Delimited formats
  - JSON Data

- Databases
  - SQL
  - MongoDB

Lecture 2: Data Collection from Online Sources

- Web Scraping
- Web APIs

Week 4: <2015-04-07 Tue>

Homework 3: Available on <2015-04-07 Tue>; Due on <2015-04-14 Tue>.

Lecture 1: Data Manipulation I

- Combining and Merging Data Sets
- Reshaping and Pivoting

Lecture 2: Data Manipulation II

- Data Transformation
  - Aggregation
  - Group Operations

Week 5: <2015-04-14 Tue>

Homework 4: Available on <2015-04-14 Tue>; Due on <2015-04-21 Tue>.

Lecture 1: Descriptive Statistics I

- Summarizing the data
- Correlations

Lecture 2: Descriptive Statistics II

- Data Slicing

Week 6: <2015-04-21 Tue>

Homework 5: Available on <2015-04-21 Tue>; Due on <2015-04-28 Tue>. 
Lecture 1: Graphical Analysis I
• ggplot2
  – Histograms
  – Box Plots
  – Scatter Plots
  – Line Plots

Lecture 2: Graphical Analysis II
• ggplot2
  – Series
  – Faceting
  – Statistics

Week 7: <2015-04-28 Tue>
Homework 6: Available on <2015-04-28 Tue>; Due on <2015-05-05 Tue>.

Lecture 1: Linear Regression I
• What is a linear regression
• Coefficient Interpretation

Lecture 2: Linear Regression II
• Dummy Variables

Week 8: <2015-05-05 Tue>
Lecture 1: Classification Trees
Lecture 2: Final Exam (tentative date)

Last updated: January 16, 2015