**Data Collection and Analysis Using Python**

**PHD Program – 2018 - 19**

**(10Hrs)**

**COURSE OBJECTIVES**

In this course we will learn how to use the Python programming language for our research needs. This course is especially useful if your research requires you to collect and analyze data from the web. The course is intended for students who have little to intermediate programming knowledge. It might be a little too basic for students who have a good programming background.

By the end of this course you should be able to:

1. Understand the basic programming concepts i.e datatypes, conditions, functions and modules.
2. Read and write into an CSV/excel file
3. Collect data from the internet using the two most commonly used methods – a) using web APIs b) web scrapping
4. Use Python third party modules for higher computing problems

**COURSE SCHEDULE**

The course is divided into 3 sessions and runs for a total of 10 hrs.

1. Session 1: May 3rd from 2:00 p.m. to 5:00 p.m.
2. Session 2: May 6th from 2:00 p.m. to 6:00 p.m.
3. Session 3: May 9th from 9:30 [a.m. to](http://p.m.to/) 12:30p.m.

**PRE-ASSIGNMENT**

1. Download and install Anaconda distribution for python (version 3.7) - <https://www.anaconda.com/distribution/>
2. If you do not have any programming background – Complete (atleast) “Unit 1: Python Syntax” from - https://www.codecademy.com/learn/learn-python. If you prefer, you can use any other online sources for learning Python. In any case make sure you cover the basic commands and syntax of the language. Basically, before coming to class you should be able to do the following – a) initialize a variable as an integer or as a string. b) print a sentence and a variable on the output console (the screen).

**FINAL ASSESSMENT**

The final assessment will be based on the code that you submit for each of the three Python in-class sessions. If you are not able to complete some of the exercises in class, you can submit it at a later point in time. The course will be graded on a pass/fail basis.

**SESSION DETAILS**

**Session 1: Programming concepts (3 hours)**

1. Basics of python programing – Lecture (45mins): In this lecture we introduce open source, programing languages and the basic programing concepts
2. Integrated development environment (IDE) – Spyder: In this part we learn what an IDE is and how we can use this tool to build and debug code.
3. Exercise 1: We write our first code and strengthen our understanding of programming concepts.
   1. Data types and data structures
   2. Conditional statements and loops
   3. Functions and modules
   4. Basic I/O – Here we will write a program to open a CSV file and search its contents. We will use the “with open” command and “for” loop.
4. Exercise 2: We try and de-bug a code snippet that has several errors

**Session 2: Web data analysis (4 hours)**

1. Online data: (Lecture 30 mins) In this lecture we learn how data is accessed from the website (client-server model) and the formats in which data is available online. In specific, we try to understand what HTML, JSON formats are. Then we learn how we can get data from websites using either APIs or by scrapping.
2. Exercise 3: We will use the “requests” module to make a request on the Internet. And then use a web API to collect data in JSON format.
   1. Collect yelp data using their API and store it in excel sheet.
3. Exercise 4: We will use third party modules to scrape HTML data from any website.
   1. Using the “BeautifulSoup” module scrape data from a website and store in a CSV/Excel sheet
4. Exercise 5: We will try to scrape dynamic webpages and get data that is refreshed.

**Session 3: Other Python modules (3 hours)**

1. Feedback and learning: (Lecture 30 mins) In this lecture we will try and understand some of the popular machine learning methods.
2. Exercise 6: Text analytics using NLTK
   1. Use third-party modules to run basic supervised learning algorithms on text.
3. Exercise 7: Some basic classification using SKLEARN
4. If anyone has a pressing Python need, email me please. We can try and include part of that problem within this exercise or change this section to address that problem.

**NOT ALLOWED IN CLASS**

1. Throwing tomatoes at me.