INFO I590 Data Science Onramp Basics
(1-3 credit hours)

Data Science Onramp Basics contains mini courses with the goal to build and enhance your data science skills which are oftentimes demanded or desired in data science related jobs. Each mini course will be counted as one credit hour. Each time you enroll, you can select 1-3 credit hours which means that you can select 1 or 2 or 3 mini courses. You are allowed to maximally enroll twice for this whole course.

Most of the mini courses are written in text format and few in video format. We provide Teaching Assistant (TA) support and office hours. If you encounter any problems, please feel free to reach our TAs either at their office hours or schedule an appointment which fits better to your schedule. You can learn your selected mini courses in sequence or parallel. But we HIGHLY recommend parallel learning because: 1) you can participate the online discussion with other classmates; 2) our TAs will have weekly office hours and monthly live demos based on weekly and monthly contents of the mini courses; and 3) you have a good reason to get your assignment done on time rather than rush to finish them before the end of the semester.

Each mini course has its own grading policy. In general, grading is based on assignments/projects, online discussions, and quizzes. If you select more than one mini course, the average of each mini course grade will be counted towards your final grade.

Tableau

Tableau is a leading data analysis software used by analytics, banking and consulting organizations for data analysis. Tableau helps users to design/develop/deploy data science algorithms without writing huge chunks of code. The visualization of data joins and merges provides an easy way for a non-technical user to work on Data without worrying about coding in traditional scripting languages. In this course, we will learn about Tableau visualization from scratch to a professional level of understanding. We will also understand the techniques for building effective visualizations on various public data sets. The course consists of bi-weekly assignments which mainly focus towards a target problem and building visualization to discovery significant insights. There is also a final project for students to apply knowledge for a practical dataset and present their story-telling skills through interesting data visualizations.

This course will enable all the students gain all the important skills needed for building data visualizations and effective story telling. It will make the students proficient in using the tableau visualization tool and build impressive visualization story boards in their professional careers.
Course Structure:

- Introduction to data visualization and its usage
- Familiarizing with the Tableau visualization tool
- Importing data in tableau, working with sample data set, exploring features in tableau. Building simple visualizations in tableau
- Working on features like filters
- Effective use of Details feature, sorting options, view tool bar, worksheet options.
- Creating dashboard and worksheets
- Creating calculated fields, groping set, creating hierarchy
- Working with Time Series data set
- Building effective geo maps and other custom visualizations
- Implementing K-means clustering and classification, prediction in tableau
- Final Project

**Machine Learning with Python**

Machine learning is a technique which is used to teach computers, without being explicitly programmed. In this course, you will learn about basics of python and extending them to use different important packages like Matplotlib, Scikit-learn in python and about different kinds of classifications and classifiers used in machine learning.

We will begin our course with basic python programs because it is good to have some basic level of python experience before we go into advance concepts like machine learning. In industries, most of the computer programmers use two important approaches to write complex applications, recursive approach and iterative approach. You will learn about these important concepts from modules as well as from programming assignments. Scikit-learn is one of the best open source machine learning package in python with large active open source community. We will use this package to learn machine learning in applied fashion. At last, we will show you how you can build recommendation system using Scikit-learn package.

Course Structure:

- Introduction to Python
- Installing Python and setting up PyCharm IDE and Anaconda
- Python strings, constants, variables and scope
- Arithmetic and binary operations
- Control structures, functions and exception handling
- Using NumPy and Pandas library in Python
- Introduction to Matplotlib in Python
- Machine learning with Scikit-Learn and Scipy
- Concepts and implementation of linear regression using Numpy and Logistic Regression
- Introduction to Scipy
NLP in Python

Text mining starts generally with the process of information retrieval. We need to identify the source of data and then collect from this source. General sources are web, blogs, social media platforms, reviews and comments, etc. Once we collect the data, we need to clean the noise in it, such as the removal of duplicate data entries, unwanted information such as url's, image links, etc. There are number of steps involved in denoising the data and this depends on the kind of data that you have at hand. Once we clean the text data, we can apply natural language processing techniques such as parsing, pos tagging, etc.

The whole idea is to convert something not so structured into something meaningful and structured. Once we have such a structured output, we can perform various tasks such as:

- Sentiment analysis
- Topic detection
- Document summarization
- Entity relational modelling
- Pattern recognition
- Predictive analytics
- Text categorization

In this course we will cover some primary concepts in sentiment analysis. The above-mentioned tasks are extremely useful for gaining insights into textual data. We will explore the topics in detail.

Course Structure:

- Use regular expression to match string patterns
- Basic linux commend line functions
- Set up python and install packages using pip
- Basic functions in Python
- Basic python function working on strings
- Twitter APT to grab tweets
- Handle Json format and how to deal with it in Python
- Clean a tweet’s content by removing non-useful characters
- Use nltk to run semantic analysis on sentences
- Two projects
Machine Learning with Java

Machine learning is currently one of the hottest topics now. As the data science is deeply involved with machine learning algorithms and programming languages, it is important to master at least one programming language skill to play with machine learning algorithms or to solve real-life problems. In this course, we introduce how to use Java to build machine learning models to solve regression, classification, and clustering problems. We also introduce how to evaluate the machine learning models and interpret results. Although in fact there are many Java packages supporting machine learning algorithms, in this course, we only focus on the most popular one—Weka, which is a Java package containing many fancy algorithms and is widely used in recent years. One nice thing about Weka is that the package offers not only a Java library so that you can develop your own code to build a model but a well-developed GUI tool so that even for those people who are not familiar with Java, they can even build up a machine learning model very quickly by just clicking several buttons in the GUI.

Course Structure:

- What is Weka and install the GUI on your local computer
- What is Java install Java JRE and JDK so that you can run your code in the later sessions
- Input file format in Weka such as ARFF and XRFF
- Generate artificial data in Weka Java package
- Filter data in Weka
- Some basic classification methods using in Weka
- Tree based classification methods in Weka
- Advanced classification methods particularly in Weka
- Basic clustering methods using in Weka
- Learn how to visualize your results in Weka

Machine Learning with R

In this course, it is expected that you know the basic functionalities in R coding and we are going to cover the machine learning topics, how to implement them, what are the famous packages in R community, and how we can use those packages and how we can play with the different parameters in the packages which will affect the results. This will be a short course with 10 modules which will cover almost all widely used machine learning algorithms. Don’t worry, I won’t be adding a lot of theory to it rather I will be adding a lot of screenshots and code to give you a much better experience. My expectation is whatever task we are going to perform, please try to do hands-on side by side on your system. Don’t take this class as a theory lecture rather take it as a lab session.

Course Structure:

- Getting started
Web Scraping

As a Data Scientist, one is responsible for crunching humongous amounts of data to extract insights and streamline businesses based on the results. But the role of a Data Scientist doesn't start with understanding and analyzing data. Before we do any analysis, we must have data at hand. The first step to solving any data problem is to identify the problem, followed by collecting relevant data, and cleaning and representing the data in a functional form. Then we can use visualization and other analytical techniques to glean any useful insights.

It is fundamentally essential that data scientists can collect data from various sources. Data could be available in a structured form via well-defined REST APIs or unstructured (raw) data from websites, and any other type of data in-between. The Web Scraping course is all about extracting data of interest from any source.

The course will be divided into 5 parts. The first part deals with the basics of Python, which is completely optional for students with prior experience using Python. However, I recommend taking a quick glance at it unless you use Python on a day-to-day basis. The second part of the course deals with advanced Python coding necessary for web scraping. The third deals with extracting structured data using APIs. In the fourth part we throw light on basic tools and packages of Python for web and chrome development tool. Our fifth and final part deals with extracting raw data from web pages using Scrapy package.

Course Structure:

Part 1: Fundamentals of Python (Optional)

1. Using iPython notebooks
2. Control flow
3. Functions
4. Data Structures: Lists, tuples, dictionaries
5. Iterables and generators
Part 2: Essentials of Python

1. Object-oriented programming using Python
2. Error and Exception handling
3. File Input / Output
4. CSV files
5. JSON files
6. Strings and Regular Expressions

Part 3: Structured Data Extraction

1. REST APIs
2. Twitter API

Part 4: Fundamentals of Web Data and Developer Tools

1. HTML
2. XML
3. Chrome dev-tools
4. urllib package
5. BeautifulSoup package

Part 5: Building Spiders using Scrapy

1. Scrapy package

Machine Learning Principles

The goal of this course is to provide students with the knowledge and breadth of Machine Learning. This involves some of the crucial paradigms in the field such as the anatomy of Machine Learning problems, Gradient Descent, Regularization, Cross-Validation, Overfitting, Bias/Variance tradeoffs and more. Other topics covered are various practical algorithms used in Machine Learning such as Supervised Learning Problems using Linear Regression, Decision Trees, SVMs, Naive Bayes, and Logistic Regression, Unsupervised Learning Problems using K-Means and K-Nearest Neighbors, and Semi-supervised learning. Other miscellaneous topics are covered as well, such as Deep Learning, Reinforcement Learning, Ensemble Methods, and more. There are also various portions where the content is generated upon the students’ demands, where they can learn about what is trending and popular in the field of Machine Learning in the current day. All of these topics will be learned using readings,
quizzes, surveys, discussions, and of course coding assignments using Python and Jupyter Notebook!

Course Structure:

- Course Introduction and Overview
- Introduction to Machine Learning and the Development Environment for This Course
- Linear Regression
- Overfitting, Underfitting, the Bias/Variance Tradeoff, and Regularization
- Decision Trees
- Cross Validation
- Support Vector Machines
- Maximum Likelihood Estimation (MLE), Maximum A Posteriori Estimation (MAP) and Gradient Descent
- Naive Bayes
- Logistic Regression
- Supervised Learning, Unsupervised Learning, Semi-Supervised Learning, Reinforcement Learning, and Deep Learning
- K-Means Clustering
- Miscellaneous Topics 1
- Miscellaneous Topics 2
- Wrap-up week