**INSY 5378: Data Science: A Programming Approach**

**Spring 2016**

**Instructor:** Sridhar Nerur

**Office Number:** COBA (Business Building) Room 518

**Office Telephone Number:** 817-272-3530

**Email Address:** snerur@uta.edu

**Faculty Profile:** <http://wweb.uta.edu/insyopma/nerur/>

**Office Hours: M 1:00-2:00 p.m.; Wed. 4:30-5:25 p.m. or** by appointment

**Section Information:** INSY 5378 / Section 001

**Time and Place of Class Meetings:** M 2:00 – 4:50 p.m.; COBA 152

**Catalog Description**: The world is awash in data and companies are now trying to discern patterns and predict behaviors of both consumers and competitors to gain and sustain a competitive advantage. The unstructured nature of data as well as the myriad sources they come from make it particularly challenging for companies to systematically capture, cleanse, store, and analyze the data. Python is a simple yet powerful language that has a rich ecosystem to facilitate the analysis of such complex data. The aim of this course is to acquaint students with aspects of the Python language that are necessary to effectively function as a data scientist. Upon successful completion of the course, students will be familiar with data structures and programming constructs in the Python language, accessing data from files and databases, Market-Basket Analysis, Text Analytics, and Map-Reduce.

**Prerequisite**: Programming background required.

**Learning Objectives:** The aim of this course is to acquaint students with aspects of the Python language that are necessary to effectively function as a “data scientist”. Upon successful completion of the course, students will be familiar with:

a. Data structures and programming constructs in the Python language. Specifically, students will have a good grasp of lists, tuples, dictionaries, classes, selection (e.g., if ..else), and iteration (e.g., while and for loops).

b. Accessing data from files (e.g., text, csv, JSON, etc.) and databases. Students will also be exposed to the fundamentals of SQL.

c. Market-Basket Analysis using R

d. Text Analytics in Python & R, including topic modeling and sentiment analysis

e. Spark, which extends the notion of Map-Reduce, the abstraction from Google that inspired Hadoop. Specifically, we will use PySpark.

f. Machine learning algorithms using Scikit-learn.

g. Basics of Social Network Analysis using Networkx

**Required Textbooks and Other Course Materials:**

* *Data Science* by John Paul Mueller, Luca Massaron and EMC Education Services, Wiley Custom edition, ISBN: 978-1-11927-056-0
* *Mastering Python for Data Science* by Samir Madhavan, Packt Publishing Ltd., 2015, ISBN: 978-1-78439-015-0

**Other Materials/Resources:**

1. <https://www.kevinsheppard.com/images/0/09/Python_introduction.pdf>
2. <http://www.codecademy.com/en/tracks/python>
3. <https://developers.google.com/edu/python/>
4. Also check out [www.coursera.org](http://www.coursera.org) and [www.udacity.com](http://www.udacity.com) for introductory Python courses
5. <http://www.astro.washington.edu/users/vanderplas/Astr599/notebooks/17_SklearnIntro>

**Software Requirements:**

1. **Python:** Given below is a list of IDEs:

Anaconda (<http://continuum.io/downloads>) – my first choice

Enthought Canopy (<https://store.enthought.com/downloads/>)

PyCharm (<https://www.jetbrains.com/pycharm/download/>)

IDLE ([www.python.org](http://www.python.org))

1. **R (recommend R Studio as well)**
2. **Other Python modules (sklearn, gensim, networkx, rpy2, pymongo, and others)**

**Description of major assignments and examinations:**

The distribution of points will be as follows:

Homeworks 15%

Text Mining Project 10%

Machine Learning Project 10%

2 Quizzes 10%

Final Project 20%

Finals (Comprehensive) 25%

**Grading:** The following criteria will be used to assess your grade (**no rounding**):

A (>=90%), B (>=80%), C (>=70%), D (>=60%), F (<60%)

Final Project: An important component of the course is a group project. The class will be divided into groups\* of four for this purpose. Each group will analyze a dataset, write up a report, and make a presentation to the class. Groups must submit a brief description of the project to the professor by the 5th of March (2015) and get his approval before proceeding to work on the project. The project is meant to reinforce concepts taught in the lectures, slides, and/or additional readings. Points will be taken off if you do not attend **all** the presentations. Points for the project will be based on the following criteria:

1. Demo of the project (i.e., a walkthrough)
2. Depth of analysis
3. Communication – how well you present your story
4. Peer evaluations (therefore, it is possible for members of the same group to receive different scores)
5. A project report that has the following:
6. The business / research question that you are addressing
7. The implications of the project
8. Evaluation of models (why was a particular model chosen?)
9. Lesson learned
* ***Note that the same groups will work on the text mining and machine learning projects as well*.**

Further details will be provided in class and/or Blackboard.

**Attendance:** At The University of Texas at Arlington, taking attendance is not required. Rather, each faculty member is free to develop his or her own methods of evaluating students’ academic performance, which includes establishing course-specific policies on attendance. As the instructor of this section,I will consider attendance mandatory for all lectures. If you miss a class, you are responsible for the materials covered.

**Exams**: You are responsible for everything that is covered in the classroom, including additional materials that the instructor may discuss in class. There are **no make-up exams**. Under extenuating circumstances (e.g., medical emergency, family emergency, work-related travel, etc.), the average score of other exams will replace the missed exam score. You can only use this excuse for one exam. The final exam will be comprehensive covering all the contents, where regular exams will cover partial contents (as described in Course Schedule).

**Homeworks**: You will have 3 homework exercises. You will have about a week to complete the exercises. Homework must be turned in electronically (via BlackBoard) by the due date and time specified by the instructor. Late submissions will receive a score of 0 (no exceptions).

**Expectations for Out-of-Class Study**: Beyond the time required to attend each class meeting, students enrolled in this course should expect to spend at least an additional 9 hours per week of their own time in course-related activities, including reading required materials, completing assignments, and preparing for exams/quizzes.

**UNIVERSITY and COLLEGE POLICIES**

**Grade Grievances**: Any appeal of a grade in this course must follow the procedures and deadlines for grade-related grievances as published in the current University Catalog: <http://catalog.uta.edu/academicregulations/grades/#undergraduatetext>

**Drop Policy:** Students may drop or swap (adding and dropping a class concurrently) classes through self-service in MyMav from the beginning of the registration period through the late registration period. After the late registration period, students must see their academic advisor to drop a class or withdraw. Undeclared students must see an advisor in the University Advising Center. Drops can continue through a point two-thirds of the way through the term or session. It is the student's responsibility to officially withdraw if they do not plan to attend after registering. **Students will not be automatically dropped for non-attendance**. Repayment of certain types of financial aid administered through the University may be required as the result of dropping classes or withdrawing. For more information, contact the Office of Financial Aid and Scholarships (<http://wweb.uta.edu/aao/fao/>).

**Disability Accommodations:** UT Arlington is on record as being committed to both the spirit and letter of all federal equal opportunity legislation, including *The Americans with Disabilities Act (ADA), The Americans with Disabilities Amendments Act (ADAAA),* and *Section 504 of the Rehabilitation Act.* All instructors at UT Arlington are required by law to provide “reasonable accommodations” to students with disabilities, so as not to discriminate on the basis of disability. Students are responsible for providing the instructor with official notification in the form of a letter certified by the **Office for Students with Disabilities (OSD).** Students experiencing a range of conditions (Physical, Learning, Chronic Health, Mental Health, and Sensory) that may cause diminished academic performance or other barriers to learning may seek services and/or accommodations by contacting:

**The Office for Students with Disabilities, (OSD)** [www.uta.edu/disability](http://www.uta.edu/disability) or calling 817-272-3364.

**Counseling and Psychological Services, (CAPS)** [www.uta.edu/caps/](http://www.uta.edu/caps/) or calling 817-272-3671.

Only those students who have officially documented a need for an accommodation will have their request honored. Information regarding diagnostic criteria and policies for obtaining disability-based academic accommodations can be found at [www.uta.edu/disability](http://www.uta.edu/disability) or by calling the Office for Students with Disabilities at (817) 272-3364.

**Title IX:** *The University of Texas at Arlington does not discriminate on the basis of race, color, national origin, religion, age, gender, sexual orientation, disabilities, genetic information, and/or veteran status in its educational programs or activities it operates. For more information, visit*[*uta.edu/eos*](http://www.uta.edu/hr/eos/index.php)*. For information regarding Title IX, visit* [www.uta.edu/titleIX](http://www.uta.edu/titleIX).

**Academic Integrity:** Students enrolled all UT Arlington courses are expected to adhere to the UT Arlington Honor Code:

*I pledge, on my honor, to uphold UT Arlington’s tradition of academic integrity, a tradition that values hard work and honest effort in the pursuit of academic excellence.*

*I promise that I will submit only work that I personally create or contribute to group collaborations, and I will appropriately reference any work from other sources. I will follow the highest standards of integrity and uphold the spirit of the Honor Code.*

UT Arlington faculty members may employ the Honor Code as they see fit in their courses, including (but not limited to) having students acknowledge the honor code as part of an examination or requiring students to incorporate the honor code into any work submitted. Per UT System *Regents’ Rule* 50101, §2.2, suspected violations of university’s standards for academic integrity (including the Honor Code) will be referred to the Office of Student Conduct. Violators will be disciplined in accordance with University policy, which may result in the student’s suspension or expulsion from the University.

**Electronic Communication:** UT Arlington has adopted MavMail as its official means to communicate with students about important deadlines and events, as well as to transact university-related business regarding financial aid, tuition, grades, graduation, etc. All students are assigned a MavMail account and are responsible for checking the inbox regularly. There is no additional charge to students for using this account, which remains active even after graduation. Information about activating and using MavMail is available at <http://www.uta.edu/oit/cs/email/mavmail.php>.

**Student Feedback Survey:** At the end of each term, students enrolled in classes categorized as “lecture,” “seminar,” or “laboratory” shall be directed to complete an online Student Feedback Survey (SFS). Instructions on how to access the SFS for this course will be sent directly to each student through MavMail approximately 10 days before the end of the term. Each student’s feedback enters the SFS database anonymously and is aggregated with that of other students enrolled in the course. UT Arlington’s effort to solicit, gather, tabulate, and publish student feedback is required by state law; students are strongly urged to participate. For more information, visit <http://www.uta.edu/sfs>.

**Final Review Week:** A period of five class days prior to the first day of final examinations in the long sessions shall be designated as Final Review Week. The purpose of this week is to allow students sufficient time to prepare for final examinations. During this week, there shall be no scheduled activities such as required field trips or performances; and no instructor shall assign any themes, research problems or exercises of similar scope that have a completion date during or following this week *unless specified in the class syllabus*. During Final Review Week, an instructor shall not give any examinations constituting 10% or more of the final grade, except makeup tests and laboratory examinations. In addition, no instructor shall give any portion of the final examination during Final Review Week. During this week, classes are held as scheduled. In addition, instructors are not required to limit content to topics that have been previously covered; they may introduce new concepts as appropriate.

**Emergency Exit Procedures:** Should we experience an emergency event that requires us to vacate the building, students should exit the room and move toward the nearest exit. When exiting the building during an emergency, one should never take an elevator but should use the stairwells. Faculty members and instructional staff will assist students in selecting the safest route for evacuation and will make arrangements to assist individuals with disabilities.

**Student Support Services**:UT Arlington provides a variety of resources and programs designed to help students develop academic skills, deal with personal situations, and better understand concepts and information related to their courses. Resources include tutoring, major-based learning centers, developmental education, advising and mentoring, personal counseling, and federally funded programs. For individualized referrals, students may visit the reception desk at University College (Ransom Hall), call the Maverick Resource Hotline at 817-272-6107, send a message to resources@uta.edu, or view the information at <http://www.uta.edu/universitycollege/resources/index.php>

You may also find the following information useful:

Library Home Page <http://www.uta.edu/library>

Subject Guides [http://libguides.uta.edu](http://libguides.uta.edu" \t "_blank)

Subject Librarians [http://www.uta.edu/library/help/subject-librarians.php](http://www.uta.edu/library/help/subject-librarians.php%22%20%5Ct%20%22_blank)

Course Reserves [http://pulse.uta.edu/vwebv/enterCourseReserve.do](http://pulse.uta.edu/vwebv/enterCourseReserve.do%22%20%5Ct%20%22_blank)

Library Tutorials [http://www.uta.edu/library/help/tutorials.php](http://www.uta.edu/library/help/tutorials.php%22%20%5Ct%20%22_blank)

Connecting from Off- Campus [http://libguides.uta.edu/offcampus](http://libguides.uta.edu/offcampus%22%20%5Ct%20%22_blank)

Ask A Librarian [http://ask.uta.edu](http://ask.uta.edu/%22%20%5Ct%20%22_blank)

**Emergency Phone Numbers**: In case of an on-campus emergency, call the UT Arlington Police Department at **817-272-3003** (non-campus phone), **2-3003** (campus phone). You may also dial 911. Non-emergency number 817-272-3381

**Tentative Course Schedule**

|  |  |  |
| --- | --- | --- |
| **Week** | **Assignments and Topics** | **Readings** |
| **1** | Introduction to Big Data Analytics and Python | Syllabus, Chapter 1 |
| **2** | Python programming | Chapter 4 |
| **3** | Python programming continued | Chapter 4; *Chapter 1* |
| **4** | Quiz 1; Data Analytics Lifecycle; Advanced Python | Chapter 2; *Chapter 1* |
| **5** | Text Analytics | *Chapter 11; Slides* |
| **6** | Market Basket Analysis | Chapter 3; *Slides* |
| **7** | Working with data | Chapter 5 |
| **8** | Spring Break (No classes) March 14 - 18 |  |
| **9** | Machine Learning using Scikit-learn (sklearn) | Chapters 6-9; *Chapters 6,7,9, 10* |
| **10** | Machine Learning using Scikit-learn (sklearn) |  Same as above |
| **11** | Machine Learning using Scikit-learn (sklearn) | Same as above |
| **12** | Quiz 2; Recommendation Systems |  *Chapter 8* |
| **13** | MongoDB & PyMongo; Overview of Hadoop & PySpark | *Chapter 12; Slides* |
| **14** | Final Project presentations |  |
| **15** | Final Project presentations | Chapter 11 |
| **5/9** | **Final Exam (2:00 – 4:30 p.m.)** | Comprehensive |

**Chapters in italics are the ones from *Mastering Python for Data Science* by Samir Madhavan**

NOTE: *As the instructor for this course, I reserve the right to adjust this schedule in any way that serves the educational needs of the students enrolled in this course. Students are responsible to be aware of changes announced in class and/or via Blackboard. – Sridhar P Nerur*

**Kindly check the Academic Calendar for important dates**.