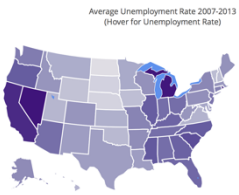
DATS 6103 – Introduction to Data Mining
CRN 54095 Section 11 ~~1957 E B14~~ (R)
Thu 6:10 PM – 8:40 PM
August 31, 2020 – December 12, 2020

INSTRUCTOR

Name: Nima Zahadat, Ph.D.
Term: Fall 2020
Campus Address: Samson Hall Suite 313
E-mail: nzahadat@gwu.edu
Office hours: By appointment

RESOURCES



Advanced Data Mining (Required)
TopHat Publishing
ISBN 978-1-77330-624-7

A Programmer's Guide to Data Mining
The Ancient Art of the Numerati (Optional)

COURSE DESCRIPTION

This course is a survey of concepts, principles, and techniques in data mining, including classification, association, and cluster analyses. Students learn to apply data mining methods to real-world problems with minimal rigorous mathematical understanding of the underpinnings of the methods. The course helps build a good foundation for taking advanced courses in the data science curriculum and for applying the basic techniques to practical problems. Data based examples and exercises using Python and other tools are integrated into class activities.

SOFTWARE

We will utilize Anaconda Python for this class along with Sublime text editor, SQLite browser, Plotly, Twitter, NLP, Excel, VirtualBox, other Office applications, and other tools as they become necessary.

COURSE PREREQUISITES

You are expected to have a basic knowledge of statistics (e.g., at the level of STAT 2118 Regression Methods that covers analysis of research data through simple and multiple regression and correlation). The course prerequisite is DATS 6101 or permission of the instructor.

OBJECTIVES

Be able to:

1. Develop code using Python
2. Clean and preprocess complex data using code
3. Work with external libraries
4. Visualize data using variety of library tools
5. Explain and use the mining process for descriptive and predictive analytics
6. Explore data using various mining and visualization techniques
7. Understand and apply the core data mining methods of classification, association and analysis

TOPICS

Data Mining and Analysis Overview: the nature of data, Python, and other tools; graphical display of data; classification and estimation; association and forecasting.

Basic Concepts of Pattern Recognition: basic statistical descriptions of data; data visualization; measuring similarity and dissimilarity; regression analysis; optimization; histogram.

Data Processing: data quality, cleaning, noise; data reduction, integration; and online analytical processing; types of data; outliers and robustness; corrupted, noisy, expensive, and heterogeneous data.

OUT OF CLASS TIME INVESTMENT

To get the most out of the class, students are required to dedicate at least 3-5 hours outside of the classroom, dedicated to doing research and working on their projects.

TENATIVE WEEKLY SCHEDULE (TOPICS WILL CHANGE THIS TERM)

Week	Assignment	Research
01 (08/31 – 09/04)	Syllabus Getting setup	learnpython.org codeacademy.com
02 (09/07 – 09/11)	Learn video capture Structure, loops, conditionals, collections Retirement calculator File operations Presidents code Splitter/Joiner	Start researching your first project data
03 (09/14 – 09/18)	Start reading first chapter of Ancient Art File operations review Splitter/Joiner interactive Python object model Port scanner Web server	Continue working on your first project
04 (09/21 – 09/25)	Continue reading Ancient Art Student database using text files Student database using SQLite	Continue working on your first project
05 (09/28 – 10/02)	Read chapter 2 of Ancient Art Data mining using Pandas Data frames FBI homicide dataset Baby names dataset	Continue working on your first project
06 (10/05 – 10/09)	Read chapter 3 of Ancient Art Terrorism dataset Python docx library	Continue working on your first project
07 (10/12 – 10/16)	Read chapter 4 in your book Watch Git and GitHub videos Learn github.io publishing Learn Zenodo publishing Unemployment dataset Homes dataset	Watch videos on Orange posted on Blackboard Project 1 due
08 (10/19 – 10/23)	Read chapter 4 of Ancient Art HTML review and Web Scraping	Research project 2
09 (10/26 – 10/30)	Read chapter 5 of Ancient Art Phishing dataset Climate change dataset	Work on project 2
10 (11/02 – 11/06)	Read chapter 6 of Ancient Art	Work on project 2

	Twitter API	
11 (11/09 – 11/13)	NLP	Project 2 due
12 (11/16 – 11/20)	Review of statistical concepts Data mining concepts	Research project 3
Off (11/23 – 11/27)	Thanksgiving; no class	
13 (11/30 – 12/04)	Facts of life presentation	Work on project 3
14 (12/07 – 12/11)	Makeup class if needed	Work on project 3
15 (12/14 – 12/18)	No class	Project 3 due

ASSESSMENTS

There will be three individual projects.
There will be weekly assignments.

GRADING

Grade scale is as follows:

97 – 100%	A+
93 – 96%	A
90 – 92%	A-
87 – 89%	B+
83 – 86%	B
80 – 82%	B-
77 – 79%	C+
73 – 76%	C
70 – 72%	C-
< 70%	F

Your projects are graded based on the following rubric:

Please note:

1. This is a Data Mining project. The focus is mining data, finding patterns, associations, meaning, forecasting, and/or making predictions, perhaps over time or in general
2. You must have a fairly complex dataset that by itself would be difficult or impossible to infer information from
3. You must program using Python and Python packages. While additional tools (R, MySQL, Excel, etc.) can be used to enhance your work, Python and its packages are required and must form the backbone of your work
4. Everything must work, including your programs

If even a single one of the above conditions is not met, the project gets a zero (0). If all the above conditions are met, then project 1 will be graded based on the following rubric:

- Followed directions and files properly packaged and named (5)
- Research (5)
- Design thinking (5)
- Data processing via code (5)
- Code organization and sophistication (5)
- Visualization (variety) (5)
- Visualization (meaningful) (5)
- Visualization (clear and easy to use) (5)
- Programs and tools (interface) (5)
- Programs and tools (easy to use) (5)
- Technical knowledge (5)
- Clarity of concepts and analyses (5)
- Focused? (5)
- Review and demo of code (5)
- Code commented professionally (5)
- Error free (5)
- Key findings (5)
- Conclusions (5)
- Presentation design, quality, professionalism (5)
- Presented within time (5)

Note that a project that doesn't work or is fairly incomplete will not receive partial credit and will be given a grade of 0. The rubrics apply only to working projects.

For projects 2 and 3, the rubric is below:

- Followed directions and files properly packaged and named (5)
- Topic selection (originality) (5)
- Topic selection (creativity) (5)
- Research (10)
- Dataset complex? (10)
- Design thinking (10)
- Data processing via code (10)
- Code organization and sophistication (10)
- Visualization (variety) (10)
- Visualization (meaningful) (10)
- Visualization (clear and easy to use) (10)
- Programs and tools (interface) (10)
- Programs and tools (easy to use) (10)
- Technical knowledge (10)
- Clarity of concepts and analyses (10)

- Focused? (10)
- Review and demo of code (5)
- Code commented professionally (5)
- Error free (5)
- Key findings (5)
- Conclusions (5)
- Presentation (organization) (5)
- Presentation (professionalism) (5)
- Presentation (quality) (5)
- Published to github.io and Zenodo (10)
- Presented within time (5)

Note that a project that doesn't work or is fairly incomplete will not receive partial credit and will be given a grade of 0. The rubrics apply only to working projects.

INDIVIDUAL PROJECTS

The individual projects will constitute the following:

1. The first project will be a presentation on the military spending of at least the top 10 or more countries in the world (100 points)
2. The second project will be a working data mining project on a topic of your choice that is approved by the instructor (200 points)
3. The second project will be a working data mining project on a topic of your choice that is approved by the instructor (200 points)

ASSIGNMENT SUBMISSION

-Assignments are due on time; see due dates on Blackboard
-Be sure to put all your files into a **folder** and name the folder like this

"DATS 6103 - Individual Project # - First Last"
as in
"DATS 6103 - Individual Project 1 - Bugs Bunny"
if your name is Bugs Bunny; otherwise use your own name

-Pay attention to the spacing and capitalization; do not add your own formatting
-Zip the folder and name it with the same formatting
-Unless your name is Bugs Bunny, use your own name
-**Upload your file to Blackboard**

READING ASSIGNMENTS

You are required to read roughly one chapter in your optional book every other week. You are required to read one chapter per week in your required book including watching the videos.

EMAIL ETIQUETTE

In the age of technology, when most forms of communication are electronic, it is important to adopt a proper etiquette to communicate with one another. It is asked that students use salutation when sending emails to their instructors and also make sure to SIGN their name and include their class/section at the end of the email. The instructor reserves the right NOT to reply to emails that are not properly addressed or do not have a signature. Students should also use their GWU email for any correspondence with the instructors. Students are required to check their emails daily and especially the morning before class.

ACADEMIC INTEGRITY

Students are responsible for understanding the George Washington University's Honor Code's provisions. In the spirit of the code, a student's word is a declaration of good faith acceptable as truth in all academic matters. Cheating and attempted cheating, plagiarism, lying, and stealing of academic work and related materials constitute Honor Code violations. These will not be tolerated. The code states: "Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information." For the remainder of the code, see: http://www.gwu.edu/~ntegrity/code.html

SUPPORT FOR STUDENTS OUTSIDE THE CLASSROOM

DISABILITY SUPPORT SERVICES (DSS)
Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Marvin Center, Suite 242, to establish eligibility and to coordinate reasonable accommodations. For additional information please refer to: http://gwired.gwu.edu/dss/

UNIVERSITY COUNSELING CENTER (UCC) ***202-994-5300***
The University Counseling Center (UCC) offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include:

- crisis and emergency mental health consultations
- confidential assessment, counseling services (individual and small group), and referrals
  http://gwired.gwu.edu/counsel/CounselingServices/AcademicSupportServices

SECURITY

In the case of an emergency, if at all possible, the class should shelter in place. If the building that the class is in is affected, follow the evacuation procedures for the building. After evacuation, seek shelter at a predetermined rendezvous location.