BIA-660-C Web Mining

School of Business
2022 Fall

Instructor:	Dr. Jingyi Sun
Course Website:	https://sit.instructure.com/courses/63228/
Meeting Times:	Wednesday 12:00-2:30PM
Classroom Location:	North Building 105
Contact Info:	jingyi.sun@stevens.edu
Office Hour:	Wednesday 4-5PM
Office Hour Location:	Babbio Center 645
Office Hour Zoom:	https://stevens.zoom.us/j/6363169775
Teaching Assistant:	Thirunaavukkarasu Murugesan, tmuruges@stevens.edu

Prerequisites:

- **Students must have programming experience**. Programming experience with Python, R, Java etc. is preferred.
- The primary programing language for this class is Python.
- This is **not an introductory course of Python programming**. Instead, this course focuses on web data retrieval, social network analysis and text mining along with extensive hands-on exercises using Python. If you haven't programmed with Python, you'll need to take extra effort to master python programming in order to keep up with the course progress.
- It is also highly recommended for students to have taken Multivariate Data Analytics (BIA 652), Social Network Analytics (BIA 658), Data Analytics & Machine Learning (MIS 637) in order to grasp the class materials.

COURSE DESCRIPTION

The Internet is essentially our largest source of data. In this course, through hands-on experience, students will learn the pipeline of analyzing web data, including:

- Reviewing Python programming concepts, useful modules, and analytics packages.

- Crawling data from the web including general websites and API.
- Extracting social networks from web data and network mining.
- Parsing, transforming, and annotating text data.
- Extracting features from the data to prepare for text mining.
- Conducting supervised and unsupervised learning to derive insights and help business decisions.

In addition to hands-on exercises, students need to work in team of 3-4 to collaborate on projects during the course. Through the projects, students will apply learned techniques and possibly research on other techniques to creatively solve some realistic business problems.

STUDENT LEARNING OUTCOMES

During the course, students will be able to develop the following useful skills:

- *Data collection and preprocessing skills*: students will learn how to identify and profile candidate sources of valuable data, as well as how to automatically collect and manage the information they need for their analytics tasks.
- *Social network analysis skills*: students will learn how to describe and analyze social networks extracted from web data.
- *Text mining skills*: students will be exposed to a wide range of well-accepted techniques for automated textual analysis.
- *Team skills*: students will be organized as teams to collaborate on course project. Every student in a team needs to make significant contribution to the team, as his or her contribution will be evaluated by the other members.

COURSE MATERIALS

- Lecture Notes will be posted on Canvas.
- Most lecture notes are Jupyter Notebooks. Slides will also be provided occasionally.
- Please review lecture notes and try all Python code in the lecture notes before class.
- If you run into problems with Python code, contact me or our TA to fix them before class.

TENTATIVE COURSE SCHEDULE

Module	Lecture#	Date	Topic	Assignments out
Python	1	1.18	Introduction & Python basics	Assignment 0
	2	1.25	Python basics	Assignment 1
Data Retrieval	3	2.1	Data scraping from websites	
	4	2.8	Data scraping from API	Assignment 2
	5	2.15	Network mining	
Networks	6	3.1	Preprocessing	Assignment 3
Text Mining	7	3.8	Preprocessing	
	8	3.22	Classification	Assignment 4
	9	3.29	Classification	
	10	4.5	Clustering	Assignment 5
	11	4.12	Topic modeling	
	12	4.19	Sentiment	Assignment 6
	13	4.26	Word vectors	
	14	5.3	Final project presentations	

COURSE REQUIREMENTS

Homework Assignments

- 6 assignments will be given. The best 5 assignments will be counted in your final grade. Assignment 0 is applying for a Twitter developer account.
- Assignments may include bonus questions for extra credits. Bonus points will be counted in the final grade.
- You'll need to write program code in most assignments. Your program code must be executable. **No credits will be given to program code that fails to execute** (No partial grade for failed program!). Submission guideline will be provided for each assignment.
- **Each assignment needs to be completed independently.** Although you may discuss with peers or use internet resources to get some general ideas, once start coding, you should work independently.
    - **Never ever copy others' work** (same logic with minor modification, e.g., changing variable names).
    - **Anti-Plagiarism software will be used to check all submissions**.
    - By "Graduate Student Code of Academic Integrity" (see below), in this class, **a student with plagiarism for the first time receives 20%-50% grade deduction of the assignment in question. The second violation leads to a failing grade for this course**.
    - **Never ever disclose your solutions to others. Always keep your solution confidential. If you allow someone to copy your assignment, no matter intentionally or unintentionally, once similar assignments are identified, you will receive 20%-50% grade deduction as well.**

- Each assignment will be distributed as scheduled on the corresponding lecture day. Assignments **are due in two weeks by 11:59 p.m. Eastern Time** before the lecture day. **10% penalty will be applied for each day of being late**. Please note, students living in distance time zones or overseas must comply with this due date deadline policy.
- You may ask for an extension in some situations (e.g. job interview) **within 5 days after an homework assignment is given**. After that period, no extension will be given (no extension in the last minute!)

Group project on text mining (also refer to the project guideline)

- *Midterm submission*: Collect, clean and organize online data from one or more websites of your choice. The deliverable includes a cleaned dataset, scripts for data scraping and cleaning, exploratory data analysis, and a short proposal on what insights can be derived from the dataset and what methods can be used to obtain the potential insights. Midterm report is due on **3/19**.
- *Final submission*: Choose an important research question that emerges in the context of the dataset collected for the midterm project. Develop, apply and record an analytics methodology to address your question. The entire project work will be presented in the last class. Final report is due on **5/7**.

*Important requirements and timeline for the group project***:**

- Find a team and write a proposal. Discuss your proposal with the instructor.
- Lay out a project plan, list detailed tasks and deadlines, and assign the tasks to each team member.
- Log project progress (completed tasks and signatures of task owners).
- Evaluate peer members.
- Submit your project, peer evaluation, project progress logs, and sign your submission.
- **Never ever copy (or imitate) others' projects (e.g. GitHub, Blogs). You need to show your own ideas in the class project. By "Graduate Student Code of Academic Integrity" (see below), such a violation will lead to a failing grade for all the team members (you are responsible for what your teammates submit!).**

GRADING PROCEDURES

Grades will be based on:

Homework	(50%)
Mid-term Project	(20%)

Final-term Project (20%)

Class attendance and participation (10%)

Grading scale:

Grade	Score
A	94-100
A-	90-93
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
F	<70

ACADEMIC INTEGRITY

Undergraduate Honor System

Enrollment into the undergraduate class of Stevens Institute of Technology signifies a student's commitment to the Honor System. Accordingly, the provisions of the Stevens Honor System apply to all undergraduate students in coursework and Honor Board proceedings. It is the responsibility of each student to become acquainted with and to uphold the ideals set forth in the Honor System Constitution. More information about the Honor System including the constitution, bylaws, investigative procedures, and the penalty matrix can be found online at http://web.stevens.edu/honor/

The following pledge shall be written in full and signed by every student on all submitted work (including, but not limited to, homework, projects, lab reports, code, quizzes and exams) that is assigned by the course instructor. No work shall be graded unless the pledge is written in full and signed.

"*I pledge my honor that I have abided by the Stevens Honor System*."

Reporting Honor System Violations
Students who believe a violation of the Honor System has been committed should report it within ten business days of the suspected violation. Students have the option to remain anonymous and can report violations online at www.stevens.edu/honor.

Graduate Student Code of Academic Integrity

All Stevens graduate students promise to be fully truthful and avoid dishonesty, fraud, misrepresentation, and deceit of any type in relation to their academic work. A student's submission of work for academic credit indicates that the work is the student's own. All outside assistance must be acknowledged. Any student who violates this code or who knowingly assists another student in violating this code shall be subject to discipline.

All graduate students are bound to the Graduate Student Code of Academic Integrity by enrollment in graduate coursework at Stevens. It is the responsibility of each graduate student

to understand and adhere to the Graduate Student Code of Academic Integrity. More information including types of violations, the process for handling perceived violations, and types of sanctions can be found at www.stevens.edu/provost/graduate-academics.

Special Provisions for Undergraduate Students in 500-level Courses

The general provisions of the Stevens Honor System do not apply fully to graduate courses, 500 level or otherwise. Any student who wishes to report an undergraduate for a violation in a 500-level course shall submit the report to the Honor Board following the protocol for undergraduate courses, and an investigation will be conducted following the same process for an appeal on false accusation described in Section 8.04 of the Bylaws of the Honor System. Any student who wishes to report a graduate student may submit the report to the Dean of Graduate Academics or to the Honor Board, who will refer the report to the Dean. The Honor Board Chairman will give the Dean of Graduate Academics weekly updates on the progress of any casework relating to 500-level courses. For more information about the scope, penalties, and procedures pertaining to undergraduate students in 500-level courses, see Section 9 of the Bylaws of the Honor System document, located on the Honor Board website.

LEARNING ACCOMODATIONS

Stevens Institute of Technology is dedicated to providing appropriate accommodations to students with documented disabilities. The Office of Disability Services (ODS) works with undergraduate and graduate students with learning disabilities, attention deficit-hyperactivity disorders, physical disabilities, sensory impairments, psychiatric disorders, and other such disabilities in order to help students achieve their academic and personal potential. They facilitate equal access to the educational programs and opportunities offered at Stevens and coordinate reasonable accommodations for eligible students. These services are designed to encourage independence and self-advocacy with support from the ODS staff. The ODS staff will facilitate the provision of accommodations on a case-by-case basis.

Disability Services Confidentiality Policy

Student Disability Files are kept separate from academic files and are stored in a secure location within the Office of Disability Services. The Family Educational Rights Privacy Act (FERPA, 20 U.S.C. 1232g; 34CFR, Part 99) regulates disclosure of disability documentation and records maintained by Stevens Disability Services. According to this act, prior written consent by the student is required before our Disability Services office may release disability documentation or records to anyone. An exception is made in unusual circumstances, such as the case of health and safety emergencies.

For more information about Disability Services and the process to receive accommodations, visit https://www.stevens.edu/office-disability-services. If you have any questions please contact: Phillip Gehman, the Director of Disability Services Coordinator at Stevens Institute of Technology at pgehman@stevens.edu or by phone (201) 216-3748.

INCLUSIVITY

Name and Pronoun Usage

As this course includes group work and in-class discussion, it is vitally important for us to create an educational environment of inclusion and mutual respect. This includes the ability for

all students to have their chosen gender pronoun(s) and chosen name affirmed. If the class roster does not align with your name and/or pronouns, please inform the instructor of the necessary changes.

Inclusion Statement

Stevens Institute of Technology believes that diversity and inclusiveness are essential to excellence in academic discourse and innovation. In this class, the perspective of people of all races, ethnicities, gender expressions and gender identities, religions, sexual orientations, disabilities, socioeconomic backgrounds, and nationalities will be respected and viewed as a resource and benefit throughout the semester. Suggestions to further diversify class materials and assignments are encouraged. If any course meetings conflict with your religious events, please do not hesitate to reach out to your instructor to make alternative arrangements.

You are expected to treat your instructor and all other participants in the course with courtesy and respect. Disrespectful conduct and harassing statements will not be tolerated and may result in disciplinary actions.

MENTAL HEALTH RESOURCES

Part of being successful in the classroom involves a focus on your whole self, including your mental health. While you are at Stevens, there are many resources to promote and support mental health. The Office of Counseling and Psychological Services (CAPS) offers free and confidential services to all enrolled students who are struggling to cope with personal issues (e.g., difficulty adjusting to college or trouble managing stress) or psychological difficulties (e.g., anxiety and depression). Appointments are can be made by phone (201-216-5177).

EMERGENCY INFORMATION

In the event of an urgent or emergent concern about the safety of yourself or someone else in the Stevens community, please immediately call the Stevens Campus Police at 201-216-5105 or on their emergency line at 201-216-3911. These phone lines are staffed 24/7, year round. Other 24/7 resources for students dealing with mental health crises include the National Suicide Prevention Lifeline (1-800-273-8255) and the Crisis Text Line (text “Home” to 741-741). If you are concerned about the wellbeing of another Stevens student, and the matter is *not* urgent or time sensitive, please email the CARE Team at care@stevens.edu. A member of the CARE Team will respond to your concern as soon as possible.