NYU SCHOOL OF PROFESSIONAL STUDIES

Special Topics: Text Data Mining

MASY1-GC 5000 | 101 | Spring 2024 | 01/25/2024 - 05/02/2024 | 3 Credits
Modality: In-person
Course Site URL: https://brightspace.nyu.edu/

General Course Information

Name/Title: Dr. Andres Fortino, Clinical Associate Professor, He/Him/His
NYU Email: agf249@nyu.edu
Class Meeting Schedule: 01/25/2024 - 05/02/2024 / Thursday | 02:00pm -- 04:35pm
Class Location: Bldg: MIDC Room 528
Office Hours: By appointment, NYU Zoom.

Description

This seminar will enhance the curriculum by identifying, analyzing, and applying special topics pertinent to the Management and Systems degree. The specific titles and content of each seminar will change to reflect emerging areas of interest, which can only be determined at the time of offering. The course may be used to satisfy the elective degree requirement. Applicability to specific concentrations will be noted in the course schedule and is at the department's discretion.

This course will cover the primary techniques for data mining and analyzing text data to discover interesting patterns, extract practical knowledge, and support decision-making, emphasizing statistical approaches that can be generally applied to arbitrary text data in any natural language with no or minimum human effort. Detailed analysis of text data requires an understanding of natural language text, which is a difficult task for computers. However, several statistical approaches have been shown to work well for the "shallow" but robust analysis of text data for pattern finding and knowledge discovery. You will learn the basic concepts, principles, and major algorithms in text mining and their potential applications. We shall learn to perform keyword analysis, semantic analysis, create visual representations of the text, perform qualitative data analysis, similarity scoring of texts, entity and topic extraction, and latent semantic analysis of text data. We shall use Python as our basic analysis tool via the open-source visual programming interface Orange 3.

Prerequisites

1210 – Quantitative Models for Decision Makers
1240 – Information Technology and Data Analytics

Learning Outcomes

- At the conclusion of this course, students will be able to:
- Construct applications using unstructured data like news articles and tweets.
- Apply machine learning classifiers to categorize documents by content and author.
- Practice using document similarity and topic models to work with large data sets.
- Visualize and interpret text analytics, including statistical significance testing.
- Perform sentiment analysis of product reviews and social media postings.

- Use Python text analysis via visual programming

Communication Methods
This course describes a range of business opportunities and solutions centered around the use of text. It also identifies sources of competitive intelligence in text and provides solutions for parsing and storing incoming knowledge. It is based on the merging technology of natural language processing. It uses real-world case studies, and the course provides examples of the most useful statistical and machine learning techniques for handling text, semantic, and social data. We then describe how and what you can infer from the data and discuss practical approaches for visualizing and communicating the results to decision-makers.
This course will use the NYU Brightspace LMS for the delivery of course materials and for course communications.

Important course information, announcements, updates, course presentations and other materials will be posted on the NYU Brightspace LMS. Course announcements will also be simultaneously forwarded by the NYU Brightspace LMS to students' NYU email addresses. Students are expected and required to be aware of any such announcements or communications and are advised to check the announcements as well as their NYU email address regularly during the 14-week duration of the course as well as afterward while conducting their Applied Project.

Credit students must use their NYU email to communicate. NYU Brightspace LMS course-mail supports student privacy and FERPA guidelines.

The instructor's email address is agf249@nyu.edu and it is checked regularly and frequently; students will usually receive a reply within 12 hours during the workweek. The instructor does not have access to an NYU telephone number.

The instructor will conduct office hours using Zoom, by telephone, or in-person at the NYU campus—by appointment. If you would like to schedule a meeting, please send an email to the instructor at least two days prior to the date you would like to meet. You should also suggest an alternative date in case the first date is not available. Discussions through online platforms will require that you have speakers and a microphone. A video camera is highly recommended.

Structure | Method | Modality
This course will be conducted in person once a week on Tuesdays for 14 weeks. The class will encompass lectures, assignments, class team workshop exercises, examples and demos, and a final exam, and a team project. All class content and assignments will be made available online via Brightspace. The student should check the website daily for any updates or announcements.

Expectations

Learning Environment

You play an important role in creating and sustaining an intellectually rigorous and inclusive classroom culture. Respectful engagement, diverse thinking, and our lived experiences are central to this course and enrich our learning community. As graduate students, you are expected to conduct yourselves in a professional manner and engage and collaborate with your classmates. SPS classrooms are diverse and include students who range in age, culture, learning styles, and levels of professional experience. To maintain an inclusive environment that ensures all students can equally participate with and learn from each other, as well as receive feedback and instruction from faculty during group discussions in the classroom, all course-based discussions and group projects should occur in a language that is shared among all participants.

Participation

To receive full credit for class participation, you should attend all classes since much of the learning occurs during class lectures, presentations, and class discussions. You must contribute and engage in class dialogue during every class session for the course. Please contact the instructor if you anticipate missing any part of the class. Please arrive on time so as not to disturb the flow of the lecture. Excessive lateness may result in a lower overall grade. Please contact the instructor if you anticipate missing any part of the class.

Participation grades will be based on:

A. Involvement in class discussions, dialogues, and activities during each session
B. Participation demonstrates the integration of reading, classwork, relevance, and application.
C. Willingness to learn by accepting feedback, trying new skills and approaches, etc.
D. Quality/quantity of providing effective and balanced feedback.

Students who join the course during add/drop are responsible for ensuring that they identify what assignments and preparatory work they have missed and complete and submit those per the syllabus.

Assignments and Deadlines

Students are expected to participate in each class session by understanding the subject, sharing ideas, or discussing/commenting on other students' comments. In addition, students must complete and submit all assigned homework on time. Late submission of homework will either not be accepted or will result in a 10% loss of credit. Students are also expected to develop with and present a team project with other students and take and pass a final exam.

See full detail of expectations under “Assessment Strategy” below. Further information about specific assignments can also be found in the “Course Outline” section.

Course Technology Use

We will utilize multiple technologies to achieve the course goals. I expect you to use technology in ways that enhance the learning environment for all students. All class sessions require the use of Zoom. All class sessions require the use of technology (e.g., laptop, computer lab) for learning purposes. You must bring a laptop to class.

The Use of AI

You are expected to use AI (ChatGPT and code generation tools) appropriately in this class. In fact, some assignments will require it. Learning to use AI is an emerging skill, and I provide tutorials on how to use them. I am happy to meet and help with these tools during office hours or after class. Be aware of the limits of ChatGPT:

- If you provide minimum effort prompts, you will get low-quality results. You will need to refine your prompts in order to get good outcomes. This will take work.
- Don't trust anything it says. If it gives you a number or fact, assume it is wrong unless you either know the answer or can check in with another source. You will be responsible for any errors or omissions provided by the tool. It works best for topics you understand.
- Be thoughtful about when this tool is useful. Only use it if it is appropriate for the case or circumstance.
- AI is a tool, but one that you need to acknowledge using. *Please include a paragraph at the end of any assignment that uses AI explaining what you used the AI for and what prompts you used to get the results, as given below. Please do so in compliance with academic honesty policies.*
- Each assignment should have a properly annotated version of the following statement at the end of the assignment, which you, in good faith, are pledging is true:

> *"I/We have used AI and AI-assisted technologies as per the course policy in preparing this assignment. Specifically, I/we employed* ***[Insert Name of AI Tool/Service]*** *for* ***[Insert Purpose]****. This tool/service assisted in* ***[elaborating specific contributions - e.g., drafting initial ideas, improving language]****. Following the use of the AI tool, I/we have thoroughly reviewed, edited, and critically analyzed the content to ensure it aligns with the learning objectives and academic integrity standards of the course. The final submission represents my/our original thought and analysis, with AI tools serving only as an aid in the process. I/We take full responsibility for the content of this assignment."*

Feedback and Viewing Grades

I will provide timely meaningful feedback on all your work via our course site in NYU Brightspace. You can access your grades on the course site Gradebook.

Attendance

I expect you to attend all class sessions. Attendance will be taken into consideration when determining your final grade. Refer to the SPS Policies and Procedures page for additional information about attendance.

Excused absences are granted in cases of documented serious illness, family emergency, religious observance, or civic obligation. In the case of religious observance or civic obligation, this should be reported in advance. Unexcused absences from sessions may have a negative impact on a student's final grade. Students are responsible for assignments given during any absence.

Each unexcused absence or being late may result in a student's grade being lowered by a fraction of a grade. A student who has three unexcused absences may earn a Fail grade.

Refer to the SPS Policies and Procedures page for additional information about attendance.

Textbooks And Course Materials

Required:

Fortino, Andres, Text Data Mining- A Case Study Approach, Mercury Publishers, 2021

Course Chatbot:

Use this ChatGPT chatbot to ask questions form the course materials, including the syllabus, course notes, session transcripts, grading rubrics and assignments

https://chat.openai.com/g/g-GpgVwhNQj-text-data-mining-course-companion

SOFTWARE

Required

Voyant - https://voyant-tools.org/
Orange3 - https://orangedatamining.com/
ChatGPT: https://openai.com/
Additional open-source programs will be required and installed as instructed in class.

Grading | Assessment

Final Assignment – (20%). There will be one final assignment in the form of a team analytics deliverable to ensure that the student has mastered the material presented. Instructions for the assignment are posted on the class website. The final assignment is due on the last day of the semester and will not be accepted late.

Final Team Case Study - (20%)
Part A – Proposal (5%)
Part B – Final (15%)

Labs – 10 required labs (50% total, 5% each). Two additional optional labs are provided for practice and additional topic coverage. There is a lab due every week. The top 10 out of 12 lab grades will be retained to contribute to the final grade; the lowest two lab grades will be dropped. Student answers to the labs will be entered in the appropriate Assignment in the Brightspace class website. They are due one week after the class. There is a 10% penalty for a late assignment posting for up to a week, and no credit will be given for a lab assignment delivered after that.

Lab 1 -	Framing Questions	(5%)
Lab 2 -	Tools, Techniques and Data Preparation	(5%)

Lab 3 -	Word Frequency Analysis	(5%)
Lab 4 -	Keywords Analysis	(5%)
Lab 5 -	Sentiment Analysis	(5%)
Lab 6 -	Visualizing Text Data	(5%)
Lab 7 -	Coding for Qualitative Data Analysis	(5%)
Lab 8 -	Entity Extraction	(5%)
Lab 9 -	Topic Recognition	(5%)
Lab 10 -	Text Similarity Scoring	(5%)
Lab 11 -	Fuzzy Logic (optional)	(5%)
Lab 12 -	Practice Final Exam (optional)	(5%)

Team Class Workshops – (10% total, 1% each). 10 required team workshop deliverables An additional optional workshop on visual programming is provided for practice and additional topic coverage. There is a team workshop due every week. The top 10 out of 11 lab grades will be retained to contribute to the final grade; the lowest team workshop grade will be dropped. Student answers to the team workshops will be entered in the appropriate Assignment on the Brightspace class website. They are due one day after the class. The assignments are done by the team at the end of each class so there is no need for extra time to complete assignments No credit will be given for a lab assignment delivered after that.

Lab 1 -	Framing Questions	(1%)
Lab 2 -	Tools, Techniques and Data Preparation	(1%)
Lab 3 -	Word Frequency Analysis	(1%)
Lab 4 -	Keywords Analysis	(1%)
Lab 5 -	Sentiment Analysis	(1%)
Lab 6 -	Visualizing Text Data	(1%)
Lab 7 -	Coding for Qualitative Data Analysis	(1%)
Lab 8 -	Entity Extraction	(1%)
Lab 9 -	Topic Recognition	(1%)
Lab 10 -	Text Similarity Scoring	(1%)
Lab 11 -	Fuzzy Logic	(1%)

Final – (10%). There will be a short 1-hour in-class final. A practice final exam will be made available for you to practice taking the final exam: Lab 12.

Review Quizzes - (10%). There are 10 out of 12 required REs (Reflection Exercises), 1% each for taking them; credit is not based on the score. This is not an exercise to measure what you know but to help you transfer knowledge from short-term memory to long-term memory. Students who used these exercises got as much as a 30% increase in their final exam grades in the past. The quizzes are open online for a whole week, and they are timed to maximize knowledge transfer. Students are advised to take each Quiz when it is available.

See the "Grades" section of Academic Policies for the complete grading policy, including the letter grade conversion, and the criteria for a grade of incomplete, taking a course on a pass/fail basis, and withdrawing from a course.

Course Outline
Start/End Dates: 01/25/2024 - 05/02/2024 / Thursday
Time: 02:00pm -- 04:35pm
No Class Date(s): Thursday – 03/21/2024
Special Notes: Spring Break 03/18/24 - 03/24/24

Session 1 - 01/25/24
Topic Description: Introduction
What is text data mining?
Assignments: Read Fortino CH 1

Session 2 – 02/01/24
Topic description – Framing Analytical Questions
How to frame good analytical questions?
Assignments: Lab 1: Framing Questions

Session 3 – 02/08/24
Topic description – Preparing the data files
What are the tools, data formats and data preparation processes?
Assignments: Lab 2: Preparing data sets

Session 4 – 02/15/24
Topic description – Word Frequency Analysis
How do we discover the most frequent words in text data?
Assignments: Lab 3: Word Frequency Analysis

Session 5 – 02/22/24
Topic description – Keyword Analysis
How do we discover the keywords that characterize a document?
Assignments: Lab 4: Keyword Analysis

Session 6 – 02/29/24
Topic description – Sentiment Analysis
How do we discover what people are telling us in their texts?
Assignments: Lab 5: Sentiment Analysis

Session 7 – 03/07/24
Topic description – Coding Qualitative Data
How do we analyze qualitative data using coding techniques?
Assignments: Lab 6: Coding Qualitative data

Session 8 – 03/21/24
Topic description – Visualizing Text Data
How do we convert our text analysis to a visual?
Assignments: Lab 7: Visualizing Text

Session 9 – 03/28/23
Topic description – Text Similarity Scoring
How do we compare the similarities of the two text documents?
Assignments: Lab 8: Text Similarity Scoring

Session 10 – 04/04/24
Topic description – Named Entity Recognition
How do we extract the entities from textual data?
Assignments: Lab 9: Entity Recognition

Session 11 – 04/11/24
Topic description – Topic Recognition
How do we extract the topics covered by a document?
Assignments: Lab 10: Topic Recognition

Session 12 – 04/18/24
Topic description – Fuzzy Logic
Introduction to Computing with Words (CCW) fuzzy logic techniques for text mining
Assignments: Lab 11: Fuzzy Logic for Text Data Mining

Session 13 – 04/25/24
Topic description – Visual Programming
Introduction to graphical phyton tool for text mining - Orange3
Assignments: Lab 12: Final Exam Practice (optional)

Session 14 – 05/02/24
Topic description – Exam- Online
FINAL EXAM – and FINAL ASSIGNMENT DUE
Assignments: Final Assignment

NOTES:

The syllabus may be modified to better meet the needs of students and to achieve the learning outcomes.

The School of Professional Studies (SPS) and its faculty celebrate and are committed to inclusion, diversity, belonging, equity, and accessibility (IDBEA), and seek to embody the IDBEA values. The School of Professional Studies (SPS), its faculty, staff, and students are committed to creating a mutually respectful and safe environment (*from the SPS IDBEA Committee*).

New York University School of Professional Studies Policies

1. Policies - You are responsible for reading, understanding, and complying with University Policies and Guidelines, NYU SPS Policies and Procedures, and Student Affairs and Reporting.

2. Learning/Academic Accommodations - New York University is committed to providing equal educational opportunity and participation for students who disclose their dis/ability to the Moses Center for Student Accessibility. If you are interested in applying for academic accommodations, contact the Moses Center as early as possible in the semester. If you already receive accommodations through the Moses Center, request your accommodation letters through the Moses Center Portal as soon as possible (mosescsa@nyu.edu | 212-998-4980).

3. Health and Wellness - To access the University's extensive health and mental health resources, contact the NYU Wellness Exchange. You can call its private hotline (212-443-9999), available 24 hours a day, seven days a week, to reach out to a professional who can help to address day-to-day challenges as well as other health-related concerns.

4. Student Support Resources - There are a range of resources at SPS and NYU to support your learning and professional growth. For a complete list of resources and services available to SPS students, visit the NYU SPS Office of Student Affairs site.

5. Religious Observance - As a nonsectarian, inclusive institution, NYU policy permits members of any religious group to absent themselves from classes without penalty when required for compliance with their religious obligations. Refer to the University Calendar Policy on Religious Holidays for the complete policy.

6. Academic Integrity and Plagiarism - You are expected to be honest and ethical in all academic work. Moreover, you are expected to demonstrate how what you have learned incorporates an understanding of the research and expertise of scholars and other appropriate experts; and thus recognizing others' published work or teachings—whether that of authors, lecturers, or one's peers—is a required practice in all academic projects.

Plagiarism involves borrowing or using information from other sources without proper and full credit. You are subject to disciplinary actions for the following offenses which include but are not limited to cheating, plagiarism, forgery or unauthorized use of documents, and false form of identification

Turnitin, an originality detection service in NYU Brightspace, may be used in this course to check your work for plagiarism.

Read more about academic integrity policies at the NYU School of Professional Studies on the Academic Policies for NYU SPS Students page.

7. Use of Third-Party Tools - During this class, you may be required to use non-NYU apps/platforms/software as a part of course studies, and thus, will be required to agree to the “Terms of Use” (TOU) associated with such apps/platforms/software.

These services may require you to create an account but you can use a pseudonym (which may not identify you to the public community, but which may still identify you by IP address to the company and companies with whom it shares data).

You should carefully read those terms of use regarding the impact on your privacy rights and intellectual property rights. If you have any questions regarding those terms of use or the impact on the class, you are encouraged to ask the instructor prior to the add/drop deadline.