

**Special Topics – Text Data Mining**

# GENERAL COURSE INFORMATION

Faculty/Title: Andres Fortino, Clinical Associate Professor

Email: agf249@nyu.edu

Course title/number: Special Topics -Text Data Mining – MASY-GC5000-101

Credits as applicable: 3

Semester/Year: Spring 2022

Session: 1

Number of Classes: 14

Dates/Day(s): 1/25/22- 5/3/22

Time: 2:00pm -- 4:35pm EST

No Class Dates: Tuesday March 15, 2022 – Spring recess

Mode of Delivery: In Person

Special Notes: Office Hours Email me to request an appointment – agf249@nyu.edu

Appointments to be requested and scheduled before class starts.

# COURSE DESCRIPTION

*Special Topics Course 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.

*Text Data Mining Course Description*

This course will cover the primary techniques for 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.

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.

# COURSE PREREQUISITES:

GC1240 – Information Technology and Data Analytics

GC1210 – Quantitative Models for Decision Makers

**COURSE STRUCTURE/METHOD**

This course will be delivered Online - Synchronous, once a week on Wednesdays. Zoom is the remote instruction platform used at NYU. The class will encompass lectures, assignments, examples and demos, midterm and final exams, 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.

**COURSE 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.

**COMMUNICATION POLICY**

Credit students must use their NYU email to communicate. Non-degree students do not have NYU email addresses. Brightspace course-mail supports student privacy and FERPA guidelines. All email inquiries will be responded to within 24 hours from Monday through Friday at 5 pm. Email sent on Saturday or Sunday will not be responded to until Monday. I will respond to you using NYU's email.

Students have the opportunity to add their pronouns and the pronunciation of their names into Albert. Students can have this information displayed to faculty in Albert, Brightspace, and other NYU systems. Students can also opt-out of having their pronouns viewed by their instructors.

<https://www.nyu.edu/students/student-information-and-resources/registration-records-and-graduation/forms-policies-procedures/change-of-student-information/pronouns-and-name-pronunciation.html>

**COURSE EXPECTATIONS**

## Attendance

Students are expected to attend all classes. 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 responsibility, 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 may result in a student’s grade being lowered by a letter grade. A student who has three unexcused absences may earn a Fail grade.

University Calendar Policy on Religious Holidays: <https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html>

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.

## Classroom Expectations

As graduate students, you are expected to conduct yourselves professionally and engage and collaborate with your classmates in the Zoom meeting room. Here are our guiding principles:

* Dress as if you are in the Classroom.
* Keep your microphone muted unless asking a question or engaging in discussion.
* Check your video and audio when entering your class session.
* Think background, minimize distractions around you.
* Look into the camera instead of looking at the screen.
* Type quietly, mute if necessary.
* Don’t eat during a Zoom class session and refrain from engaging in any activity such as smoking, consuming alcohol, etc., that you would not engage in if the class was in person.

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 and 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.

## Assignments

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 lower grade. Students are also expected to develop with and present a team project with other students and take and pass a midterm exam and 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.

## Technology Policy

Use of either PC or Mac desktop or laptop will be needed to access all course content/materials, notes, assignments, and examples and demos. All class sessions require the use of Zoom.

**Online Technical Support:**

**IT Service Desk**

(212)-998-3333

24 hours a day, 7 days a week

Email: [AskIT@nyu.edu](mailto:AskIT@nyu.edu)

**Zoom Support**

* [NYU Zoom Guide for Students](https://nyu.service-now.com/servicelink/search_results.do?sysparm_document_key=kb_knowledge,bd4d4732dbf34f008fd2a2364b961964&sysparm_search=zoom)
* Make sure you are using [NYU Zoom](https://www.nyu.edu/life/information-technology/communication-and-conferencing/meetings-chat-conferencing/nyu-zoom.html) to log-in for class
* Check the [NYU Zoom site](https://t.e2ma.net/click/pvdgzd/148ffmn/lccv1i) often for updates. (To update Zoom, you can also open from your desktop and click menu, then “Check for Updates.”)

**Brightspace Support**

* Log-in to the [Brightspace](https://brightspace.nyu.edu/d2l/home)platform or visit the [Student Training](https://www.nyu.edu/life/information-technology/instructional-technology-support/instructional-technology-tools-and-services/nyu-lms-brightspace/student-training-lms-brightspace.html) website.

## Class 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:

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

# REQUIRED AND RECOMMENDED MATERIAL

Required

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

Recommended

* Miner, Gary, John Elder IV, Andrew Fast, Thomas Hill, Robert Nisbet, and Dursun Delen. *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.
* The instructor will also provide session-by-session content, which will be posted online.

SOFTWARE:

* Required – R, RStudio, JASP, Jamovi, RCommander, <https://cran.r-project.org/>
* Voyant - <https://voyant-tools.org/>
* Orange3 - <https://orangedatamining.com/>
* Additional open-source programs will be required and installed as instructed in class

# ASSESSMENT STRATEGY

**Assessments:**

**Final Assignment** – 1 Team Assignment 20%

**Labs** – 10 required labs/12 50% total, 5% each).

**Team Class Workshops** 10 workshops/11 10% total, 1% each).

**Reflection Exercises** 10 Exercises/13 10%

**Final Exam** 10%

**Final Assignment** – 1 Assignment (20%)

There will be one final assignment in the form of a team analytics deliverable to assure that the student has mastered the material presented. Instructions for the assignment are posted on the class website. Late assignments will receive a 20% penalty in the grade. 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** – Text Analytics with Visual Programming (optional) (5%)
* **Lab 12** - Practice Final Exam (optional) (5%)

**Team Class Workshops** – (10% total, 1% each). 10 required team workshop deliverables One 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 team 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 in 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 assignment 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** – Text Analytics with Visual Programming (optional) (1%)

**Reflection Exercises** (10%): There are 10 out of 12 required REs (Reflection Exercises), 1% points 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.

**Final Exam** (10%): There will be a 1 hour, 25 multiple choice question exam administered on week 14. A practice final exam will be made available for you to practice taking the final exam: Lab 12.

# NYUSPS Policies:

“NYUSPS policies regarding the Family Educational Rights and Privacy Act (FERPA), Academic Integrity and Plagiarism, Students with Disabilities Statement, and Standards of Classroom Behavior among others can be found on the NYU Brightspace LMS Academic Policies tab for all course sites as well as on the University and NYUSPS websites. Every student is responsible for reading, understanding, and complying with all of these policies.”

The full list of policies can be found at the web links below:

University: http://www.nyu.edu/about/policies-guidelines-compliance.html

NYUSPS: https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html#Graduate1

# CENTER FOR STUDENT ACCESSIBILITY

If you are a student who is requesting accommodations, please contact New York University’s Moses Center for Students Accessibility (CSA) at 212-998-4980 or mosescsa@nyu.edu. You must be registered with CSA to receive accommodations. Information about the Moses Center can be found at www.nyu.edu/csa. The Moses Center is located at 726 Broadway on the 3rd floor.

# 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.

# ACADEMIC INTEGRITY AND PLAGIARISM POLICY

All students are expected to be honest and ethical in all academic work. This trust is shared among all university community members and is a core principle of American higher education. Any breaches of this trust will be taken seriously. A hallmark of the educated student and good scholarship is the ability to acknowledge information derived from others. Students are expected to be scrupulous in crediting those sources that have contributed to developing their ideas.

Plagiarism involves borrowing or using information from other sources without proper and full credit. Students are expected to demonstrate how what they 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.

Students are subject to disciplinary actions for the following offenses, which include but are not limited to: Cheating; Plagiarism; Forgery or unauthorized use of documents; False form of identification

Use the link below to read more about Academic Integrity Policies at the NYU School of Professional Studies. Academic Policies for NYU SPS Students

<https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html#Graduate1>

# TURNITIN

TurnItIn is a plagiarism detection software used to verify academic originality. It is available only to degree courses and students. Some assignments in this course may be checked by the instructor for plagiarism using Turnitin. Most assignments will be checked using the plagiarism tool in the grammar checking software Grammarly.

**NYU School of Professional Studies Graduate Grading Scale**

Grading for graduate programs is by letter grade: A, A-, B+, B, B-, C+, C, C-, and F. For NYUSPS’s complete graduate grading policies, including criteria for a grade of incomplete, taking a course on a pass/fail basis, and withdrawing from a course, see: <https://www.sps.nyu.edu/homepage/student-experience/policies-and-procedures.html#Graduate1>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Letter** | **%** | **GPA** | **Descriptions** | **Definitions** |
| A | 95-100 | 4.0 | Exceptional | Demonstrates exceptional mastery of all learning outcomes of the course and thorough and complete understanding of all concepts. |
| A- | 90-94 | 3.7 | Excellent | Demonstrates highly competent mastery of all learning outcomes of the course and strong understanding of all concepts. |
| B+ | 87-89 | 3.3 | Very Good; exceeds course standards | Demonstrates mastery of all learning outcomes of the course and understanding of core concepts. |
| B | 83-86 | 3.0 | Good; meets course standards | Demonstrates mastery of some learning outcomes; understanding of some core concepts could be improved. |
| B- | 80-82 | 2.7 | Somewhat Satisfactory; meets some course standards and requires improvement | Demonstrates basic understanding of some learning outcomes; improved understanding of all core concepts is needed. |
| C+ | 77-79 | 2.3 | Less than Satisfactory; requires significant improvement | Demonstrates partial understanding of all learning outcomes and core concepts; requires significant improvement. |
| C | 73-76 | 2.0 | Unsatisfactory; requires substantial improvement | Demonstrates partial understanding of some learning outcomes and core concepts; requires substantial improvement. |
| C- | 70-72 | 1.7 | Unsatisfactory; requires extensive improvement | Demonstrates poor understanding of all learning outcomes and core concepts; requires extensive improvement. |
| F | Below 70 | 0.0 | Fail | Demonstrates minimal to no understanding of all key learning outcomes and core concepts; work is unworthy of course credit towards the degree. |

# COURSE OUTLINE

Table

Description automatically generated