THE COLLEGE OF NEW JERSEY
SCHOOL OF BUSINESS

COURSE NAME

ISTG 640, Text Mining

INSTRUCTOR

Dr. Zoe Huang, huangz@tcnj.edu
Zoom ID: 223 288 6231
Password: DrHuang

Use Canvas Inbox for All Email or send open questions to Canvas Chat

PRACTITIONERS - TENTATIVE & SUBJECT TO CHANGE

Jason Boyce, Entrepreneur and CEO with nearly 20 years of e-commerce and Amazon Marketplace experience, founded Avenue7Media.com in 2019 and co-founded Dazadi.com in 2002.

OFFICE HOURS

You can meet with me after the virtual class ends on Thursday nights or Saturday afternoons through Zoom. If you need to meet outside this timeframe, Saturday and Sunday mornings work best for me. You need to email me to schedule the meetings beforehand so that I could reserve the time.

COURSE DESCRIPTION

Course Unit – One Full Unit

Prerequisites – NA

Catalog – This blended course will cover the fundamental concepts and practical applications of text mining and analytics. Students will explore the process pipeline from converting unstructured text to structured data, and, leveraging Natural Language Processing to extract useful and interesting information such as text categories, text topics, trends and user sentiments. Students will learn to analyze results and gain practical skills through the use of Python.

COURSE MATERIALS

Text Analytics with Python: A Practitioner's Guide to Natural Language Processing (Second Edition) by Dipanjan Sarkar, 2019, Apress. ISBN: 978-1-4842-4354.

Software: Setup your Python environment from *Anaconda*. Anaconda Python distribution has a complete Python distribution with over hundreds of packages specially built for data science and AI. You can either use *Spyder* from the Anaconda package or download *PyCharm* CE with Anaconda plugin to write and execute your code (they are both IDEs for Python).

Download Anaconda Python 3.7 from https://www.anaconda.com/products/individual

COURSE REQUIREMENTS

This fully remote course will take place entirely remotely with no in-person meetings. This course will utilize live Zoom sessions. When lecturing or providing instruction, the class may be recorded. During this time, only the active speaker and the shared screen will be recorded. Breakout rooms, labs, and casual conversations will not be recorded. *You are not allowed, at any time, to record in part or in whole any class meeting or course image. This includes taking pictures of your screen.*

If you become ill during this class or are deeply affected by someone who becomes ill, notify me as soon as possible through the Canvas inbox. We will work something out. If I become ill, the school will notify you and make some accommodations. Life is extremely uncertain right now. Let us put our best foot forward and hope for the best.

Classes will be devoted to covering materials from the prescribed text. However, since text mining is an emerging topic, additional up-to-date study materials could be used at the discretion of the instructor to enhance the quality of learning. The instructor will emphasize the general concepts, techniques, and demonstrations during class and will reinforce the techniques through hand-on exercises. Learning will be assessed through quizzes, programming assignments, and group project. You will also participate in regular discussions both in class and in an online discussion forum. Periodically, you will present your work to the whole class.

The following describes the type of work you will complete in this course.

Class Participation

Class participation will be a qualitative measurement made by me about your participation in the course. Participation may include involvement in in-class discussions, presentations, online discussion forum, and individual conversations with me. For presentations, each student will have a chance to present your programming assignment to the whole class; you will also need to present your processes and results of group project. For online discussion forum on Canvas, you will utilize it to discuss about upcoming assignments, and share your questions and thoughts about text mining concepts and coding with Python.

Assessment will focus on quality and not quantity. You are expected to be prepared before each class, open your camera and interact with me and others in a respectful manner online.

Quizzes

There will be three quizzes given on Canvas during the semester. Each quiz includes only multiple-choice questions about basic text mining concepts. The goal of these quizzes is to ensure that you have reviewed class materials and know how to build and interpret text models.

Programming Assignments

The programming assignments are designed to provide you with an opportunity to practice using Python to do real text mining cases. All assignments will be submitted to Canvas before due date. 5% of the total possible points per late day will be deducted for late submissions.

Assignments will be presented by students and discussed in class after due dates. Each assignment should be done individually, which means you should NOT discuss with others personally or share your codes with others. You can only use our specific online discussion forum to discuss about upcoming or past assignments. Review the policy on Academic Integrity listed below for more information on this issue. Do not wait until the last minute to do the assignments. Start the assignments early so you can ask questions and get help in a timely manner.

Group Project

During the semester, you will be required to complete a series of tasks that lead into a final group project. The fundamental purpose of this project is to allow you the opportunity to be creative and methodical in retrieving and analyzing text data. Please work hard as a team (two to three students) on this project and take this effort very seriously. The following tasks will be included.

1. Find your own text data from www.kaggle.com or collect text data that you are interested in by yourselves. Twitter data is highly recommended.
2. Define the questions you want to get answers from text data. Your questions should be specific and cover at least two of the major text mining models, which may include text classification, text topics, text similarity, sentiment analysis, and contextual text mining.
3. Retrieve relevant text data, clean the data, and preprocess the data for your project.
4. Perform text mining techniques to answer your questions.
5. Write a report. The following sections should be included in the report:
    a. Introduction – Identify the resource of your text data, data background, why you select these data, your research questions and motivation.
    b. Analysis of text data – Describe how you analyze the text data, show the corresponding results, and interpret the results. Don't paste your code here.

c. Conclusion – Summarize the main findings of your research and indicate the practical implications/contributions from your findings. You may also include limitations and future research interests.

d. Appendix – Attach your Python code in this section.

6. Present your project to the whole class. Your presentation should be focused and meet requirements posted to Canvas including time limits. Prepare slides and try to use graphs or tables that quickly build understanding.

COURSE PURPOSE, LEARNING GOALS AND ASSESSMENT PLANS

Applied to a corpus or body of information, text mining can be used to make large quantities of unstructured data accessible and useful by extracting useful information hidden in text content and revealing patterns, trends and insight in large amounts of information. Specifically, this course allows students to understand NLP and text syntax, semantics and structure, discover text cleaning and feature engineering, review text classification and text clustering, assess text summarization and topic models, and leverage text mining techniques in Python. Students will work individually and in small teams to apply the knowledge and skills gained in this course to analyze and solve actual case studies. Student will present their work periodically during the semester including a comprehensive and final project presentation to the whole class.

Learning Goals

Upon completion of this course, students will be able to demonstrate mastery in the following areas:

1. Demonstrate an understanding of current analytics techniques used with textual data to discover patterns, extract knowledge, and support decision making.
2. Apply statistical machine learning models to understand textual data and predict future outcomes.
3. Critique current text analytics technologies and modeling techniques and communicate issues relevant to their effective implementation and operation.

Assessment Plan for Learning Goals

To assess student knowledge after completing this course, students will be able to:

1. Learning Outcome #1 – Participate in online discussions and presentations, complete quizzes about basic text mining concepts;
2. Learning Outcome #2 – Practice using Python to do real text mining cases, finish programming assignments individually;
3. Learning Outcome #3 – Develop a comprehensive research paper in small teams, discuss and defend the techniques and findings in a formal presentation.

For learning outcome #3, rubric has been developed and implemented in the course management system.

COURSE SCHEDULES

The tentative schedule of topics and student activities is listed below. The master schedule can be found in Canvas.

Disclaimer

Depending on the needs of the class, unplanned complications, and resources available to us, the master schedule may change during the progression of the course. As with all contemporary institutions, we need to remain nimble and react to evolving environmental issues.

Note: All online meetings occur on Thursday nights from 6:30pm to 8:50pm and may be preceded or followed by office hours. *Campus meetings on Saturday afternoons have been changed to online meetings* due to the recent pandemic.

Date	Class Plans	Submission/ Assessment
INTRO Aug 25	No class	Online Discussion
Week 1 Aug 31	Thursday: First Class - Introduction to Text Mining; Download and Install Python	
Week 2 Sep 7	Thursday: Python Syntax and Structure, Regular Expressions Saturday: Python Demonstration and Practice; Find text data for group project	Quiz 1
Week 3 Sep 14	Thursday: Natural Language Processing (NLP) Basics	Programming Assignment 1 – Information Extraction (Due Thursday)
Week 4 Sep 21	Thursday: Processing and Understanding Text Saturday: Student presentation on 1st assignment; Define research questions for group project	
Week 5 Sep 28	Thursday: Text Representation	Programming Assignment 2 – Basic NLP (Due Thursday)

Date	Class Plans	Submission/ Assessment
Week 6 Oct 5	Thursday: Fall break, no class Saturday: Student presentation on 2nd assignment; Text data preparation for group project	
Week 7 Oct 12	Thursday: Word Association Mining (Paradigmatic relation and Syntagmatic relation); Text Similarity	Quiz 2
Week 8 Oct 19	Thursday: No class Saturday: Student presentation on 3rd assignment; Perform text processing techniques to deal with your text data for group project	Programming Assignment 3 – Advanced NLP (Spelling Recommender), Document Similarity (Due Thursday)
Week 9 Oct 26	Thursday: Topic Models	
Week 10 Nov 2	Thursday: Text Classification Saturday: Student presentation on 4th assignment; Perform text mining techniques to deal with your text data for group project	Programming Assignment 4 – Topic Modeling (LDA) (Due Thursday)
Week 11 Nov 9	Thursday: Sentiment Analysis; Text-based Prediction and Contextual Text Mining	
Week 12 Nov 16	Thursday: No class Saturday: Student presentation on 5th assignment; Group project discussion and finalization	Programming Assignment 5 – Text Classification (Due Thursday)
Week 13 Nov 23	Thanksgiving, No class	
Week 14 Nov 30	Thursday: Last class - Group project presentation; Course Wrap up	Quiz 3 Group project report

Grading

A weighted scale will be used to calculate final grades.

Category	Percentage of Grade
Class Participation (Qualitative Measure)	10%
Quizzes (*3)	15%
Programming Assignments (*5)	50%
Group Project	25%

Calculating Final Grades

Your grade for the course will be determined according to the following scale.

Final Grade	Average Points	Final Grade	Average Points
A	94-100	C+	77-79.99
A-	90-93.99	C	74.76.99
B+	87-89.99	C-	70-73.99
B	84-86.99	D+	66-69.99
B-	80-83.9	D	60-65.99

Course Policies and Shared Expectations

The College of New Jersey has developed shared policies that apply to all of our courses. Click here for a complete listing of these policies. In addition, here is a link to the Graduate Bulletin where you can find contact information and services regarding graduate education. Here are policies designed for this course.

Academic Integrity – You are responsible to know the Academic Integrity policy published by the college. You may only represent work that is your own. Cheating on tests, failing to cite sources, or submitting someone else's work are just a few examples that may result in failing the entire course or dismissal from the college. In addition to academic performance, you are expected to demonstrate the qualities of honesty and integrity. All submissions by you or your team are expected to be your original work. Material that, in any way, volatiles this principle, or any form of dishonesty, cheating, fabrication, the facilitation of academic dishonesty, and/or plagiarism, may result in you receiving a failing grade for the assignment, quiz, test, or the course. In addition, further appropriate disciplinary action may be initiated.

Attendance – During the weeks where we do not have scheduled Saturday classes, plan to attend weekly synchronous lectures. You are expected to attend all classes and lectures. Dates are provided in the course schedule. If you must miss a class, send an email to me beforehand stating the reason for missing it. I realize you work full-time and may be pulled away from this program because of it, but those occasions should be rare. At times, you will also be expected to attend team meetings. These meeting times should be scheduled by your team at a time that works for all of you.

Contact Information – The best way to contact me is through the Canvas Inbox. Notification is quick and your message won't get mixed up with other email. You can also send me an email directly, and please include your name and class in the body of the message since it may not be obvious from your e-mail address. I will respond to you within 24 hours.

Course Evaluations – I believe in continuous improvement and I learn about potential improvements from your course evaluations. At the end of the semester, you are expected to complete an evaluation for this course. Please take the time for this task and provide textual feedback whenever you can.

Faculty Schedule & Availability – There will be time towards the end of our weekly virtual meeting and our scheduled Saturday meetings to talk. You can also schedule a personal Zoom meeting. Send me an email through the Canvas Inbox suggesting a date and time to speak with you. Generally, I am available during weekend mornings.

Getting Technical Support – For problems with course projects, contact me through the Canvas Inbox. For immediate problems with campus systems, contact the Help Desk at (609) 771-2660, visit https://tcnj.teamdynamix.com/TDClient/Home/, or send an email to mbahelp@tcnj.edu describing technical issue.

Teams – Building relationships with your cohorts can be invaluable to you. In addition, businesses rely heavily on teams working cooperatively to work on the type of projects you learn in this course. You will be assigned to work in a small team to complete a project. Each team member will be responsible to contribute to the group project and someone from that team should be assigned the role of Project Leader. Your team will be assigned virtual space in Canvas that only your team and I have access to. You can use this space to email team members, store and collaborate on team documents, develop study guides, leave messages, and document group activities and meeting times.

If you experience problems with your team, you should seek my advice quickly. I maintain the right to remove a team member or members from a group and either place them in another group or require that they complete the work independently. You may request to be reassigned to another team during the semester. This request will be honored when reasonable and possible.

Writing - Because writing is a fundamental business skill, your grade for each assignment will reflect, among other things, your ability to write, even for assignments with minimum writing. Feedback on your writing will be provided as deemed necessary and, if your writing needs improvement, you should seek help from someone who writes well or some other writing source. The responsibility to write well is yours. My responsibility is to hold you accountable for how well you write. Poor writing will be reflected in your final grade.

ZOOM MEETINGS

Most of our classes will be held online. To ensure that these classes run well, you need to follow these rules.

1. If you are not familiar with Zoom, go to this link and acquaint yourself with this application.
2. Be on time and, if you need to be late, notify me beforehand through the course email system. Our meeting software alerts us when someone joins a meeting so I will need to minimize the disruptions your lateness will cause in the meeting.
3. Keep your video on so that we can see you. Keep your microphone muted unless you are speaking.
4. Prepare yourself and your work area for a meeting. It should be free from external distractions (e.g., family members, pets, phones) and, if you need water, make sure you have it beforehand. Take necessary breaks before the meeting; your full attention is needed, and expected, in this class.
5. Keep a professional demeanor at all times. The tone of your voice and the words you used should be the same as if you were participating in an important business meeting. We have limited time, so your stories and discussions points need to be concise and to the point.
6. When we are having a discussion, you need to participate without dominating the conversation. All students will be expected to contribute to our meetings and be respectful of their peers' contributions. Before you speak, state your name so that your peers and I know who to respond to.
7. If you are asked to prepare a slide for a discussion, use fonts, color and layout effectively so that the slide is easy to read by others. Practice your presentation beforehand to ensure your words and ideas are easily understood by others. Readings certain facts and figures from your slide is fine but you should not need to read it word for word.
8. When someone else is speaking, wait until they are finished before you comment or follow-up on an idea. You can raise your hand in Zoom. When I see multiple raised hands, I will generally call on you in the order the hands were raised.