Syllabus: CAP 4770 Data Mining and Text Mining

Fall semester 2021

Course Information

- **Course Number and Title**: CAP 4770 Data Mining and Text Mining
- **Credit Hours**: 3 credits
- **Current Academic Term**: FALL 2021

Instructor Information

- **Instructor**: TBA
- **Office**: TBA
- **Office Hours**: TBA
- **Office Phone**: TBA
- **E-mail**: TBA@floridapoly.edu
- **Class Meeting**: Tuesday/Thursday 5:30 PM - 6:45 PM, IST 1015

Course Details

- Delivery Mode: The class will be delivered in a face-to-face format where students are expected to attend all of their scheduled University classes and to satisfy all academic objectives as defined by the instructor.

- **Course Website**: https://floridapolytechnic.instructure.com/courses/5495

- **Official Catalog Course Description**:
  This course addresses the knowledge discovery process and the use of data mining concepts and tools as part of that process. In depth analysis of processes for extracting useful unknown information from data sources and using the information to make decisions is also covered.

- **Prerequisites**: COP 3710 Database 1 and (QMB 3200 Advanced Quantitative Methods or MAS 3114 Computational Linear Algebra)

- **Communication/Computation Skills Requirement (6A-10.030)**: No

- **Required Texts**:
  "Text Mining with R: A Tidy Approach" by Julia Silge and David Robinson
  https://www.tidytextmining.com/

  "R for Data Science" by Garrett Grolemund and Hadley Wickham
  https://r4ds.had.co.nz/

  "Data Mining: Examples and Case Studies" by Yanchang Zhao
  http://www2.rdatamining.com/uploads/5/7/1/3/57136767/rdatamining-book.pdf

- **Equipment and Materials**:
  We will use the R programming language and RStudio. Both of these are free. The course covers fundamental and popular R packages for data mining and text mining, introduced as working examples. The format of the course will include lectures by the instructor, class discussions, directed readings, and students' presentations.

  Suggested: You can simply create an account in https://rstudio.cloud to access a cloud-based version of RStudio.
  Alternative: a local installation of R can be completed by downloading R from the R Project web site. After installing R, a free and open-source Integrated Development Environment (IDE) for R can be downloaded from the RStudio web site.

- **Course Objectives:**
  This course covers principles, concepts, and methods in the fields of data mining and knowledge discovery. Algorithm development, current tools, and real-world applications are explored. Topics include: data visualization, exploration, clustering, classification, association rule mining, and anomaly detection, among others.

- **Course Learning Outcomes**:
  Upon successfully completing this course, learners will be able to:
  1. **Explain** principles, concepts, methods, and techniques and trends in the fields of data mining, and knowledge discovery. *(Comprehension)*

  2. **Apply** transformations to data, including discretization to numeric attributes, numeric coding of nominal attributes, data preprocessing and data cleansing techniques, and selection of appropriate features. *(Application)*

  3. **Apply** clustering algorithms, including iterative clustering, hierarchical clustering, and probability-based clustering, and density-based cluster analysis; and generate rules by induction, classification rules, and association rules. *(Application)*

  4. **Extract** knowledge from unstructured text collections, and demonstrate the use of sequence mining algorithms to discover patterns across time or positions in a given dataset. *(Synthesis)*

  5. **Assess** the significance of mined frequent patterns, as well as the association rules derived from them by using rule and pattern assessment measures and employing algorithms to test for the statistical significance of rules and patterns. *(Evaluation)*

- **Alignment with Program Outcomes**:

	Course Learning Outcome				
Computer Science ABET Student Outcomes	1	2	3	4	5
(1) Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.	X	X			
(2) Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program's discipline.	X	X	X	X	X
(3) Communicate effectively in a variety of professional contexts.	X				X
(4) Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.	X			X	
(5) Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline.	X				
(6) Apply computer science theory and software development fundamentals to produce computing-based solutions.		X		X	X

	Course Learning Outcome				
Data Science Program Student Outcomes	1	2	3	4	5
(1) Apply current data science concepts, techniques, and practices to solve complex problems.	X	X	X	X	
(2) Analyze a given data science problem and formulate a solution in terms of the datasets needed, the techniques required or the technologies to be utilized.	X	X		X	X
(3) Communicate effectively insights, analysis, conclusions, or solutions to a diverse audience.	X				X

	Course Learning Outcome				
Business Analytics Program Student Outcomes	1	2	3	4	5
(1) Apply current business analytics concepts, techniques, and practices to solve business problems.	X	X	X	X	
(2) Analyze a given business problem using appropriate analytics techniques to generate insights and solutions.	X		X		X
(3) Communicate effectively insights, analysis, conclusions, and solutions to a diverse audience.	X				X

Academic Support Resources

- **Library**: Students can access the Florida Polytechnic University Library through the University website and Canvas, on and off campus. Students may direct questions to Academic Success Center success@floridapoly.edu or by email, library@floridapoly.edu.
- **ASC**: The Academic Success Center, located in the IST and at ASC East, provides a range of services. Students may direct questions to success@floridapoly.edu.

Course Policies

Attendance

- Please see **University Policy**, which reads "Students are expected to attend all of their scheduled University classes and to satisfy all academic objectives as defined by the instructor." Attendance in this environment does not, of course, mean actual physical attendance in the classroom, although it may include that.

- If you know that you will miss a class for any reason discuss the situation with your instructor in a timely manner.

- **Students Feeling Sick: I am a student; what should I do if I think I may have COVID-19?**
  Students who are showing symptoms or who have been exposed to COVID-19 are expected to stay in their residences (at home or in their dorm rooms) and immediately notify the FL Poly CARE manager at care@floridapoly.edu. The CARE manager will work with each student to triage their individual situation and will notify faculty of students who are not attending courses due to COVID-19 symptoms.

Late Work/Make-up work

- Each student must keep current on assignments. Late assignments are not graded, unless permission has been obtained from the instructor. In case of a medical emergency, please notify your instructor as soon as possible who will evaluate any exceptions on a case by case basis.

Grading Scale

- Grades will be determined according to the following scale:

A	93% – 100%	B	83% – 85%	C	73% – 75%	D	63% – 65%
A–	90% – 92%	B–	80% – 82%	C–	70% – 72%	D–	60% – 62%
B+	86% – 89%	C+	76% – 79%	D+	66% – 69%	F	0% – 59%

Assignment/Evaluation Methods

- Participation in all course activities is a very important element of this course and is a basic expectation. Course participation consists of active and respectful involvement in class discussions, presentations, peer feedback, postings, replies, projects, and other interactions. The discussion grade considers quality, quantity, and timeliness of student participation.

Assignment	*Percentage*
Discussions	10%
Midterm Exam	20%
Final Exam	20%
Quizzes	15%
Homework	20%
Final Project	15%
Total	100%

A note on the Final Project
In the final project you will show your knowledge and skills in data mining and text mining, using any combination of the different tools and topics discussed throughout the semester applied to an area/field of your interest.

- Final Project Report
  Your goal is to submit a cohesive project report that conveys that you have mastered the techniques discussed during the semester.
- Final Project Presentation
  You will present your final project and summarize your findings. The final project presentation accounts for 15% of your final project grade.

Your instructor will provide you with specific guidelines for the final project report and final project presentation shortly after the first few weeks of classes (format and length, call for proposals, reference materials, presentation guidelines and logistics, rubric, etc.)

Sample final project topics from previous years include*:*

- Text mining for analysis of topics discussed in social media platforms
- Finding patterns in performance of recent winning sports teams
- Clustering and recommendation algorithms for video streaming services
- Analysis of purchasing patterns for retail customers
- Sentiment analysis of lyrics from top songs in recent years
- Characterization of street network spatial features
- Clustering of traffic crashes and their relationship with inclement weather

University Policies

Basic rules for in the classroom, IST, and Campus

1. We highly recommend, until further notice, that you wear your face-covering during class and throughout the building at all times.
2. Absolutely **no eating or drinking** during class.

Reasonable Accommodations

Florida Polytechnic University is committed to assisting students with disabilities and offering reasonable accommodations to those with documented eligibility. The Office of Disability Services (ODS) coordinates accommodations for students with disabilities in accordance with the ADA Amendments Act of 2008 (ADAAA), the Americans with Disabilities Act of 1990 (ADA), and Section 504 of the Rehabilitation Act of 1973. Reasonable accommodations are determined on an individual basis through an interactive process between you, ODS, and your instructor(s). If you have already registered with ODS, please ensure that you have requested an accommodation letter for this course and communicate with your instructor about your approved accommodations at your earliest convenience. If you are not registered with ODS but believe you have a temporary health condition or permanent disability requiring an accommodation, please contact ODS as soon as possible.
The Office of Disability Services (ODS):
DisabilityServices@floridapoly.edu
(863)874-8770
ASC East building
ODS website: www.floridapoly.edu > Student Affairs > Health Wellness > Disability Services

Accommodations for Religious Observances, Practices and Beliefs

The University will reasonably accommodate the religious observances, practices, and beliefs of individuals in regard to admissions, class attendance, and the scheduling of examinations and work assignments. (See University Policy.)

Title IX

Florida Polytechnic University is committed to ensuring a safe, productive learning environment on our campus that prohibits sex discrimination and sexual misconduct, including sexual harassment, sexual assault, dating violence, domestic violence and stalking. It is important for you to know that there are resources available if you or someone you know needs assistance. You may speak to your professor, but your professors have an obligation to report the incident to the Title IX Coordinator. It is an educational goal that you feel able to share information related to your life experiences in classroom discussions and in one-on-one meetings. However, it is requirement for university employees to share information with the Title IX Coordinator regarding disclosure. However, please know that your information will be kept private to the greatest extent possible. You will not be required to share your experience. If you want to speak to someone who is permitted to keep your disclosure confidential, please seek assistance from the Florida Polytechnic University Ombuds Office, BayCare's Student Assistance Program, 1-800-878-5470 and locally within the community at Peace River Center, 863-413-2707 (24-hour hotline) or 863-413-2708 to schedule an appointment.

Academic Integrity

All students must commit to the highest ethical standards in completion of all academic pursuits and endeavors, whether in classroom or online environments: Academic Integrity.

Student Record of Lectures

Students may, without prior notice, record video or audio of a class lecture for a class in which the student is enrolled for their own personal educational use.

Recordings may not be used as a substitute for class participation or class attendance. Recordings may not be published or shared in any way, either intentionally or accidently, without the written consent of the faculty member. Failure to adhere to these requirements is a violation of state law (subject to civil penalty) and the student code of conduct (subject to disciplinary action).

Recording class activities other than class lectures, including but not limited to lab sessions, student presentations (whether individually or part of a group), class discussion (except when incidental to and incorporated within a class lecture), and invited guest speakers ***is prohibited****.* For further information, go to the Registrar's webpage and click on HB233 Guidance.

Course Schedule

- I reserve the right to modify this schedule as required by the progression of the class.
- Exercises listed as suggested problems may be used as part of some of the homework assignments and/or quizzes this term.
- Coursework is due at 11:59PM Eastern Standard Time (EST) on the date indicated.
- A tentative course calendar is included below.

Week/Date	Lesson/Topic	Suggested Problems and Readings	Assignments (tentative)
1 Aug. 24 – Aug. 29	***Overview of data mining and text mining*** * RStudio setup (packages, rstudio.cloud option) * The nature of data * Canvas and other course resources	https://rstudio.cloud/learn/primers/1	
2 Aug. 30 – Sept. 5	***Data collection and pre-processing*** * Algorithms overview * Normalization and Standardization Examples and applications	https://www.reisanar.com/courses/dmtm/index.html#data-exploration https://www.reisanar.com/courses/dmtm/index.html#data-preparation	HW 1
3 Sept. 6 – Sept. 12	***Dimensionality Reduction*** * Fundamentals of linear algebra * Principal Component Analysis. * Summarizing/aggregating data	https://www.reisanar.com/courses/dmtm/index.html#matrix-factorization https://www.reisanar.com/courses/dmtm/index.html#pca-1	HW 2 Discussion 1
4 Sept. 13 – Sept. 19	***Modern Tools for Data Analysis*** * The `tidyverse` family of packages * Data wrangling and Data Visualization fundamentals	https://www.reisanar.com/courses/dmtm/index.html#pca-2 https://www.reisanar.com/courses/dmtm/index.html#pca-3 https://rstudio.cloud/learn/primers/2	Quiz 1
5 Sept. 20 – Sept. 26	***Association Rules*** * The Apriori algorithm * Rules generation & interpretation.	https://www.reisanar.com/courses/dmtm/index.html#apriori-1	HW 3 Discussion 2
6 Sept. 27 – Oct. 3	***Anomaly detection*** * Motivation and algorithms * Modern methods	https://www.reisanar.com/courses/dmtm/index.html#apriori-2	Quiz 2
7 Oct. 4 – Oct. 10	***Clustering*** * k-means and related methods * hierarchical methods	https://www.reisanar.com/courses/dmtm/index.html#clustering	

Week/Date	Lesson/Topic	Suggested Problems and Readings	Assignments (tentative)
8 Oct. 11 – Oct. 17	***Clustering*** * Density-based methods * Fuzzy clustering, spectral clustering	https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/	HW 4 Discussion 3
9 Oct. 18 – Oct. 24	***Text, web and social media analytics*** * Motivation and modern applications * Classical definitions and methods	https://juliasilge.com/blog/tidy-text-classification/	Quiz3
10 Oct. 25 – Oct. 31	***Tidy text format*** * Introduction to tidy text mining tools * Tidy principles and data transformation	https://cran.r-project.org/web/packages/tidytext/vignettes/tidytext.html	Mid-Term 1&2 (tentative)
11 Nov. 1 – Nov. 7	***Sentiment analysis*** * Applications of sentiment analysis * Lexicons and tokenization	https://www.tidytextmining.com/sentiment.html	
12 Nov. 8 – Nov. 14	***Term frequency and inverse document frequency*** * Text analytics * Visualization tools and best practices	https://www.tidytextmining.com/tfidf.html	Discussion 4
13 Nov. 15 – Nov. 21	***n-grams and correlations*** * Relationships between words using n-grams * Network visualization	https://juliasilge.com/blog/word-associations/	Quiz 4
14 Nov. 22 – Nov. 28	***Topic modeling*** * Latent Dirichlet allocation (LDA)	https://www.tidytextmining.com/topicmodeling.html	Quiz 5
15 Nov. 29 – Dec. 5	***Project Presentation*** ***Final Review***		
16 Dec. 6 – Dec. 16	***Final Exam***		

Final Project

In the final project you will show your knowledge and skills in data mining and text mining, using any combination of the different tools and topics discussed throughout the semester applied to an area/field of your interest.

- Final Project Report
  Your goal is to submit a cohesive project report that conveys that you have mastered the techniques discussed during the semester.
- Final Project Presentation
  You will present your final project and summarize your findings. The final project presentation accounts for 15% of your final project grade.

Your instructor will provide you with specific guidelines for the final project report and final project presentation shortly after the first few weeks of classes (format and length, call for proposals, reference materials, presentation guidelines and logistics, rubric, etc.)

Sample final project topics from previous years include*:*

- Text mining for analysis of topics discussed in social media platforms
- Finding patterns in performance of recent winning sports teams
- Clustering and recommendation algorithms for video streaming services
- Analysis of purchasing patterns for retail customers
- Sentiment analysis of lyrics from top songs in recent years
- Characterization of street network spatial features
- Clustering of traffic crashes and their relationship with inclement weather

Important Dates

September 6	M	Labor Day Holiday - No Classes
October 18	M	Mid-term Grades Due
November 11	W	Veteran's Day Holiday (Observed) - No Classes
November 23	W	Withdrawal Without Academic Penalty Deadline (W assigned)
November 24-26	W-F	Thanksgiving Holiday Break - No Classes
December 8	W	Last Day of Classes
December 4-5	Th-F	Reading Days - No Classes
December 11, 13-16	S, M-Th	Final Exams
December 17	F	End of Semester
December 20	M	Final Grades Due from Faculty by 4 p.m.
December 22	W	Final Grades Available Online

Sample Rubric for Report and Presentations

The final presentations and reports will be evaluated using the rubrics included below.

Sample Report Rubric

Objective	Category	Below Expectations	Weak	Average	Good	Excellent
	Score	**1**	**2**	**3**	**4**	**5**
Students can write professional quality documents	Introduction	Opening is off-topic and inappropriate to the purpose, not concise and no clarity	Opening is somewhat related to the topic and appropriate to the purpose but is not concise and clear	Opening is related to the topic and appropriate to the purpose. Somewhat clear and concise	Opening is related to the topic and appropriate to the purpose. Clear and concise	Strong opening that is clear and concise
	Organization	Disorganized; incorrect format; unclear direction	Somewhat organized; incorrect format; unclear direction	Organized; correct format; unclear direction	Organized; correct format; clear direction	Correct formatting, strong clarity and organization in the development of main points
	Literature Review	Does not present information from any source	Presents information from irrelevant sources representing limited points of view/approaches	Presents information from relevant sources representing limited points of view/approaches	Presents in-depth information from relevant sources representing limited points of view/approaches	Synthesizes in-depth information from relevant sources representing limited points of view/approaches
	Research Design (weighted twice)	Does not provide information on research design	Inquiry design demonstrates misunderstanding of the methodology or theoretical framework	Critical elements of the methodology or theoretical framework are missing, incorrectly developed or unfocused	Critical elements of the methodology or theoretical framework are appropriately developed however, more subtle elements are ignored or unaccounted for	All elements of the methodology or theoretical framework are skillfully developed and may be synthesized from across disciplines or relevant subdisciplines
	Analysis (weighted twice)	Incorrect, Irrelevant, no supporting evidence	Correct, irrelevant, no supporting evidence	Correct, relevant, no supporting evidence	Relevant and correct with supporting evidence	Relevant, correct, complete, incorporates innovative insights
	Next Steps	Missing or content does not support conclusion	Conclusion irrelevant to the findings	Conclusion somewhat relevant to the findings	Conclusion relevant to the findings	Strong conclusion that is clear, complete and compelling
	Grammar & Spelling	Uses language that often impedes meaning due to errors	Uses language that often sometimes meaning due to errors	Uses language that generally conveys meaning to readers with clarity, although writing includes some errors	Uses straightforward language that conveys meaning to readers. Language has few errors	Uses graceful language that communicates meaning to readers with clarity and fluency and is virtually error free
	Reference Style (APA)	Did not follow APA style	Numerous errors in APA style, did not cite sources correctly, formatting issues	Some errors in APA style, cited correctly but formatting issues persist	Minimum errors in style and formatting but does not detract from readability	No errors in APA style
Total points for Report = 50						

Presentation Rubric

Objective	Category	Below Expectations	Weak	Average	Good	Excellent
	Score	1	2	3	4	5
Students can demonstrate mastery of communication technology	Use of Media	Lack of media detracts from the presentation objective	Misuse of media that detracts from the presentation objective	Use of media barely supports and contributes to the presentation objective	Use of media supports and contributes to the presentation objective	Use of media supports, clarifies and reinforces the presentation objective
	Quality of Slides	Very poor quality. Not enough or too much colors, fonts and animations that detract from project objective	Poor quality. Not enough or too much colors, fonts and animations that detract from project objective	Fonts, colors and animations barely support the presentation objective	Fonts, colors and animations support the presentation objective	Fonts, colors and animations support, clarify and reinforce the presentation objective
Students can develop and deliver a compelling oral talk with relevant facts and information	Opening statement	Opening is off-topic and inappropriate to the purpose, not concise and no clarity	Opening is somewhat related to the topic and appropriate to the purpose but is not concise and clear	Opening is related to the topic and appropriate to the purpose. Somewhat clear and concise	Opening is related to the topic and appropriate to the purpose. Clear and concise	Strong opening that is clear and concise
	Organization	Disorganized; incorrect format; unclear direction	Somewhat organized; incorrect format; unclear direction	Organized; correct format; unclear direction	Organized; correct format; clear direction	Correct formatting, strong clarity and organization in the development of main points
	Literature Review	Does not present information from any source	Presents information from irrelevant sources representing limited points of view/approaches	Presents information from relevant sources representing limited points of view/approaches	Presents in-depth information from relevant sources representing limited points of view/approaches	Synthesizes in-depth information from relevant sources representing limited points of view/approaches
	Analysis	Incorrect, Irrelevant, no supporting evidence	Correct, irrelevant, no supporting evidence	Correct, relevant, no supporting evidence	Relevant and correct with supporting evidence	Relevant, correct, complete, incorporates innovative insights
	Next Steps	Missing or content does not support conclusion	Conclusion irrelevant to the findings	Conclusion somewhat relevant to the findings	Conclusion relevant to the findings	Strong conclusion that is clear, complete and compelling
	Timing	Presentation is too short, insufficient coverage of material	Presentation is too long. Unable to cover all the material	Able to cover all the material within five extra minutes	Utilizes allotted time to provide sufficient coverage of material	Well-paced coverage of material within the allotted time
Students can deliver an oral talk with clarity and appropriate poise	Delivery Techniques	Does not participate in the oral presentation	Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) detract from the understandability of the presentation, and speaker appears uncomfortable.	Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) make the presentation understandable, and speaker appears tentative.	Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) make the presentation interesting, and speaker appears comfortable.	Delivery techniques (posture, gesture, eye contact, and vocal expressiveness) make the presentation compelling, and speaker appears polished and confident.
	Peer Evaluation	5 points				
Total Points = 50						