A picture containing logo

Description automatically generated

Syllabus: CAP 5771 Data Mining and Text Mining

**Fall semester 2023**

## Course Information

* **Course Number and Title**: CAP 5771 Data Mining and Text Mining
* **Credit Hours**: 3 credits
* **Current Academic Term**: **Fall 2023**
* **Class Meeting**: **Tuesdays and Thursdays 4:00 PM - 5:15 PM, ARC 1158**

## Instructor Information

* **Instructor**: Parisa Hajibabaee, Ph.D.
* **Office**: ARC-1107
* **Office Hours**: Tuesday, Wednesday, Thursday 1:00 – 2:00 PM or by appointment
* **Office Phone**: 863-874-8559
* E-mail: [phajibabaee@floridapoly.edu](mailto:phajibabaee@floridapoly.edu)

## 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/7253>
* 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: None
* Communication/Computation Skills Requirement (6A-10.030): No
* Recommended Texts:

*“*Data Mining and Machine Learning: Fundamental Concepts and Algorithms*”* by Mohammed J. Zaki and Wagner Meira, Jr

<https://dataminingbook.info/>

*“Data Mining: Examples and Case Studies”* by Yanchang Zhao

[http://www2.rdatamining.com/uploads/5/7/1/3/57136767/rdatamining-book.pdf](about:blank)

*“Dive into Deep Learning”* by Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J.

[https://d2l.ai/d2l-en.pdf](about:blank)

A classic by Hastie, Tibshiriani, and Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, available here: [https://hastie.su.domains/ElemStatLearn/](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fhastie.su.domains%2FElemStatLearn%2F&data=05%7C01%7Cnhbidoki%40floridapoly.edu%7C2f0902d0116540a774f608db8c547ae7%7C8d84067d9ad745729b10133d36462aaa%7C1%7C0%7C638258066710621022%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=NfOjN9O426uaSbdFSx2qe%2Bc9dEkGKLM4M04EmKp09bM%3D&reserved=0)

* Equipment and Materials:

We will use Python, and Google Colab. All are free. The course covers fundamental and popular Python and Python 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.

* 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. Use state-of-the-art principles in the field of data mining and text mining. *(Application)*
2. Evaluate clustering methods, association rule mining, and dimensionality reduction to real-world problems. *(Application)*
3. Analyze unstructured text data to produce valuable insights. *(Analyze)*
4. Design data-driven solutions to pattern recognition and data mining applied problems.

* **Alignment with Program Outcomes**:

|  |
| --- |
|  |
| Computer Science Student Outcomes | 1 | 2 | 3 | 4 |
| (1) Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions. | Applying | Applying |  |  |
| (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. | Applying | Applying | Applying | Applying |
| (3) Communicate effectively in a variety of professional contexts. | Synthesis |  |  |  |
| (4) Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles. | Synthesis |  |  | Synthesis |
| (5) Function effectively as a member or leader of a team engaged in activities appropriate to the program’s discipline. | Synthesis |  |  |  |
| (6) Apply computer science theory and software development fundamentals to produce computing-based solutions. |  | Applying |  | Applying |

|  |
| --- |
|  |
| Data Science Program Student Outcomes | 1 | 2 | 3 | 4 |
| (1) Demonstrate mastery in analyzing complex problems and applying knowledge of data science to formulate solutions. | Applying | Applying | Applying | Applying |
| (2) Communicate data science information clearly and effectively through presentations and technical writings to both expert and non-expert audiences. | Synthesis |  |  |  |
| (3) Demonstrate critical evaluation of recent research literature. | Synthesis |  |  |  |
| (4) Identify a novel relevant research problem in a chosen data science research field, perform the literature survey for the problem, create a plan to solve the problem, carry on the plan,  and defend the research. | Synthesis |  |  | Synthesis |
| (5) Recognize appropriate practices in the field of data science and their ethical implications. |  | Synthesis |  | Synthesis |

## 

## Academic Support Resources

* **Library**: Students can access the Florida Polytechnic University Library through the University website and [Canvas](about:blank), on and off campus. Students may direct questions to Academic Success Center [success@floridapoly.edu](about:blank) or by email, [library@floridapoly.edu](about:blank).
* **Peer Learning Strategists**: These are specially trained student leaders who help their peers strategize approaches to course content and work through solution methods. PLS students work in collaboration with the courses they support so the content and methods are aligned with your instructors’ expectations. The PLS room is located on the first floor of the IST in the center hallway.
* **Writing Center**: Located on the second floor of the IST (2059/2061), the Writing Center helps students to develop their writing and presentation skills. Consultations are available in person and virtually. For more detail, visit [https://floridapolytechnic.libguides.com/writingservices](about:blank).

## Course Policies

### Attendance

* Please see [University Policy](about:blank), 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.

### 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% |

### Grading Information Specifically for Graduate Students

* The grades of “A” through “C,” and “SR” are passing grades. The grades of "B-," "C+," and "C" are considered passing for graduate students but indicate weak performance for a graduate student and may not be accepted for some programs. The grades of "C-," "D+," "D," "D-," "F," and "UR" are failing grades.

### 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.

|  |  |
| --- | --- |
| *Assignment* | *Percentage* |
| Attendance and Participation | 5% |
| Homework | 10% |
| Quizzes | 5% |
| Final Project Presentation | 10% |
| Final Project Report | 10% |
| Exam 1 | 15% |
| Exam 2 | 15% |
| Final Exam | 30% |
| 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

Reasonable Accommodations

The University is committed to ensuring equal access to all educational opportunities. The University, through the Office of Disability Services (ODS), facilitates reasonable accommodations for students with disabilities and documented eligibility. It is the student’s responsibility to self-identify as a student with disabilities and register with ODS to request accommodations.

If you have already registered with ODS, please ensure that you have requested an accommodation letter for this course through the [ODS student portal](https://bear.accessiblelearning.com/FloridaPoly/) and communicate with your instructor about your approved accommodations as soon as possible. Arrangements for testing accommodations must be made in advance. Accommodations are not retroactive.

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

The Access Point

[ODS website: www.floridapoly.edy/disability](https://floridapoly.edu/student-affairs/health-wellness/disability-services.php)

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](https://floridapoly.edu/wp-content/uploads/FPU-3.009-Reasonable-Accommodations-6.22.17.pdf).)

## 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. Resources are available if you or someone you know needs assistance. Any faculty or staff member you speak to is required to report the incident to the Title IX Coordinator. Please know, however, 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. The Title IX Coordinator is available for any questions to discussion resources and options available.

## Academic Integrity

The faculty and administration take academic integrity very seriously.  Violations of [academic integrity regulation](https://floridapoly.edu/wp-content/uploads/2017/07/FPU-5.005-Academic-Integrity-7.29.14.pdf#search=academic%20integrity) include actions such as cheating, plagiarism, use of unauthorized resources (including but not limited to use of Artificial Intelligence tools), illegal use of intellectual property, and inappropriately aiding other students. Such actions undermine the central mission of the university and negatively impact the value of your Florida Poly degree. Suspected violations will be fully investigated, possibly resulting in an academic integrity hearing and sanctions against the accused student if found in violation. Sanctions range from receiving a zero on the exam or assignment, to expulsion from the university.  Repeat offenders are subject to more severe sanctions and penalties. Do not compromise your integrity for a perceived short-term gain.

## Recording 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.

## Course Schedule

* I reserve the right to modify this schedule as required by the progression of the class.
* Coursework is due at 11:59PM Eastern Standard Time (EST) on the date indicated.
* A tentative course calendar is included below.

| **Week** | **Lesson/Topic** | **Assignments**  (tentative) |
| --- | --- | --- |
| 1 | ***Overview of data mining and text mining***   * \* Introduction to data mining   \* Nature of data  \* Data cleaning  \* Data preprocessing  \* Google Colab and Python basic |  |
| 2 | ***Exploratory Data Analysis (EDA) and Visualization***  \* Google Colab and Python basic  \* Tell a story with your data  \* *Data manipulation*  \* *Data visualization*  \* Examples and applications using Python packages | HW 1 |
| 3 | ***Dimensionality Reduction***   * \*Review of linear algebra, covariance matrix, eigenvalues, and eigenvectors * \* Principal Component Analysis. (PCA)   \* Linear Discriminant Analysis (LDA)   * \* t- distributed stochastic neighbor embedding (t-SNE) * \* Examples and applications using Python packages * \* Final Project Reviews | * HW 2 |
| 4 | ***Frequent Pattern Mining & Association Rules***  \* The Apriori algorithm  \* The Eclat algorithm  \* The FP-Growth algorithm   * \* Rules generation & interpretation * \* Examples and applications using Python packages | Quiz 1 |
| 5 | ***Time-series Mining***  \* Decomposition  \* Autoregressive (AR)  \* Moving Average (MA)  \* ARMA  \* ARIMA   * \* Examples and applications using Python packages | * HW 3 * **Exam 1** (tentative) |
| 6 | ***Clustering***  \* k-means and related methods   * \* Hierarchical Clustering   \* Examples and applications using Python packages |  |
| 7 | ***Clustering***  \* Density-based methods  \* Spectral and Graph clustering  \* Examples and applications using Python packages |  |
| 8 | ***Anomaly detection***  \* Introduction and Motivation  \* Time Series Anomaly Detection  \* Clustering-Based Anomaly Detection   * \* Examples and applications using Python packages | * HW 4 |
| 9 | ***Text, web and social media analytics***  \* Motivation and modern applications  \* Classical definitions and methods  **\*** Python module re (Regular expression (RegEx) operations)  \* Examples using Python RegEx | * Quiz2 |
| 10 | ***Natural Language Processing (NLP)***  \* Intro to NLP  \* Basic text processing  \* Text normalization  \* Word normalization and stemming | * Exam 2 (tentative) |
| 11 | ***Word Embedding Techniques and Text Vectorization***  \* Relationships between words using n-grams  \* One-hot encoding  \* Term frequency and inverse document frequency (TFIDF)  **\* Examples and applications using Python packages** |  |
| 12 | ***Word Embedding Techniques and Text Vectorization***  \* Bag of Words (BOW)  \* Word2Vec  \* GloVe  \* Examples and applications using Python packages | Quiz 3 |
| 13 | ***Sentiment Analysis***  \* Applications of sentiment analysis  \* Text cleaning, tokenization, and stemming/lemmatization  \* Sentiment Lexicons like VADER and AFINN  \* Examples and applications using Python packages |  |
| 14 | ***Natural Language Processing (NLP) with Transformers***  \* BERT  \* Named Entity Recognition (NER)  \* Q&A | Quiz 4 |
| **15** | ***The power of NLP***  \* Generative Pre-trained Transformer (GPT) |  |
| **16** | * ***Final Presentations*** |  |
| **17** |

***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***

August 22-28, 2023 T-M Drop/Add Week

September 4, 2023 M Labor Day Holiday - No Classes

October 16, 2023 M Mid-term Grades Due

November 10, 2023 F Veteran’s Day Holiday (Observed) - No Classes

November 21, 2023 T Withdrawal Without Academic Penalty Deadline (W assigned)

November 22-24, 2023 W-F Thanksgiving Holiday Break - No Classes

December 6, 2023 W Last Day of Classes

December 7-8, 2023 Th-F Reading Days - No Classes

December 9, 11-14, 2023 S, M-Th Final Exams

December 15, 2023 F End of Semester/Diploma Date

December 18, 2023 M Final Grades Due from Faculty by 4:00 p.m.

December 20, 2023 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 | | | | | | |