HMG6583C/CAPS4792C DATA MINING WITH SOCIAL DATA

Department of Tourism, Hospitality, and Event Management
College of Health & Human Performance; University of Florida

INSTRUCTOR
Andrei P. Kirilenko, Ph.D.
Associate Professor
190B Florida Gym; 352.294.1648; andrei.kirilenko@ufl.edu
Office Hours: Thursdays, 10-12 pm

DEPARTMENT CHAIR
Rachel Fu, Ph.D.
Professor
Room FLG 240D; racheljuichifu@ufl.edu

COURSE DESCRIPTION

The course is designed as a foundational experience for upper-level undergraduate and graduate students from non-technical fields who want to acquire a basic understanding of data science and learn practical skills in data analysis. This course distinguishes itself by combining theoretical knowledge with practical applications, bridged through extensive Python programming exercises. To accommodate social science needs, the course emphasizes the analysis of textual data, especially those acquired from surveys and social media. For those without prior programming experience, the course provides instruction on using an AI-assisted Python programming tool, following the learn-by-doing methodology of acquiring new skills through experience. The overall learning goal of the course is to develop a conceptual understanding of data mining as well as the technical skills necessary for real-world data analysis.

The course combines lecture and lab instruction and is centered on building practical skills, requiring the students to complete a series of projects, concentrating on the analysis of tourism-related social network data. The students will learn the elements of programming (Python) required to automate data acquisition, storage, and analysis. Note that this is an introductory course, and many essential topics on Big Data such as distributed file systems, parallel computing, MapReduce, Hadoop, and similar are not covered; the CS “Introduction to Data Science” course is highly recommended as an elective for those students who want to get advanced knowledge in the subject.

COURSE OBJECTIVES

- Be able to use computational tools for data mining
- Be able to apply the basics of opinion analysis and sentiment analysis
- Identify tools to download and filter social network data from online sources
- Be able to develop tools for data analysis

By the end of the course, students should gain basic knowledge of data acquisition, pre-processing, and data mining techniques, including social media data, and be able to apply these skills to effectively carry out and present research projects in tourism and destination management.

PREREQUISITES

LEI4880 or HFT 4746 or HFT 4442 or HFT 4446C or APK4050 or STA 2023 or QMB 3250 or ADV3500 or ALS 3200C or ISM 3004 OR consent of the instructor based on taking courses on research methods, introductory statistics and data analysis.

COMPUTERS AND SOFTWARE

Personal computers are recommended. The UF computer labs will have the necessary software, but you must follow the labs' reservation schedule to complete your assignments. Note that I will provide instructions and help for Windows PC.

The course uses Python with Anaconda cloud programming environment with AI programming assistant. ***If you want to expand your capabilities in data mining by learning Python:***

- Register for the interactive Python course (I will provide advice on data mining libraries): https://www.codecademy.com/learn/python
- A good introductory course on data analysis with Python from Coursera will teach you popular tools such as pandas and matplotlib: https://www.coursera.org/learn/python-data-analysis . Select "audit" for a free course.

TEXTBOOKS

Required

- Andrei Kirilenko. Practical Data Mining with AI for Social Scientists. https://link.springer.com/book/9783031896880 ($79.99)

Note: Publication date has been pushed back by the publisher from June to October. To manage the setback, the instructor will provide pdfs of the relevant chapters.

Optional reading for deeper understanding

- Witten, Frank. Data Mining. Practical Machine Learning Tools and Techniques, 3rd ed. A hard copy from Amazon.com is approx. $25. ISBN-13: 9780123748560
- Kotu, Deshpande. Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner. E-book from the University library. Permalinks: https://bit.ly/2YXjMma; https://bit.ly/2MSHYjh
  ISBN-13: 9780128014608

- Tan, Introduction to Data Mining (Chapters 3, 7). Free download from http://www-users.cs.umn.edu/~kumar/dmbook/index.php ISBN-13: 9780133128901
- Foster Provost, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. ISBN-13: 9781449361327
- Jennifer Golbeck. Analyzing the Social Web. ISBN-13: 9780124055315
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. The free book is available online: http://www.nltk.org/book/ ISBN-13: 9780596516499
- Reza Zafarani, Mohammad Ali Abbasi, Huan Liu. Social Media Mining. An Introduction. The free book is available online: http://dmml.asu.edu/smm/SMM.pdf ISBN-13: 9781107018853
- Matthew A. Russell. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. Free older edition is available online: https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition ISBN-13: 9781491985045
- Al Sweigart, Automate the Boring Stuff with Python - Practical Programming for Total Beginners. Get for free at https://automatetheboringstuff.com/ ISBN-13: 9781593279929

ASSIGNMENTS AND EVALUATION

There will be home assignments, occasional quizzes, student presentations (only grad students), and term projects for this class. The total grade G (0-100%) will be a weighted mean of the grades in the following categories:

Grad students	Undergrad students
1. Lab assignments (20%) 2. Research paper presentation (20%) 3. Quizzes (20%) 4. Term Project (40%)	1. Lab assignments (25%) 2. Quizzes (25%) 3. Term Project (50%)

The final percentage points are translated into the letter grades using the following scheme:

Percentage	*Letter Grade*	*GPA*	*Percentage*	*Letter Grade*	*GPA*
90 – 100	A	**4**	**70 – 76.99**	C	2
87 – 89.99	B+	**3.67**	**67 – 69.99**	D+	1.33
80 – 86.99	B	**3**	**60 – 66.99**	D	1
77 – 79.99	C+	**2.33**	**Below 60**	E	0

If you notice a scoring error, notify the instructor within one week that a scoring error is made. **For the UF grading policies, see: https://catalog.ufl.edu/ugrad/current/regulations/info/grades.aspx**

Quizzes

An occasional short quiz will mostly cover the material presented after the previous quiz, but expect occasional questions related to earlier topics. **The quizzes will be closed-book and present a combination of multiple-choice and open-ended questions.** The 100% grade will require a correct answer to all questions; no answer or all incorrect answers will be graded 0%, and reasonable progress toward answering the questions will be evaluated in between.

Lab Assignments

The lab work has to be submitted through the course management system as an assignment. The acceptable submission format is a Word file. If the lab work is not finished in class, it has to be completed at home. A 100% grade will require full answers to all questions of the lab, no returned assignment will be evaluated 0%, and reasonable progress towards answering the questions will be evaluated somewhere in between.

Project

During the course, the students will work on group projects on a problem of their interest. The project should follow the steps outlined during the lectures: literature review, research design, data collection, data analysis, and research presentation. Project results should be presented in a research report (due **before** the date and time of the final exam as specified by the provost's office) AND an oral presentation.

50% of the project grade will come from the research report. Out of that, 20% will come from the literature review and research design parts, 20% from project results and result discussion, and 10% from the overall project presentation, including formatting, tables, and figures. Expect a full grade for using multiple sources of information to prepare your report, professional data analysis, and in-detail written presentation.

The other 50% of the project grade will come from the final oral presentation. Out of that, 20% will come from clearly presenting project ideas and outcomes, 20% will come from the overall quality of slides, and 10% will come from presentation professionalism: keeping within the allocated time, answering the questions, and distributing the presentation load across the entire project group.

Criteria and Weight	Excellent (100%)	Proficient (80%)	Satisfactory (60%)	Needs Improvement (40%)	Unsatisfactory (0%)
Written report					
Literature review 20%	Multiple relevant references introducing scientific publications relevant to the project (>5). Literature review leads to formulating research questions and explaining why the research is important.	Same, but few relevant references (3-5)	There are relevant references (1-2) but no clear connection to the research questions	No relevant scientific references, but references to non-science publications explain the research question	No references and/or poor presentation of the research questions

Project results and result discussion 20%	Results are clearly presented in a way that allows replicability of the study. When models are used model details and validation should be present. Discussion explains importance and relevance of the project in the context of reviewed literature.	Same, but discussion is incomplete.	Poor presentation of otherwise correct results and/or extremely poor discussion.	Same, but with errors in data analysis or poor analysis presentation in a way it prevents their understanding.	No results or no discussion
Overall project presentation 10%	Results and discussion are supported with tables and plots. The formal is uniform, clear, and easy to read. Tables and plots are with self-explanatory legends and titles. High-quality writing with proper sentences, paragraphs, and minimal grammar errors	Same, but with few drawbacks. For example, plots are not self-explanatory, or the format is hard to read or tables are messy (e.g., x=0.93843654632772). Good-quality writing with minor issues in sentences, paragraphs, and grammar	Same, but with multiple drawbacks. Adequate writing quality with noticeable issues in grammar and structure.	Very hard to read, missing tables and plots. Limited quality of writing with significant issues in grammar and structure	Nearly impossible to understand the project as written. Poor quality of writing, lacking proper sentences, paragraphs, and grammar
Oral presentation					
clearly presenting project ideas and outcomes 20%	outstanding presentation of project ideas and outcomes	Good effort and commendable quality of delivering the talk.	Adequate effort with room for improvement	Limited effort with considerable improvement needed	Minimal effort and significant improvement needed
quality of slides 20%	Outstanding quality of slides with appropriate use of tables, plots, and illustrations	Good effort and commendable quality of slides	Adequate effort with room for improvement	Limited effort with considerable improvement needed	Minimal effort and significant improvement needed
presentation professionalism 10%	Outstanding effort and exceptional quality of delivering the talk.	Good effort and commendable quality of delivering the talk.	Adequate effort with room for improvement	Limited effort with considerable improvement needed	Minimal effort and significant improvement needed

Research paper presentation (only grad students)

The students will be asked to make presentations on methods or research papers. Expect full grade for:

- Making good, professionally sound 20-25 min presentation;
- Successfully connecting the presentation to the topics discussed in class and to other peer-reviewed literature; answering the questions in a clear, professional manner.

CLASS POLICIES

If you are not able to make it to the class

You should always contact me through Canvas if you are going to miss a class or are unable to return an assignment on time.

Late assignment submission or skipping a quiz

Closely follow the course logistics concerning the submission of your work. All assignments are due before the beginning of the next class, except for in-class quizzes. Late submissions are penalized: up to 48 hours later is -20%. No make-up assignments or quizzes will be allowed except when required by the University Policies. An example of an allowed missed assignment is a student athlete's game travel, as requested by their trainer's email. Please get in touch with the Dean of Students' office to confirm a medical reason or family emergency preventing your class attendance. I will follow their guidance. Please note that I cannot accept a doctor's notice due to privacy concerns. Requirements for class attendance and make-up exams, assignments, and other work in this course are consistent with university policies that can be found at: https://catalog.ufl.edu/ugrad/current/regulations/info/attendance.aspx.

If you cannot deliver the final presentation due to a confirmed medical reason or family emergency, your presentation will be rescheduled for a later date if possible; otherwise, a 0% or "incomplete" grade for the class will be assigned. If you have not participated in the group project, a 0% or "incomplete" grade for the class will be assigned.

Note that minor sickness or a short travel will not be considered an excuse for not returning the homework. The point deduction is because you will always be given enough time to complete and return an assignment a few days before the due date; please plan ahead for possible emergencies.

Miscellanea

Water in bottles and spill-proof cups is allowed; food is not allowed. Remember: soft drink spills kill computer equipment!

Please turn off the sound on your phones and refrain from using the Internet, playing games, reading books, or engaging in any other activity unless it is directly related to the course.

CAMPUS RESOURCES and UF ACADEMIC POLICIES

https://syllabus.ufl.edu/syllabus-policy/uf-syllabus-policy-links/

Appendix A. Term project

Introduction

During the course, you will be doing a group project on a topic of your interest. Imagine that you are a group of scientists collaborating on a project. Your goal is to analyze the literature in your field of expertise, formulate a sound research proposal, collect the data, perform statistical data analysis, write a project report, and make a research presentation.

Report structure

- Abstract
- Introduction (Statement of the problems and Literature review)
- Data collection
- Data analysis
- Discussion
- References

Report writing

Writing responsibilities can be distributed among the students as they see fit. I suggest that one of the students become a project leader, responsible for project integrity. All parts have to be completed, and the text should flow seamlessly between them.

Final presentation

The students will individually present the project; that is, if there are three students in a group, there should be one presentation with the students taking turns. Make sure that your individual talks make one integrated presentation. For example, the project leader may introduce the project and tell why it is interesting/important, the next student will talk about data collection, and the last one will talk about data analysis. When four students work on one report, the fourth student may, e.g., discuss the study's implications.

Project discussions

Be prepared to discuss the project report in class, but also plan to meet outside the class: class meetings are to exchange ideas and outcomes with a larger audience.

Appendix B. Course Schedule (subject to change)

Reading: AK – Kirilenko; Additional reading: N – North; W – Witten; K – Kotu; T - Tan

Week	Lecture	Reading	Assignment due
1	Introduction. Syllabus. Course structure. Project. Reports. Introduction to Data Mining and the CRISP-DM Process	AK Chapter 1 Provost, Fawcett.pdf; W 1	-
2	Data Pre-processing and Scrubbing - Data complexity. - Data type and level of measurement. Metadata. - Sources of data errors and inconsistency. - Methods of data quality analysis. - Data cleaning: handling missing data, noise, and inconsistent data. - Data reduction. - Data scaling.	AK Chapter 2 K 3	Lab 1
3	Introduction to Data Analytics. Association Rules - Overview of data analytics methods. - Informal definition of association rules. - Examples of association rules applications: market basket analysis and recommender systems. - Formal definition of association rules. Itemset, support count, support, frequent itemset. - Measures of interestingness: Support, confidence, Laplace correction, and Piatetsky-Shapiro. - Measures of interestingness: Lift. - Measures of interestingness: Conviction and Gain. - Approaches to mine association rules. - A priory principle works for finding frequent itemsets. - Data considerations	AK Chapter 3 W 4.5, 6.3; K 6.	Lab 2
4	Decision Trees in Data Analytics - Introduction to decision trees. - Methods of decision tree node splitting. - Handling categorical and continuous data. - Pruning the trees. - From trees to random forest.	AK Chapter 4 T 4.3 – 4.4	Lab 3; Quiz
5	Clustering. K-means. DBSCAN. Hierarchical clustering - Basics of cluster analysis. - Partitional clustering: K-means. - Data considerations and limitations of K-means. Overcoming limitations. - Density-based partitioning: DBSCAN. - Introduction to hierarchical clustering. - Agglomerative and divisive approaches. - Distance measures. - Applications for text and image analysis.	AK Chapter 5, 6 K 7 – 7.3; T 7	Lab 4
6	Predictive analytics. Supervised learning. Naïve Bayes. - Supervised vs. unsupervised learning.	AK Chapter 7 W 4.2, 6.7; K. 4.4	Lab 5; Quiz

	- Introduction to conditional probability. Breast cancer testing example. - Bayesian approach. - Naïve Bayes classifier. - Dealing with continuous variables. - Back to K-Means: supervised K-Means classifier. - Overview of other supervised classification methods.		
7	Holiday		
8	Method presentations by graduate students		Lab 6
9	Validation and evaluation methods - Understanding the difference between validation and evaluation. - Model performance evaluation metrics: accuracy, precision, recall, F-measure. - Accounting for by-chance agreement: Kohen's kappa. - Holdout methods of performance evaluation. Cross-validation. - Working with small samples: Bootstrapping. - Cumulative evaluation methods of model performance: Gain and lift charts. - Response Operator Curve and Area Under Curve: ROC and AUC.	AK Chapter 8 W 5.1 – 5.6; K 8	
10	Introduction to Web Data Scraping - Overview of web data scraping. - Introduction to API. - Understanding API documentation. - Parsing API response: JSON and XML. - Tools and libraries: Postman, Request. - Rate Limiting and Pagination: handle common API limits. Exercise: Write a Python script to scrape headlines from the Reddit travel subreddit. Data: subreddit r/Travel.	AK Chapter 9 Thelwall et al., 2010	Lab 7
11	Introduction to Text Mining - Overview of text mining goals - Text cleaning and visualization methods - Examples from marketing and business analytics.	AK Chapter 11 K 9.2; Dean, "Text mining"	Lab 8; Quiz
12	Introduction to Sentiment Analysis - Introduction to sentiment analysis. - Three types of sentiment analysis. Cumulative, feature-based, and comparative sentiment mining. - Emotion analysis: Plutchik's wheel of emotions. - Lexicon-based sentiment and emotion analysis. - Machine learning sentiment analysis. - Off-the-shelf solutions.	AK Chapter 10 Liu, "A survey of opinion mining and sentiment analysis"	Lab 9
13	Topic Modeling - From text mining to topic modeling. - Main assumptions of topic modeling. - Simple approaches to topic inference: word frequency, word cloud, change-tracking. - Clustering and classification in topic modeling.	AK Chapter 12, 13	Lab 10

	- Topic inference based on latent variables. LDA. - OpenAI's ChatGPT: Topic modeling with ChatGPT and GPT-4 API. - Google's BERT: Topic modeling with BERTopic.		
14	Project consultations		Lab 11; Quiz
15	**ORAL PROJECT REPORT**		
16	**Written report due**	**Finals week**	**Written report due date**