HMG6583c/LEI4905 Data Mining with Social Data

Department of Tourism, Recreation & Sport 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: Friday 1 – 3 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 to 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 network data from online sources
- Be able to develop your own tools for data acquisition and warehousing

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

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 would need to follow the labs' reservation schedule to complete your assignments. Note that I will provide instructions and help for Windows PC; Mac *should be* ok: the software we will be using works on either platform.

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

- Witten, Frank. Data Mining. Practical Machine Learning Tools and Techniques. A hard copy from Amazon.com is approx. $25.

- 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
- Tan, Introduction to Data Mining (Chapters 3, 7). Free download from http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Optional reading for deeper understanding

- Foster Provost, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.
- Jennifer Golbeck. Analyzing the Social Web.
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. Free book is available online: http://www.nltk.org/book/
- Reza Zafarani, Mohammad Ali Abbasi, Huan Liu. Social Media Mining. An Introduction. Free book is available online: http://dmml.asu.edu/smm/SMM.pdf
- 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
- Al Sweigart, Automate the Boring Stuff with Python - Practical Programming for Total Beginners. Get for free at https://automatetheboringstuff.com/

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. Student presentations (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. No issues regarding scoring will be reviewed beyond this period or after midnight of the last day of the Examination week, whichever comes first. **For UF grading policies see https://catalog.ufl.edu/ugrad/current/regulations/info/grades.aspx**

Quizzes

An occasional short quiz will usually cover the material from the previous theme but expect occasional questions related to the earlier topics. The quizzes will be closed book. The exams will have the same format (with a few more problems to solve) and may cover any topic in the course. **For full credit make sure the instructor is able to read through your handwriting**. A 100% grade will require full answers to all questions, a returned blank paper will be evaluated as 0%, and reasonable progress toward answering the questions will be evaluated somewhere in between.

Assignments

The lab work has to be submitted through the course management system as an assignment. The acceptable submission format is a Word or PDF 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 or a returned blank paper will be evaluated as 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, that is, literature review, research design, data collection, data analysis, and research presentation. Project results should be presented in the form of a research report (due **before** the date and time of the final exam) AND an oral presentation.

Expect a 100% grade for using multiple sources of information for preparation of your report, professional data analysis, in-detail presentation of the topic, intelligent answers to the questions, and active engagement in the discussion of the projects during the project meetings. See Appendix A for clarifications. For participation in

project discussion, expect full grades for asking questions, submitting answers, sharing your opinions, and similar class-time participation.

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

Always contact me through Canvas if going to miss a class or are unable to return an assignment in time.

Late assignment submission or skipping a quiz

Closely follow the course logistics concerning the submission of your work. All assignments (quizzes, problems from the textbook, and SPSS labs) are due before the beginning of the next class. Late submissions are penalized: Up to 48 hours later -20%. No make-up assignments or quizzes will be allowed except as required by the University Policies. An example of an allowed missed assignment is a student athlete's game travel, as requested by his/her trainer's email. **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**.

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

Presentations

If you are unable to deliver a presentation due to a confirmed medical reason or family emergency, it will be rescheduled for a later date if possible; otherwise, a 0% credit or an "incomplete" grade will be assigned.

Food

Water in bottles and spill-proof cups is allowed by the class policies, but may be prohibited in a specific room; food is not allowed. Remember: soft drink spills kill computer equipment.

Special accommodations

Students requesting special classroom accommodations must first register with the Dean of Students Office. Pease let the instructor know your needs ASAP.

Miscellanea

1. Please switch off the sound on your phones and refrain from using the Internet, playing games, reading books, and other activities unless it is directly related to the course.
2. Unless an urgent business requires my attention, I will be available for questions **after** the lecture hours. For more complex questions that require substantial time please secure an appointment by sending in an email.
3. Students are expected to provide feedback on the quality of instruction in this course by completing **online evaluations at https://evaluations.ufl.edu**. Evaluations are typically open during the last two or three weeks of the semester, but students will be given specific times when they are open. Summary results of these assessments are available to students at https://evaluations.ufl.edu/results/."

FINE PRINT

Group work and academic honesty

The plagiarism and other violations of the academic honesty will be punished with 0% grade for the assignment; the offender will be reported to the head of department and/or graduate school for possible actions. The UF defines plagiarism in the following way (https://www.dso.ufl.edu/sccr/process/student-conduct-honor-code):

"(a) Plagiarism. A student shall not represent as the student's own work all or any portion of the work of another. Plagiarism includes but is not limited to:

1. Quoting oral or written materials including but not limited to those found on the internet, whether published or unpublished, without proper attribution.

2. Submitting a document or assignment which in whole or in part is identical or substantially identical to a document or assignment not authored by the student."

Further, each student is expected to abide by the Honor Code: "We, the members of the University of Florida community, pledge to hold ourselves and our peers to the highest standards of honesty and integrity" (https://www.dso.ufl.edu/sccr/process/student-conduct-honor-code/). Furthermore, you are obligated to report any condition that facilitates academic misconduct to appropriate personnel. Please refer to the abovementioned Honor Code for a complete explanation of the University of Florida Academic Honesty Policy.

CAMPUS RESOURCES

Health and Wellness

U Matter, We Care: If you or a friend is in distress, please contact umatter@ufl.edu or 352 392-1575 so that a team member can reach out to the student.

Counseling and Wellness Center: http://www.counseling.ufl.edu/cwc/Default.aspx, 392-1575.

University Police Department, 392-1111 (or 9-1-1 for emergencies). http://www.police.ufl.edu/

Sexual Assault Recovery Services (SARS): Student Health Care Center, 392-1161.

Disability resource center: https://drc.dso.ufl.edu, 392-8565, accessUF@ufsa.ufl.edu.

Academic Resources

E-learning technical support, 352-392-4357 (select option 2) or e-mail to Learning-support@ufl.edu. https://lss.at.ufl.edu/help.shtml.

Career Resource Center, Reitz Union, 392-1601. Career assistance and counseling. http://www.crc.ufl.edu/

Library Support, http://cms.uflib.ufl.edu/ask. Various ways to receive assistance with respect to using the libraries or finding resources.

Teaching Center, Broward Hall, 392-2010 or 392-6420. General study skills and tutoring. http://teachingcenter.ufl.edu/

Writing Studio, 302 Tigert Hall, 846-1138. Help brainstorming, formatting, and writing papers. http://writing.ufl.edu/writing-studio/

Student Complaints Campus: https://www.dso.ufl.edu/documents/UF_Complaints_policy.pdf; On-Line Students Complaints: http://www.distance.ufl.edu/student-complaint-process

Student Complaints

Campus students: https://www.dso.ufl.edu/documents/UF_Complaints_policy.pdf;
On-Line students: http://www.distance.ufl.edu/student-complaint-process

UF Covid-19 class expectations

In response to COVID-19, the following practices are in place to maintain your learning environment, to enhance the safety of our in-classroom interactions, and to further the health and safety of ourselves, our neighbors, and our loved ones.

- If you are not vaccinated, get vaccinated. Vaccines are readily available at no cost and have been demonstrated to be safe and effective against the COVID-19 virus. Visit this link for details on where to get your shot, including options that do not require an appointment: https://coronavirus.ufhealth.org/vaccinations/vaccine-availability/. Students who receive the first dose of the vaccine somewhere off-campus and/or outside of Gainesville can still receive their second dose on campus.
- You are expected to wear approved face coverings at all times during class and within buildings even if you are vaccinated. Please continue to follow healthy habits, including best practices like frequent hand washing. Following these practices is our responsibility as Gators.
    - Sanitizing supplies are available in the classroom if you wish to wipe down your desks prior to sitting down and at the end of the class.
    - Hand sanitizing stations will be located in every classroom.
- If you sick, stay home and self-quarantine. Please visit the UF Health Screen, Test & Protect website about next steps, retake the questionnaire and schedule your test for no sooner than 24 hours after your symptoms began. Please call your primary care provider if you are ill and need immediate care or the UF Student Health Care Center at 352-392-1161 (or email covid@shcc.ufl.edu) to be evaluated for testing and to receive further instructions about returning to campus. UF Health Screen, Test & Protect offers guidance when you are sick, have been exposed to someone who has tested positive or have tested positive yourself. Visit the UF Health Screen, Test & Protect website for more information.
    - Course materials will be provided to you with an excused absence, and you will be given a reasonable amount of time to make up work.
    - If you are withheld from campus by the Department of Health through Screen, Test & Protect you are not permitted to use any on campus facilities. Students attempting to attend campus activities when withheld from campus will be referred to the Dean of Students Office.
- Continue to regularly visit coronavirus.UFHealth.org and coronavirus.ufl.edu for up-to-date information about COVID-19 and vaccination.

MISC.

Students are expected to provide professional and respectful feedback on the quality of instruction in this course by completing course evaluations online via GatorEvals. Guidance on how to give feedback in a professional and respectful manner is available at https://gatorevals.aa.ufl.edu/students/. Students will be notified when the evaluation period opens, and can complete evaluations through the email they receive from GatorEvals, in their Canvas course menu under GatorEvals, or via https://ufl.bluera.com/ufl/. Summaries of course evaluation results are available to students at https://gatorevals.aa.ufl.edu/public-results/.

Students with disabilities requesting accommodations should first register with the Disability Resource Center (352-392-8565, www.dso.ufl.edu/drc/) by providing appropriate documentation. Once registered, students will receive an accommodation letter which must be presented to the instructor when requesting accommodation. Students with disabilities should follow this procedure as early as possible in the semester.

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

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 in a project. You goal is to analyze the literature in your field of expertize, formulate a sound research proposal, collect the data, perform statistical data analysis, write project report, and make a research presentation.

1. Report structure

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

2. Report writing

The writing responsibilities can be distributed between the students as they see fit. I suggest that one of the students becomes project leader, responsible for project integrity. All parts have to be completed; there should be seamless flow of the text between the parts.

3. 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 implications of the study.

4. Project discussions

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

5. Project grading

50% of the grade will be group-assigned based on the quality of the final report;

50% of the grade will be individually assigned based on the quality of presentation.

Final report grading (up to 100%):

90-100: Excellent. Excellent, scholarly, and advanced college-level work. Original, insightful ideas, in-depth discussion. Well organized and structured. Very good grammar, careful formatting;

80-90: Good. Good college-level work that exceeds requirements. Original, well organized. Good comprehension of the topic is demonstrated. Acceptable grammar. Some areas are noticeably weaker than others;

70-80: Satisfactory. Average work. Assignment is not thought through and/or presentation is not cohesive. Improvement is needed on depth, originality of thought, structure, and presentation;

60-70: Marginal. Below-average work. Substantial improvements are needed in the areas of content, reasoning, and delivery, as well as grammar and formatting. There is a missing section in the report;

0: Failure. The assignment is not submitted or is significantly incomplete (more than one missing section).

Oral report grading (up to 100%):

The following scoring rubric will be used for the oral presentation grading with 0-20 percentage points in each category:

Report content : Appropriate introduction, data, analysis, and outcomes sections in the report. Well defined technical terms. Good summary of the work at the end;

Data visualization: Appropriate use of tables, maps, and scientific graphics for information delivery;

Professional delivery: clear, audible voice, appropriate gestures and eye contact that engage the audience, seamless switching between presenters, no between-reports pause for missing presentation, not playing video or similar issues;

Answering questions: questions from the audience are answered in a way that shows that presenter is familiar with the subject;

Time management: presentation time limit is not exceeded and also is not significantly shorter than allocated. For a full grade, presentation time should not be over 10% shorter or 5% longer.

Appendix. Course schedule (subject to change).

Reading books: W – Witten; K – Kotu; T - Tan

#	Date	Lecture	Reading	Assignment due
1	Aug. 26	Introduction. Syllabus. Course structure. Project. Reports. Introduction to Data Mining and the CRISP-DM Process	Provost, Fawcett.pdf; W 1 Lecture notes Chapter 1	-
	Sept 2	Holiday		
2	Sept 9	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.	K 3 Lecture notes Chapter 2	Lab 1
3	Sept. 16	Introduction to Data Analytics. Association Rules - Overview of data analytics methods. - Informal definition of association rules. - Examples of association rules applications: market basket analysis, 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	W 4.5, 6.3; K 6. Lecture notes Chapter 3	Lab 2
4	Sept. 23	Decision Trees in Data Analytics - Introduction to decision trees. - Methods of decision tree node splitting. - Handling categorical and continuous data. - Pruning the trees.	T 4.3 – 4.4 Lecture notes Chapter 4	Lab 3; Quiz lectures 1-3

		- From trees to random forest.		
5	Sept 30	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.	K 7 – 7.3; T 7 Lecture notes Chapter 5, 6	Lab 4
6	Oct 7	Predictive analytics. Supervised learning. Naïve Bayes. - Supervised vs. unsupervised learning. - 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.	W 4.2, 6.7; K. 4.4; Lecture notes Chapter 7	Lab 5; Quiz lectures 4-5
	Oct 14	Student method presentations (only grad students)		Lab 6; oral presentation
7	Oct 21	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.	W 5.1 – 5.6; K 8 Lecture notes Chapter 8	
8	Oct. 28	Introduction to Web Data Scraping - Overview of web data scraping. - Introduction to API. - Understanding API documentation.	Sentiment Analysis for Social Data Thelwall et al., 2010 Lecture notes Chapter 9	Lab 7 Quiz lectures 6-7

		- 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.		
9	Nov 4	Introduction to Text Mining and Sentiment Analysis - Overview of text mining goals. - Examples from marketing and business analytics. - 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.	K 9.2; Dean, "Text mining"; Liu, "A survey of opinion mining and sentiment analysis" Lecture notes Chapter 10, 11	Lab 8
	Nov. 11	Holiday		Lab 9
10	Nov. 18	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. - 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.	Lecture notes Chapter 12, 13	Quiz
	Nov. 25	Holiday		
11	Dec 2	PROJECT REPORT		Lab 10
	Dec. 9	Written report due	Written report due	Written report due