ITAO 70250 – Unstructured Data Analytics

Spring 2018 (Module 3)

Professor	Dr. Timothy E. Carone
Class Location and Time	Mendoza L003; MW 03:30-05:20 PM
Phone (Mobile)	1-847-226-0659
Phone (Office)	1-574-631-9322
Email	Tcarone1@nd.edu
Course Webpage	Via Sakai
Office	Mendoza 327
Office Hours	Monday: By appointment or telephonically Tuesday: 7-9PM Wednesday: Office 10AM-12PM, 7-9PM Thursday: After 5PM, Telephonically Only (on travel) Friday/Saturday/Sunday: 7AM-10PM, Telephonically and Zoom/Hangout

COURSE OVERVIEW

This course introduces students to the process of performing high-valued analytics using unstructured data to support business decisions. Unstructured data (text, tweets, posts, video, audio, …) consist of over 80% of the data being produced every day. Most companies have minimal competencies in using unstructured data analytics for new product development, customer retention, workforce optimization, and a myriad of other areas. This course will get you started in how to lead efforts at your company to monetize their data and achieve your career objectives.

This course will not cover data cleaning. The data you will use will be ready to go or almost ready to go.

COURSE OBJECTIVES

By the end of the semester, I hope that you will have achieved the following objectives:

1. LEARN HOW TO SOLVE BUSINESS PROBLEMS THAT NEED ANALYTICS OF UNSTRUCTURED DATA AS PART OF THE SOLUTION.

2. UNDERSTAND THE VARIOUS KEY ALGORITHMS USED TO CREATE THE ANALYTICS.
3. UNDERSTAND THE USE OF ANALYTICS IN MANY DIFFERENT BUSINESS CONTEXTS TO OBSERVE THE SIMILARITIES AND DIFFERENCES IN HOW ANALYTICS IS USED TO SOLVE PROBLEMS.

ACADEMIC CODE OF HONOR

Everyone in the course is expected to adhere to the University's Academic Code of Honor, which can be found online at http://www.nd.edu/~hnrcode/docs/handbook.htm#1. Acts of academic dishonesty such as plagiarizing (whether from work submitted in prior terms of the class, online sources such as Wikipedia, or elsewhere) or copying off of others during exams will not be tolerated, and will be penalized as specified by university policy. The following pledge, which can be found in the Honor Code, sums it up: "As a member of the Notre Dame community, I will not participate in or tolerate academic dishonesty."

GENERAL COMMENTS

I am looking forward to teaching this class. I believe unstructured data analytics is one of the most interesting and practical classes you will take during your college careers. I hope that by the end of the semester I can convince you to share this belief.

I care about your progress over the semester and will do everything I can to help you succeed. My office hours this semester are listed above. Please do not hesitate to visit if you have any questions or concerns, or even if you just want to chat about the course. I can, of course, meet with you outside of these times. I am not in my office on Fridays. I am typically available Sunday through Saturday from 7:00AM to 10:00PM ET. The best way to get in touch with me is to send me a text (1-847-226-0659). I will respond as quickly as I can (quickly = seconds to a day if it is a weekend) with either a phone call or to suggest a time for a call. When you send a text you need to start off with "Tim – (*insert your name*) here." so I will know who you are and that you are from this class. You can also e-mail or speak to me before/after class if you would like to set up a meeting. I am also willing to help (to the best of my ability) with career issues, job interview preparation, and graduate school questions.

I regularly post materials on Sakai during the semester, including the syllabus, lecture slides, readings, and other information. Please check our course website for these materials.

There is no need to let me know that you will be absent from class, but do remember that to earn an A you need to be present. If you miss class, please get class notes and any other relevant materials from a team member.

MATERIALS

Required

READINGS - Articles that provide real world examples of the topics we will be covering in class are in Sakai under RESOURCES. You need to read them prior to class as they will be discussed in class and your ability to contribute to the class discussions will figure into your Class Participation grade.

TEXTBOOKS:

1. ***TEXT MINING WITH R: A TIDY APPROACH***, 1ST EDITION, JULIA SILGE AND DAVID ROBINSON, ISBN-13: 978-1491981658 (HTTPS://WWW.AMAZON.COM/TEXT-MINING-R-TIDY-APPROACH/DP/1491981652/REF=SR_1_1?S=BOOKS&IE=UTF8&QID=1508424166&SR=1-1&KEYWORDS=TEXT+MINING+WITH+R)
2. **R COOKBOOK: PROVEN RECIPES FOR DATA ANALYSIS, STATISTICS, AND GRAPHICS**, 1ST EDITION, PAUL TEETOR, ISBN-13: 978-0596809157 (HTTPS://WWW.AMAZON.COM/COOKBOOK-ANALYSIS-STATISTICS-GRAPHICS-COOKBOOKS/DP/0596809158/REF=CM_CR_ARP_D_PRODUCT_TOP?IE=UTF8)

COURSE GRADING

The components of your grade are as follows:

COMPONENT	WEIGHT
Class Participation	20%
Homework	50%
Final Exam	30%

1. Class Participation

This portion of the grade will reflect the extent to which you offer meaningful insights on the day's topic including questions on the case studies. Specifically, I evaluate class contribution based on the extent to which your contributions consistently increase the average class understanding of the discussion at hand rather than how frequently you participate. Rich discussions makes class more enjoyable for everyone, and helps you and your fellow classmates learn more effectively than simply listening to me lecture. Below is a non-exhaustive list of things that constitutes good contribution:

1. MAKING EVIDENCE-BASED COMMENTS AND RECOMMENDATIONS
2. DEMONSTRATING AN UNDERSTANDING OF THE READINGS
3. BUILDING ON THE COMMENTS OF OTHERS
4. BEING A GOOD LISTENER AND RESPECTING YOUR PEERS' OPINIONS
5. ASKING THOUGHTFUL QUESTIONS
6. A HIGH CONTRIBUTION-TO-WORDS RATIO (I.E., AVOID FILLING "AIR TIME")

I use a variety of methods to determine your class contribution grades, including your comments you discuss with me or questions. Specifically, after every class I will give each of you a contribution score. A class score is based on the following scale:

CONTRIBUTE ATTRIBUTE	SCALE
Absent without excuse or not prepared to discuss the case study or other reading assigned	0
Disrupting attention (e.g., excessive side discussions with your neighbors)	2-4
Partial attention (e.g., arriving late, sleeping in-class) and nameplate visible	4-6
Full attention and nameplate visible	6-7
Full attention with contribution demonstrating comprehension and nameplate visible	8-10

Consequently, simply showing up on time, displaying your nameplate, and paying attention in class gets you a 7 out of 10. To move up from there, you need to get involved and bring something to the day's discussion. I am more than happy to talk with you about your class contribution at any point during the semester and am happy to provide feedback about how you are doing in regard to this component of your grade.

Please find a seat you like by the start of the second class (Wednesday, August 23). This will be your seat for the semester. **Always display your nameplate**.

2. Homework

Your homework will vary between questions that cover current affairs in analytics)and how they are used and misused today) as well as programming problems.

Three of your homework grades will be the successful completion of the three Datacamp courses.

3. Final Exam

The final exam will have two parts. There will be a take home programming problem and an in-class exam.

The take home exam will require you to solve a problem and implement the solution in R. I am not concerned about your level of sophistication with R. I care about your solution and implementing the solution in R.

The in-class final exam will contain a combination of problem exercises, multiple choice questions, and, perhaps, short essays that are designed to test your knowledge of concepts and terminology covered in class and in the readings. Similar to the mid-term exam, students should bring a pencil/pen to the exam as well as a calculator (this is the only electronic device that may be used during the exam). **The final exam has been scheduled from 3:30-5:30PM on Thursday March 1st in Mendoza L061.**

Barring a documented emergency, missing either exam will result in a grade of zero. Please contact me immediately (or have a relative or friend contact me) in the event of a medical or family emergency.

COURSE MATERIALS

1. LECTURE 1 HANDOUT – THIS DOCUMENT COVERS MATERIAL THAT IS NOT COVERED IN ANALYTICS TEXTS. IT DISCUSSES THE PROCESS TO SOLVE PROBLEMS IN ANALYTICS AND ISSUES THAT OCCUR SUCH AS BIAS.
2. R SOLUTIONS – WE WILL COVER MANY TOPICS AND YOU WILL BE PROVIDED WITH R CODE THAT IS HEAVILY COMMENTED AS WELL AS THE DATA BEING USED.
3. READINGS – EACH CLASS WILL START WITH AN ASSIGNED READING. MUCH OF YOUR CLASS PARTICIPATION GRADE WILL DEPEND ON YOUR KNOWLEDGE OF THE ARTICLE AND CONTRIBUTING ADDITIONAL MATERIALS

ELECTRONIC DEVICE POLICY

Electronic devices are allowed in the class to the extent they are used for class purposes as I would like to reduce the use of paper. Class purposes can be referencing the syllabus, accessing Sakai or taking notes. Please keep in mind that use of a device might be offensive to classmates in your vicinity so keep that in mind. If you use a device during class for non-class purposes it will always significantly negatively impact your class contribution grade.

Phones should stay in your pocket or bag for the duration of class and set to vibrate. If you have an emergency where your phone needs to be out during class, please let me know before class begins. If you use a phone during class it will always significantly negatively impact your class contribution grade.

ADDITIONAL NOTES

I believe everyone has the right to take the class without undue hardship deriving from conditions such as physical or learning disabilities. If you have any such condition, please notify me in the first week of class and I will strive to make the appropriate accommodations.

This syllabus is subject to change. For example, I may find a better case study for you to use rather than the one listed above. Any change will be communicated to you as soon as it occurs.

Unstructured Data Analytics – Carone – Spring 2018 (Section 1)

CLASS	DATE	TOPIC	READINGS TO DO BEFORE CLASS	HOMEWORK PROBLEMS
1	1/15 Mon	Introduction to Unstructured Data Analytics	Lecture 1 Handout	
2	1/17 Wed	Text Analysis	Chap. 3 – Analyzing Word and Document Frequency Reading - Real News on Fake Data in China	
3	1/22 Mon	Text Analysis	Chap. 3 – Analyzing Word and Document Frequency Reading - Why 2016 election polls missed their mark	Datacamp: Introduction to R
4	1/24 Wed	Text Analysis	Chap. 4 – Relationships Between Words: N-Grams and Correlations Reading - US considers mortgage credit check shake-up	
5	1/29 Mon	Supervised Learning	Algorithms: K Nearest Neighbor, Logistic Regression, Neural Networks, Support Vector Machine (SVM), Naïve-Bayes Reading - We snoop to conquer - Retail technology	Datacamp: Intermediate R
6	1/31 Wed	Case Study	Predicting March Madness Reading - New big data trend tracks 'digital footprints'	Datacamp: Data Visualization with ggplot2 (Part 1)
7	2/5 Mon	Case Study	Classification of Complex Legal Documents Reading - Law firms try self-analysis, 'Moneyball' style	Musical Instrument Identification Using Neural Networks
8	2/7 Wed	Unsupervised Learning	Algorithms: K-Means, EM, DBSCAN Reading - The Power of Learning - Data Analytics	
9	2/12 Mon	Case Study	Identify Undervalued Baseball Players Reading - Push my buttons - Retailing	Real Estate Price Prediction with Regression and Classification
10	2/14 Wed	Case Study	Modeling Political Identity Reading - The Great Chain of Being Sure about Things – Blockchains Reading - The trust machine - The promise of the blockchain	
11	2/19 Mon	Case Study	Predicting Gross Box Office Reading - Battle of the brains - Artificial intelligence	Identifying Authors of the Federalist Papers
12	2/21 Wed	Advanced Topics – Geospatial Data	Finding Poverty in Satellite Images Reading - Living with technology - The data republic	
13	2/26 Mon	Final Exam Prep	Covers Classes 1-12	
14	3/1 Thurs	Final Exam	Covers Classes 1-12	
OFFICE HOURS (847) 226-0659		Monday	7AM-3PM: Telephonically Only	
		Tuesday	Office: 7-9PM	
		Wednesday	Office: 10AM-12PM, 7-9PM	
		Thursday	After 5PM: Telephonically Only	
		Friday/Saturday/Sunday	7AM-10PM: Telephonically and Zoom/Hangout	