BMIS 625
Text Mining of Unstructured Data

UNIVERSITY OF MONTANA
COLLEGE OF BUSINESS
MASTER OF BUSINESS ANALYTICS

Course Description

This course applies the skills and technology of data science to text mining and explores unstructured data generally. Most data stored and created today is *unstructured*. Unstructured data is data that is not stored in a regularized way such as a relational database or a series of key-value pairs. Common examples of unstructured data include text, images and video. In this course we'll focus principally on text data.

There are several fields of textual analysis in which one can get a Ph.D (e.g., natural language processing, linguistics, computational linguistics). The central methods to text analysis, however, are accessible very quickly. In this class we will begin assuming no previous experience working with text analysis and finish with skills that could be deployed in the workplace. We will not delve deeply enough to become a professional in the field, but you'll see what it would take by the end of the course.

This course will be taught mostly in Python, as that is the standard for text analysis in the industry. We will also take a couple of weeks working through the R packages that handle text mining, as these are powerful, intuitive, and pretty fun to use.

Course Objectives

1. Students will be familiar with the broad themes and problems of text mining.
2. Students will be introduced to the theory behind language processing.
3. Students will be familiar with key functionality from Python's Natural Language Tool Kit (NLTK).
4. Students will practice techniques such as sentiment analysis, tokenization, term-document frequency analysis, and Naïve Bayes classification.
5. Students will be familiar with approaches to topic discovery.

Assessment

- ### Class Participation: 25%

  Show up ready to work hard. Ask questions when you're confused. Contribute to discussions either in-person or online.

- ### Individual Project: 75%

  Students will complete an individual assignment working with text data. This data will be assembled as part of Applied Data Analytics (BMKT 670). If you are taking Text Mining *before* Applied Data Analytics, let me know and I'll help you find a dataset. Ideally, this work can be progress against the MSBA capstone.

Course Outline

The following is a rough outline of the topics to be covered, by week.

	Language(s)	Topics	Due Dates
Week 1	Python, SQL	Introduction, Working with Text Data	
Week 2	Python	Character Encodings, Tokenization	
Week 3	Python	Parsing, Stemming	
Week 4	Python	APIs, Web Scraping	
Week 5	Python	Regular Expressions	
Week 6	Python	Spelling Correction	
Week 7	Python	Probabilistic models for text, Naïve Bayes	
Week 8	Python	Clustering, Similarity Measures	
Week 9	Python	Sentiment Analysis	
Week 10	Python	Information Retrieval and the tf-idf	
Week 11	Python	Information Retrieval and the tf-idf	
Week 12	R	Text Mining with R	
Week 13	R	Text Mining with R	
Week 14	Thanksgiving		
Week 15	-	Project Presentations	

Double Dipping

A note on double dipping, which we define as submitting an assignment from one course in a second course. Here's what a recent syllabus for BMKT 680 says on the topic:

> Please note that it is a form of academic misconduct to submit work that was also used in another course, aka *"double dipping."* ***Don't do it.*** If you are trying to get synergies across your classes/assignments, just ask a professor for advice. Don't try for a two-fer without approval!

I'm generally okay with double dipping if you get my approval, but I include the above quote to highlight that my stance is anomalous. If you're interested in using a project in my class for another class, let's talk about it and decide how you'll differentiate the two bodies of work. We *expect* you to use work from ADA in your capstone and don't consider that double dipping. Good talk.

Code of Conduct

We are dedicated to providing a welcoming and supportive environment for all people, regardless of background or identity. We recognize that some groups in our community, however, are subject to historical and ongoing discrimination, and may be vulnerable or disadvantaged. Membership in such a specific group can be on the basis of characteristics such as such as gender, sexual orientation, disability, physical appearance, body size, race, nationality, sex, color, ethnic or social origin, pregnancy, citizenship, familial status, veteran status, genetic information, religion or belief, political or any other opinion, membership of a national minority, property, birth, age, or choice of text editor. We do not tolerate harassment of participants on the basis of these categories, or for any other reason.

Harassment is any form of behavior intended to exclude, intimidate, or cause discomfort. Because we are a diverse community, we may have different ways of communicating and of understanding the intent behind actions. Therefore, we have chosen to prohibit certain norms of behavior in our community, regardless of intent. Prohibited harassing behavior includes but is not limited to:

- written or verbal comments which have the effect of excluding people on the basis of membership of a specific group listed above;
- causing someone to fear for their safety, such as through stalking, following, or intimidation;
- the display of sexual or violent images;
- unwelcome sexual attention;
- non-consensual or unwelcome physical contact;
- sustained disruption of talks, events or communications;
- incitement to violence, suicide, or self-harm;
- continuing to initiate interaction (including photography or recording) with someone after being asked to stop; and
- publication of private communication without consent.

Behavior not explicitly mentioned above may still constitute harassment. The list above should not be taken as exhaustive but rather as a guide to make it easier to enrich all of us and the communities in which we participate. All interactions should be professional regardless of location: harassment is prohibited whether it occurs on or offline, and the same standards apply to both.

Enforcement of the Code of Conduct will be respectful and not include any harassing behaviors. Thank you for helping make this a welcoming, friendly community for all.

This code of conduct is a modified version of that used by PyCon, which in turn is forked from a template written by the Ada Initiative and hosted on the Geek Feminism Wiki. This specific code of conduct can be found here: Greg Wilson (ed.): How to Teach Programming (And Other Things). Second edition, Lulu.com, 2017, 978-1-365-98428-0, http://thirdbit.com/teaching.

Names and Pronouns

Many people might go by a name in daily life that is different from their legal name. In this classroom, we seek to refer to people by the names that they go by. Pronouns can be a way to affirm someone's gender identity, but they can also be unrelated to a person's identity. They are simply a public way in which people are referred to in place of their name (e.g. "he" or "she" or "they" or "ze" or something else). In this classroom, you are invited (if you want to) to share what pronouns you go by, and we seek to refer to people using the pronouns that they share. The pronouns someone indicates are not necessarily indicative of their gender identity. This statement was found at trans.umd.edu and you can visit that site to learn more.

Additional "fine print"

Professional Business Conduct in Class:

You are preparing to enter the business world as professionals and to prepare for a business career, so I expect each of you to behave in a professional manner in class.

- Arrive on time and stay for the entire class (unless excused by me).
- Behave with honesty and integrity. Don't let your team down!
- Respect everyone in class and listen openly to their ideas.
- Come to class prepared for discussion.
- Refrain from engaging in behavior that disrupts the class- this means no cell phones!

If at any time you are displaying disrespectful behavior, you may be asked to leave.

Academic Integrity: Academic misconduct is any activity that may compromise the academic integrity of the University of Montana. Academic misconduct includes, but is not limited to, deceptive acts such as cheating and plagiarism. Please note that it is a form of academic misconduct to submit work that was previously used in another course.

"Plagiarism is the representing of another's work as one's own. It is a particularly intolerable offense in the academic community and is strictly forbidden. Students who plagiarize may fail the course and be remanded to the Academic Court for possible suspension or expulsion."

"Students must always be very careful to acknowledge any kind of borrowing that is included in their work. This means not only borrowed words *but also ideas.* Acknowledgement of whatever is not one's own original work is the proper and honest use of sources. Failure to acknowledge whatever is not one's own work is plagiarism." So, ALWAYS err on the side of caution by citing the resources used in preparing your work. Moreover, always use direct quotations for exact wording taken from another source.

All students must practice academic honesty. Academic misconduct is subject to an academic penalty by the course instructor and/or a disciplinary sanction by the University. All students need to be familiar with the Student Conduct Code. The Code is available for review online at http://life.umt.edu/vpsa/student_conduct.php. It is the student's responsibility to be familiar the Student Conduct Code.

Basic Needs Security Any student who faces challenges securing food or housing, and believes that this could affect their performance in this course, is urged to contact any or all of the following campuses resources:

1. **Food Pantry Program**: UM offers a food pantry that students can access for emergency food. The pantry is open on Tuesdays from 9 to 2, on Fridays from 10-5. The pantry is located in UC 119 (in the former ASUM Childcare offices). Pantry staff operate several satellite food cupboards on campus (including one at Missoula College). For more information about this program, email umpantry@mso.umt.edu, visit the pantry's website (https://www.umt.edu/uc/food-pantry/default.php) or contact the pantry on social media (@pantryUm on twitter, @UMPantry on Facebook, um_pantry on Instagram).
2. **ASUM Renter Center**: The Renter Center has compiled a list of resources for UM students at risk of homelessness or food insecurity here: http://www.umt.edu/asum/agencies/renter-center/default.php and here: https://medium.com/griz-renter-blog. Students can schedule an appointment with Renter Center staff to discuss their situation and receive information, support, and referrals.
3. **TRiO Student Support Services**: TRiO serves UM students who are low-income, first-generation college students, or have documented disabilities. TRiO services include a textbook loan program, scholarships and financial aid help, academic advising, coaching, and tutoring. Students can check their eligibility for TRiO services online here: http://www.umt.edu/triosss/apply.php#Eligibility.

Please contact me any time for help if you are comfortable doing so. I will do my best to help connect you with additional resources.

Disability Accommodations: Students with disabilities will receive reasonable accommodations in this course. To request course modifications, please contact me within the first two weeks of class. I will work with you and Disability Services in the accommodation process. For more information, visit the Disability Services website at http://www.umt.edu/dss/ or call 406.243.2243 (Voice/Text).

COLLEGE OF BUSINESS MISSION STATEMENT

The College of Business at the University of Montana creates transformative, integrated, and student-centric learning experiences, propelling our students to make immediate and sustained impact on business and society. We nurture our students' innate work ethic to develop confident

problem solvers and ethical decision makers. We pursue thought leadership and collectively create opportunities for a better life for our students, faculty, and staff.

Email: According to University policy, faculty may only communicate with students regarding academic issues via official UM email accounts. Accordingly, students must use their GrizMail accounts (fname.lname@umontana.edu). Email from non-UM accounts will likely be flagged as spam and deleted without further response. To avoid violating the Family Educational Rights and Privacy Act, confidential information (including grades and course performance) will not be discussed via phone or email.

COLLEGE OF BUSINESS- ASSESSMENT AND ASSURANCE OF LEARNING

As part of our assessment process and assurance-of-learning standards, the Masters of Business Analytics has adopted the following Learning goals:

1. Knowledge: A deep understanding of a wide range of analytical techniques and programming tools for both structured and unstructured (e.g., text, sentiment) data

2. Application: Ability to apply appropriate analytical techniques to solve a wide variety of business/organizational problems

3. Communication/Story Telling:

- Ability to effectively communicate data analytics results and translate into effective business decision making inputs.
- Ability to effectively use data visualization techniques to illustrate results and implications.
- Ability to write an impactful narrative summarizing key insights and implications from analysis.

4. Ethics/Data Stewardship: Ability to act effectively as data stewards, applying governance techniques to secure data, to develop and promote policies for using data in an ethical manner, to respect data privacy considerations, and to enforce data compliance

5. Innovation: Ability not only to use data analytics to answer existing questions and solve known problems, but also to harness data analytics to identify new sources of value; to see patterns and anomalies; to reveal innovative insights.

Upon successful completion of this course, a student will be able to:

- Understand what makes unstructured data analysis ubiquitous, difficult, and important.
- Manage text data programmatically via Python, specifically by using the Natural Language Tool Kit (NLTK) package.
- Understand how to use tokenization and stemming to extract basic information from text data.
- Understand the basic syntax of regular expressions and use them in straightforward applications to quickly search text data.
- Use one of the key classification techniques, Naïve Bayes.
- Understand the foundational methods behind spell checkers.
- Use sentiment analysis, unlocking a fundamental technique of feature engineering (of which sentiment analysis is only the most common example).
- Learn the features of the NLTK book corpus, a data set that has utility across a variety of applications.
- Understand the theory behind information retrieval, particularly the term-frequency/inverse-document-frequency (tf-idf) statistic.
- Use the Twitter API to gather unstructured data for analysis.