Fall 9-1-2021

BMIS 625.V60: Text Mining of Unstructured Data

John W. Chandler
University of Montana, Missoula, john.chandler@umontana.edu

BMIS 625
Text Mining of Unstructured Data

John Chandler, PhD
e: john.chandler@business.umt.edu
p: 406-544-8720
CRN: 73080
Class meeting: Online (zoom link) and In Person Wednesday 4-6. Online Lab Thursday 4-6 (zoom link).
Office Hours: Tuesday 3-4 and Wednesday 9-10 at this link.

Welcome to Text Mining!

In this course, you'll learn about working with text data. Text data represents a huge and largely untapped resource for organizations of all sorts. Most analysts are comfortable only with numeric data. By the end of this course, you'll be able to do many of the analyses that will help you in future jobs:

1. **Processing text data**: As with any data set, preparing the information for analysis is a big job. Text data is no exception.
2. **Acquiring Data**: Application Programming Interfaces (APIs) and web scraping are two tools that unlock vast troves of data.
3. **Recognizing Patterns in Text**: With numeric data we practice exploratory data analysis; with text data we'll work on identifying simple patterns in a data set.
4. **Lexicon Expansion**: Often information in text will be associated with a particular vocabulary. Analyses can gain statistical power if we expand human-determined vocabularies using machine learning.
5. **Comparing Groups of Text**: One common use case is to simply compare several groups of text too each other. Often important insights can be gleaned immediately with this simple technique.
6. **Classifying Text**: Next along the chain is taking new text and putting it *into* these kinds of groups. The work from the previous item lays the foundation for classification.
7. **Sentiment Analysis**: Often we can use text mining tools to understand the sentiment of the people who created a body of text. Many organizations use this type of data to monitor consumer or member satisfaction.
8. **Topic Modeling**: Given a large body of text, it is often useful to perform an analog of cluster analysis and attempt to put the text into groups. While it is rarely possible to do this perfectly, topic models can help us understand a body of text.
9. **Vector-based Representations**: Without question the new hotness in text mining and natural language processing is to use neural networks to understand text. These types of models are far beyond the scope of our class, but I'd be negligent to not introduce you to the topics and point out how you could use the results of these models.

My goal for the course is for you to develop a good, basic toolbox that can be used to answer straightforward questions about text data. In this class we're learning how to use the hammers, saws, and wrenches of text mining, but we'll play with some power tools too.

This syllabus has the following sections:

- **Meeting and Communication** How we'll work together this semester.
- **Required Materials** The books required and recommended for the course.

- **Grading** An introduction to our contract grading this semester.
- **The Work** This section describes all the work that you could do in this course. You'll need to read this carefully before deciding on the work you'll contract for.
- **Class Format** A brief introduction to what we'll actually be *doing* in our class meetings.
- **Course Outline** An outline for what we'll be covering when during the course. Subject to change, of course.
- **Code of Contact** My expectations for you regarding things like behavior and collaboration.

Meeting and Communication

Some key things to know about how we'll meet and how we'll communicate with each other:

- **Weekly Meeting:** our class will meet weekly on Wednesdays from 4-6. We'll work to avoid penalizing students who cannot meet at this time. Expect to meet on Zoom for those of you who can join synchronously: https://umontana.zoom.us/my/chandler?pwd=MHpGVGRqbnpnM1dLNHc2TEVSVW9DZz09.
- **Weekly Lab:** There is an optional lab that will meet online on Thursdays from 4-6 to get help on assignments or exercises from class. This lab is also available to the students in Text Mining.
- **Weekly Videos:** Each week I'll post a short(ish) video telling you what's going on this coming week, what you should be thinking about, any deadlines, etc. Pleasepleaseplease watch these videos.
- **Office Hours:** I have office hours at the time indicated above, but I'm happy to meet at other times if those times don't work for you.
- **Moodle:** All file sharing will take place via Moodle[1].
- **Teams:** Teams will be our communication tool for the semester. We're going to use it to collaborate and it's a great way to get help from me and from your peers. You're a member of the UMT Analytics team (let me know if not). Please join the Applied Data Analytics channel within that team.
- **Email:** I'll also use email occasionally to make announcements when I need to make sure they reach the whole class. I'll try to also post this stuff to Slack. Reach me on email via john.chandler@business.umt.edu.
- **Text:** Having an emergency? Text me at 406-544-8720 any time you want. If I'm asleep I'll have notifications off. If you just call me without texting first, I'm unlikely to answer. Who answers unknown numbers these days?

Required Materials

1. Jurasky & Martin text 3rd edition: No need to purchase anything for this, I'll assign material off this webpage.
2. *The Infinite Gift*, Charles Yang, Scribner's, 2006.

Grading in the Course

This course uses a type of grading you may not be familiar with, called "contract grading". If you haven't used it before, it may take some getting used to. **This course does not use points or letter grades**, other than the final grade you decide on (or "contract for") at the beginning and receive at the end. Instead of grades, you'll receive feedback on your work and guidance on bringing it up to the standards of the

[1] I think this is true. If it turns out not to be true, I'll mention it several times and let's all work to amplify the message.

course. Once your finished product is approved, the work will be considered completed and you can move on. Another uncommon aspect of this course: you're going to pick your own deadlines (within boundaries). I've explained a bit about *why* I like contract grading in a video posted to Moodle. The following list details the requirements for each grade. Each item is described more fully further down. In the first week of the course you'll choose your workload, your deadlines, and your desired grade.

Passing

To pass the class, I expect the following from you:

- Honoring deadlines you're contracted for.
- Completing any reading for the week before Wednesday of that week.
- Participating fully in your topic discussion group.
- Communicating in a timely fashion with me if you are experiencing any issues with the course.
- Communicating in a timely fashion with your fellow students when you need to coordinate with them.
- Treating everyone involved in the class with respect.

Grade: C

The C level is the path of least resistance. If you complete this level, you'll be exposed to the key ideas of the course, but you'll only do the minimum amount of work. You should choose this option if this semester is shaping up to be a difficult one for you and you don't need a grade higher than a C[2].

- Everything from "Passing".
- Five class assignment submissions.
- Co-pilot on one A&A project.

Grade: B

This is the standard level in the class. This grade represents a balance between workload and comprehensiveness. If you complete this level, you'll gain deeper experience with the key ideas of the course.

- Everything from "Passing" and C.
- One "Acquire and Analyze" project.
- Read *The Infinite Gift*.
- Participate in two book club meetings about *The Infinite Gift*.
- Eight class assignment submissions.
- Share one data set.

Grade: A

This grade represents a more substantial investment in the materials of the class. This level represents the everything you would typically have the option to do in a full-semester class with traditional grading. Here, however, you must do everything to a high standard.

- Everything from "Passing", C, and B.
- Co-lead two book club meetings about *The Infinite Gift*.

[2] Note: to remain a student in good standing in the CoB graduate programs you need a minimum GPA of 3.0 (B average).

- Eleven class assignment submissions.
- Share two data sets.

Summary

The following table has a summary of the work required by grade.

	Grade		
	A	**B**	**C**
Item			
Books			
"The Infinite Gift"	1	1	0
Book Club Meetings	Lead 2	Attend 2	0
Data Set Shares			
Share & Descriptive Stats	2	1	0
Acquire & Analyze			
Project & Write-up	1	1	Co-pilot
Beautiful Mistakes	Optional	Optional	Optional
Class Assignments	11	8	5

The Work

There is a lot of work mentioned in the previous section. This section tells you more about what that work is.

Book & Book Club

The only book we'll read in its entirety for this course is *The Infinite Gift* by Charles Yang. Only students who are going for an A or B will be required to read the book, although everyone is encouraged to. We'll discuss the book during some synchronous classes.

You'll also meet twice in small groups that I'll assign you to. This is an opportunity to have a deeper discussion of the material in the book and work together a bit more. Students going for a B are required to participate in the book club; students going for an A are required to co-lead the discussion and submit a write-up for your group. There's more information in the appropriate Moodle section.

Data Set Share

In this assignment you'll share new data sets with your classmates. The basic idea behind the assignment is to get good at sourcing text-based data, to understand the kinds of analyses you might do, and to get practice at asking interesting questions about those data sets. Essentially, this is the first half of an "acquire and analyze" assignment.

When you share your data set, you'll share the actual data (or a sample of the data), as well some additional information to help orient your classmates to the data. Your share will include a handful of items submitted to the forum in the "Data Set Share" section of Moodle. Your data set should be new. Please check the previously submitted data sets to make sure you aren't duplicating work. Students going for an A, B, or C will share 2, 1, or 0 data sets, respectively.

Acquire and Analyze Project

This assignment will require you to find a data set and analyze it, though we will have a liberal interpretation of "analyze". If your data set does not require programming to assemble it, please ask for my approval before using it. For the analysis you'll apply a technique we cover in class. The deliverable for this assignment will be your code and a write-up of your project. You can do this project on one of your "Data Set Share" assignments!

For students going for a C, you're required to be a "co-pilot" on someone else's A&A project. We'll match you up with a project you're interested in around week 10. Your job is to help with coding, editing, and idea generation. Essentially, you're a member of the project team.

Class Assignment Submissions

Many weeks we'll do work in class that supports the material covered in lectures or readings. There will be at least 14 assignments that you can optionally submit to have your work checked and to fulfill the requirements of your contract. Typically, these can be finished in 2-4 hours. You will submit your code for review, and I'll ask you for revisions as needed to complete the work and format your code professionally.

Code will be submitted via an assignment in GitHub Classroom[3]

One sign of mature code is conforming to a style guide. When you work for a company, you'll probably have a style guide you use. Since we don't, I recommend the following guides for your work. Work for contract credit won't be accepted with major style guide violations.

- Python: Use the Google Style Guide
- R: Use either the Google Style Guide or the Hadley Wickham Style Guide
- SQL: I don't have a canonical style guide for SQL, but this one is quite good.

Important Note: I'll often post code to work on that accompanies the lectures. When I do this, I'll call them "exercises". These are distinct from "assignments", which you'll do to fulfill this portion of the contract. If you ever have any confusion, ask me!

Beautiful Mistakes

This is an assignment from Applied Data Analytics and is *totally optional* for our class. I'm adding that part of the syllabus here, however, to encourage you to participate. Here's what that syllabus says:

> I'm not sure where the term "beautiful mistakes" comes from, although I know Bob Ross liked to use the term "happy accidents" and I watched a lot of him on PBS as a kid. My real inspiration is a lyric from R.E.M.'s song "World Leader Pretend" off the album *Green*:
>
> > *This is my mistake, let me make it good.*
> > *I raised the wall and I will be the one to knock it down.*
>
> Making mistakes is part of learning. If you're not making mistakes in this class, you're either extremely good at everything we're doing or not pushing yourself hard enough. I want to create

[3] This is subject to change. I'm trying to bend GitHub Classroom to my will, with limited success so far.

a culture where we make mistakes, learn from them, and get better collectively. I'm going to ask you to document some of your programming mistakes in a Slack channel.

I've created a channel in our Slack called #beautiful-mistakes. During the semester I've asked you to post to that channel a variable number of times, depending on your contract. A post should look something like this:

- What you were trying to do.
- Your original code.
- A *clear explanation* of what was wrong with the original code.
- The corrected code.

These can be simple; there are many common mistakes that will be hugely beneficial.

Class Format

Class format will primarily be hands-on work. Lectures will be delivered asynchronously via posted YouTube videos. The lectures will introduce new technical material, analyze real-world implementations of data science techniques, and serve as refreshers for the advanced marketing and technical material. You will also receive code to run and modify before our classes. The hands-on work in class will extend that work, so it is critical that you seek help if you cannot get the pre-work to run on your machine. Plan on sharing your screen during class, so make sure to mute any notifications that could be embarrassing.

Classes will be recorded and Zoom links will be posted as soon as possible after the class. (Typically processing the video file takes about 30 minutes.) Asynchronous students are encouraged to "skim" the class video. I'll often start classes with some comments on the lecture and reading. Then we'll collectively discuss the topics and the work for the day. Typically, there will be some boring parts where people are working—those should be skippable. Please let me know how I can make it easier for you to consume content asynchronously.

There is an optional class meeting on Thursday, from 4:00-6:00, which is a lab. I'm happy to discuss topics from readings or lecture, but this lab will co-convene with students from the Applied Data Analytics class. This is a time to make progress on your work with the ability to get help.

Course Outline

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

	Topics	Assignments
Week 1	Processing Text Data	Spelling Bee Solver
Week 2	Processing Text Data	Spelling Bee Pt 2 & Processing Headlines
Week 3	Patterns in Text	Word Generator
Week 4	Patterns in Text	Text Patterns
Week 5	Scraping and APIs	Scraping Conventions & Wikipedia API & Pulling Twitter Data
Week 6	Regex and Character Encoding	Regex Crossword
Week 7	Lexicons	Lexicon Expansion
Week 8	Comparing Groups	Comparing Groups
Week 9	Spell Checking	Text Functions
Week 10	Classifying Text	Naïve Bayes

Week 11	Sentiment Analysis	Speaker Conventions
Week 12	Parsing	Parsing (SVO)
Week 13	Topic Modeling	Topic Modeling with LDA
Week 14	Topic Modeling Continued	-
Week 15	Vector-Based Representations	-

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.

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. You cannot submit an Acquire & Analyze project as-is for a Text Mining assignment.

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 University of Montana's College of Business is a collegial learning community dedicated to the teaching, exploration, and application of the knowledge and skills necessary to succeed in a competitive marketplace.

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 (netid@grizmail.umt.edu or 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 School of Business Administration has adopted seven learning goals for our undergraduate students:

- Learning Goal 1 – CoB graduates will possess fundamental business knowledge.
- Learning Goal 2 – CoB graduates will be able to integrate business knowledge.
- Learning Goal 3 – CoB graduates will be effective communicators.
- Learning Goal 4 – CoB graduates will possess problem solving skills.
- Learning Goal 5 – CoB graduates will have an ethical awareness.
- Learning Goal 6 – CoB graduates will be proficient users of technology.

- Learning Goal 7 – CoB graduates will understand the global business environment in which they operate.

MS in Business Analytics – Learning Goals

1. Knowledge and Application:
    - An understanding of a range of analytical and programming techniques
    - Ability to apply appropriate techniques to solve a variety of business/organizational problems
2. Communication:
    - Ability to effectively communicate data analytics results and translate into business decisions.
    - Ability to effectively use data visualization techniques.
3. Ethics/Data Stewardship:
    - An understanding of ethical implications of data stewardship and privacy.
4. Innovation:
    - Ability to harness data analytics to identify new sources of value and 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).
- Use the Twitter API to gather unstructured data for analysis.