Computational Text Analysis

(CJ 502: Special Topics in Communication)
Spring 2022
Thursdays; 4:25p.m. to 6:55p.m.; Room 219 (C&J Building)
Department of Communication & Journalism
The University of New Mexico
Phone: (505)-277-5305 | cjdept@unm.edu

Professor: Mohammad Yousuf, Ph.D.
Office: C&J Building, Room 221
Office Hours: 11 a.m. to noon on Tuesdays and Thursdays **OR** by appointment (or just walk in whenever office door is open)
Email: myousuf@unm.edu

1. COURSE DESCRIPTION

The course covers various topics relating to computational content analysis, an emerging approach to collecting and analyzing a large amount of content (e.g., text). The topics range from web scraping to corpus building to machine learning methods to ethics. This approach is very efficient in finding patterns (e.g., bias), and is becoming increasingly popular among scholars investigating media representation, framing, and public sentiment. The course focuses on concepts and methods relating to natural language processing and machine learning. Specific methods covered in the course include topic modelling, sentiment analysis, and other simpler methods such as dictionary-based text analysis and keyword-in-context. Students get hands-on experience with data collection and analysis. The course also covers some basic aspects of programming that help students understand the methods better. Overall, it equips students with some advanced analytical tools to critically analyze media content. It is important to note that these methods go very well with qualitative methods. This is NOT a quantitative methods or statistics course.

2. LEARNING OBJECTIVES

This course has several learning objectives. By the end of the semester, each student is expected to:

1. Possess intermediate level knowledge of important concepts and methods developed in computer science and adopted by art, business, and social science disciplines to analyze large textual corpora;
2. Have skills using Python, a high-level programming language;
3. Be familiar with and possess intermediate level knowledge of several Python libraries commonly used for computational content analysis;
4. Be able to analyze large corpora of texts using methods such as topic modelling and sentiment analysis; and
5. Be able to pursue self-directed learning of computational methods.

3. READINGS

The course will utilize articles and chapters from several books during the semester. Every week, the instructor will upload PDF copies or provide URLs on Learn for free access to the reading materials. Students are not required to buy any book. See the course calendar and references at the end of this syllabus for a complete reading list.

4. PREREQUISITES & REQUIREMENTS

There are no prerequisites for the course. However, some background in statistical analysis and computer programming may help understand the materials better and faster. The course has been designed to make students competent consumers of computational methods, not methodologists.

Software: The course will use Anaconda, a free distribution of several programming languages for scientific computing. Students will need personal computers with administrator privileges so that they can install and use Anaconda. Additional software (e.g., Mallet) and plug-ins for web-scraping will be used.

Overall Grade Distribution

- Attendance and participation: 20%
- Assignments: 40%
- Research Paper: 40%

Participation. This course aspires to be an open communicative space for collaborative and critical inquiry. You are responsible for engaging with all texts and materials, and coming to all course meetings prepared to discuss and apply them.

Assignments. Over the period of the course, students will complete multiple assignments designed to evaluate their ability to put the lessons into practice. Assignments will involve building text corpora and practicing data analysis using the methods taught in class.

Research Paper. Each student will prepare a research paper (about 5,000 words excluding tables and references) demonstrating their ability to apply at least one method (or methods) covered in the course. This is not a group project.

Presentation of Research Paper. There will be a 'research paper presentation day' before the due date for the Research Paper when each student will present her/his paper in front of classmates.

5. ATTENDANCE/CLASSROOM POLICIES

University-sanctioned events
Students who are part of a university-sanctioned group (e.g., sports or academic team) should provide me with an official letter from a coach/supervisor in the first week of the semester indicating the dates they will be absent.

Religious holidays
If you want to participate in any religious holiday that falls on a class day, please let me know as soon as possible.

Emergencies
If you have a major illness or any crisis such as death in the family or major accident, contact the UNM Dean of Students Office at (505) 277-3361 or doso@unm.edu.

Accommodation/Disabilities
In accordance with University Policy 2310 and the Americans with Disabilities Act (ADA), academic accommodations may be made for any student who notifies the instructor of the need for an accommodation. It is imperative that you take the initiative to bring such needs to my attention, as I am not legally permitted to inquire. Students who may require assistance in emergency evacuations should contact the instructor as to the most appropriate procedures to follow. Contact Accessibility Resource Center at (505) 277-3506 for additional information.

Withdrawal
Please consult your course catalog for withdrawal and refund deadlines. You may withdraw from a course during the first six weeks of the semester without the Dean's approval and the withdrawal (W) will not be noted on your academic record. When students leave the University during a semester and do not complete the withdrawal process, they become liable for grades of "F" in their courses, even though they may have been passing at the time of leaving.

Grade of "Incomplete"
A grade of "Incomplete" is given only when circumstances beyond the student's control have prevented completion of the work of a course within the official dates of the semester or session. Students are responsible for making arrangements with the instructor for resolving an incomplete grade. If you receive an incomplete, it must be completed within one year from the published end day of the semester in which the grade was assigned. If the work is not finished in the allotted time period, the grade changes to an "F."

Academic integrity
You are expected to maintain the highest standards of honesty and integrity in academic and professional matters. The University reserves the right to take disciplinary action, up to and including dismissal, against any student who is found guilty of academic dishonesty or otherwise fails to meet the standards. Any student judged to have engaged in academic dishonesty in coursework may receive a reduced or failing grade for the work in question and/or for the course. Academic dishonesty includes, but is not limited to, dishonesty in quizzes, tests, or assignments; claiming credit for work not done or done by others; hindering the academic work of other

students; misrepresenting academic or professional qualifications within or without the University; and nondisclosure or misrepresentation in filling out applications or other University records.

Diversity
This course encourages different perspectives related to such factors as gender, race, nationality, ethnicity, sexual orientation, religion, and other relevant cultural identities. This course seeks to foster understanding and inclusiveness related to such diverse perspectives and ways of communicating.

6. UNM POLICIES

Title IX: Gender Discrimination
In an effort to meet obligations under Title IX, UNM faculty, Teaching Assistants, and Graduate Assistants are considered "responsible employees" by the Department of Education (see pg. 15). This designation requires that any report of gender discrimination which includes sexual harassment, sexual misconduct and sexual violence made to a faculty member, TA, or GA must be reported to the Title IX Coordinator at the Office of Equal Opportunity.

Accessibility
The American with Disabilities Act (ADA) is a federal anti-discrimination statute that provides comprehensive civil rights protection for persons with disabilities. Among other things, this legislation requires that all students with disabilities be guaranteed a learning environment that provides for reasonable accommodations of their disabilities. If you have a disability requiring accommodation, please contact the UNM Accessibility Resource Center in 2021 Mesa Vista Hall at 505-277-3506. Information about your disability is confidential.

- Blackboard's Accessibility statement (https://www.blackboard.com/blackboard-accessibility-commitment)
- Microsoft's Accessibility statement (https://www.microsoft.com/en-us/accessibility/)
- Anaconda Documentation Accessibility statement (https://docs.anaconda.com/anaconda-repository/user-guide/tasks/pkgs/control-access/)

For Military-Connected Students

There are resources on campus designed to help you succeed. You can approach any faculty or staff for help with any issues you may encounter. Many faculty and staff have completed the GREEN ZONE training to learn about the unique challenges facing military-connected students. If you feel that you need help beyond what faculty and/or staff can give you, please reach out to the Veterans Resource Center on campus at 505-277-3181, or by email at vrc@unm.edu.

Administrative Mandate on Required Vaccination
All students, staff, and instructors are required by UNM Administrative Mandate on Required Vaccinations to be fully vaccinated for COVID-19 as soon as possible, but no later than

September 30, 2021, and must provide proof of vaccination or of a UNM validated limited exemption or exemption no later than September 30, 2021 to the UNM vaccination verification site. Students seeking medical exemption from the vaccination policy must submit a request to the UNM verification site for review by the UNM Accessibility Resource Center. Students seeking religious exemption from the vaccination policy must submit a request for reasonable accommodation to the UNM verification site for review by the Compliance, Ethics, and Equal Opportunity Office. For further information on the requirement and on limited exemptions and exemptions, see the UNM Administrative Mandate on Required Vaccinations.

Requirement on Masking in Indoor Spaces
All students, staff, and instructors are required to wear face masks in indoor classes, labs, studios and meetings on UNM campuses, see masking requirement. Qualified music students must follow appropriate specific mask policies issued by the Chair of the Department of Music and the Dean of the College of Fine Arts. Vaccinated and unvaccinated instructors teaching in classrooms must wear a mask when entering and leaving the classroom and when moving around the room. When vaccinated instructors are able to maintain at least six feet of distance, they may choose to remove their mask for the purpose of increased communication during instruction. Instructors who are not vaccinated (because of an approved medical or religious exemption), or who are not vaccinated yet, must wear their masks at all times. Students who do not wear a mask indoors on UNM campuses can expect to be asked to leave the classroom and to be dropped from a class if failure to wear a mask occurs more than once in that class. With the exception of the limited cases described above, students and employees who do not wear a mask in classrooms and other indoor public spaces on UNM campuses are subject to disciplinary actions.

Change in Modality
The President and Provost of UNM may direct that classes move to remote delivery at any time to preserve the health and safety of the students, instructor and community. Please check [fill in your communication system] regularly for updates about our class and please check https://bringbackthepack.unm.edu regularly for general UNM updates about COVID-19 and the health of our community.

CJ 502 Calendar for Spring 2022 (likely to change)

	Topics Covered	Readings
Week 1 (Jan. 20)	**Introduction to Computational Text Analysis** • Computational Text Analysis • History • Recent applications • Promise and pitfalls	DiMaggio, 2015; Grimmer & Stewart (2013); Nelson (2020); Tausczik & Pennebaker (2010).
Week 2 (Jan. 27)	**Web Scraping** • Building a scraper • Scraping with Python • HTML parsing • Using APIs	Mitchell (2018) (*Note*: eBook is available at library.unm.edu)
Week 3 (Feb. 3)	**Computational Linguistics** • Computational models of language • Language features • Computational semantics	Bengfort, Bilbro and Ojeda (2018), Chapter 5. Additional readings: Wintner (2013); Fox (2013);
Week 4 (Feb. 10)	**Introduction to Python and Related Libraries** • Installation of Anaconda • Jupyter notebook • Introduction to related libraries: pandas, nltk, spacy, regex • Python syntax • Data types, variable, function	https://www.python.org/; https://www.anaconda.com/; https://pandas.pydata.org/; https://www.nltk.org/; https://spacy.io/; https://docs.python.org/3/library/re.html
Week 5 (Feb. 17)	**Corpus Building and Data Preprocessing** • Accessing and processing text • Encoding • Regular expressions for pattern detection • Normalizing text • Segmentation	Bengfort, Bilbro and Ojeda (2018), Chapter 3; Bird, Klein and Loper (2009), Chapter 3; Denny and Spirling (2018)

Week 6 (Feb. 24)	**Vectorization and Transformation** • Information retrieval • Vectorization • TF-IDF • Word co-occurrences • Statistical properties of text	Bengfort, Bilbro and Ojeda (2018), Chapter 4; Manning, Raghavan, and Schutze (2009), Chapters. 1–2; Turney and Pantel (2010).
Week 7 (Mar. 3)	**Natural Language Processing** • Fundamentals of natural language processing • N-gram models • Part-of-speech tagging • Lexicons	Bengfort, Bilbro and Ojeda (2018), Chapter 7; Bird, Klein and Loper (2009), Chapter 5
Week 8 (Mar. 10)	**Introduction to Machine Learning** • Classification • Supervised learning • Unsupervised learning	Bengfort, Bilbro and Ojeda (2018), Chapters 5 – 6; TBD
Week 9 (Mar. 17)	**SPRING BREAK: NO CLASS**	
Week 10 (Mar. 24)	**Topic Modelling** • Latent Dirichlet Allocation (LDA) • LDA with Mallet • LDA with Python	Blei (2012); DiMaggio, Nag and Beli (2013); Shahin, 2016
Week 11 (Mar. 31)	**Topic Modelling** • Bi-Term modelling • Correlated modelling • Supervised modelling • Dynamic modelling • Word-Embedding	Blei and Lafferty (2006); Blei and McAuliffe (2010).; Kozlowski, Taddy and Evans (2019); Mikolov, Chen, Corrado and Dean (2013).
Week 12 (Apr. 7)	**Sentiment Analysis** • Lexicon-based methods • Vader	Baccianella, Esuli and Sebastiani (2010); Flores (2017); Hutto and Gilbert (2014); Taboada, Brooke, Tofiloski, Voll and Stede (2011)
Week 13 (Apr. 14)	**Classifiers** • Features and classes • Naïve Bayes classifiers	Bird, Klein & Loper (2009), Chapter 6

	• Nearest neighbor classifiers • Multi-class problems	
Week 14 (Apr. 21)	**Reliability and Validity of Computational Methods** • Intercoder reliability • Cross validation • Model evaluation	Bengfort, Bilbro and Ojeda (2018), Chapter 5; Lombard, Snyder-Duch and Bracken (2002).
Week 15 (Apr. 28)	**Ethics; Introduction to Deep Learning**	Bengfort, Bilbro and Ojeda (2018), Chapter 12; Salganik (2017), Chapter 6.
Week 16 (May. 5)	**Presentation of Research Papers**	
Exam Week	**Final proposal due on May 9**	

References

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).

Bengfort, B., Bilbro, R., & Ojeda, T. (2018). *Applied text analysis with python: enabling language-aware data products with machine learning (First)*. O'Reilly Media.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with python. O'Reilly Media. Available at: NLTK Book

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113-120).

Blei, D. M., & McAuliffe, J. D. (2010). Supervised topic models. arXiv preprint arXiv:1003.0783.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing Systems*, 18, 147.

Christopher, D. M., Raghavan, P., & Schutze, H. (2009). *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis, 26(*2), 168-189.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2(*2), 2053951715602908.

Flores, R. D. (2017). Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using Twitter data. *American Journal of Sociology, 123(*2), 333-384.

Fox, C. (2013). Computational semantics. In A. Clark, C. Fox, & S. Lappin (Eds.), *The handbook of computational linguistics and natural language processing* (pp. 394-428). Wiley.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267-297.

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 8, No. 1).

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84(*5), 905-949.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28(*4), 587-604.

Manning, C., & Schutze, H. (1999). *Foundations of statistical natural language processing.* MIT press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mitchell, R. E. (2018). *Web scraping with python: Collecting more data from the modern web (Second).* O'Reilly Media.

Nelson, L. K. (2020). Computational grounded theory: A methodological framework. *Sociological Methods & Research, 49(*1), 3-42.

Pratt-Hartmann, I. (2013). Computational complexity in natural language. In A. Clark, C. Fox, & S. Lappin (Eds.), *The handbook of computational linguistics and natural language processing* (pp. 43-73). Wiley.

Salganik, M. (2017). Bit by bit: Social research in the digital age. Princeton, NJ: Princeton University Press. (Available online: http://www.bitbybitbook.com/en/ethics/)

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37(*2), 267-307.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29(*1), 24-54.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research, 37*, 141-188.

Wintner, S. (2013). Formal language theory. In A. Clark, C. Fox, & S. Lappin (Eds.), *The handbook of computational linguistics and natural language processing* (pp. 11-42). Wiley.