PS9594A: Computational Text Analysis

Department of Political Science – Western University, Winter 2024
Monday 9:00-12:00am, SSC 7210

Instructor: Dr. Sebastián Vallejo Vera (`sebastian.vallejo@uwo.ca`)
Office hours: Monday 12:00-2:00pm or by appointment... mostly by appointment (SSC 7221)

Course description

One of the most abundant sources of data available to social and political scientists today is text. Recent advances in Natural Language Processing (NLP) have spearheaded a text-as-data revolution, which has led social scientists to seek out new means of analyzing text data at scale. In this course, we will learn the intuition behind—and how to implement—different computational methods to process, analyze, and classify text. The course will cover Bag-of-Words (BoW) approaches, unsupervised methods, supervised and semi-supervised methods, and generative methods that use text as data, as well as how we can interpret the results obtained from applying these methods.

Course objectives

In this course, students will:

- learn how to use text as data;
- understand the potential and the limitations of using text as data;
- get training on how to use computational text analysis techniques;
- learn how to obtain and process text data.

Acknowledgments

The organization of the first part of this course (Weeks 1 - 5) and the format of the assignments are borrowed from Christopher Barrie's excellent course on "Computational Text Analysis", a syllabus from the prolific Tiago Ventura, and Grimmer, Roberts, and Stewart's excellent book, "Text as data: A new framework for machine learning and the social sciences". The code used throughout the course is a patchwork of my own code, but my own code borrows heavily from the internet (but that's true for all code). I try my best to give credit to the original authors of the code (when and if possible).

Readings and Slides

The main external platform we will use in this class is Perusall (`www.perusall.com`). Perusall is a free collaborative annotation tool that allows you to analyze texts collaboratively. All the required texts for the class, the most important supplementary readings, the dates on which you should have them completed, and the class slides are available on Perusall.

Instructions to register on Perusall:

1. Go to www.perusall.com.
2. Create an account (you can use your institutional email as well as your personal email).
3. Accept the Terms of Services and Privacy Policies (you should read them, but you probably won't).
4. Select the option "Create or enroll in a course," and then choose "I'm a student."
5. You will be prompted to enter the course code. The course code is VALLEJO-WNW2T.
6. It will ask you to enter your Student ID in addition to your name. Enter your UWO student number.
7. Done! You should now have access to the course page.

If you couldn't access the course, you can also watch this video: https://www.youtube.com/watch?v=lbfo7Yusdi8.

Code

All the code for the class will be posted at https://svallejovera.github.io/cpa_uwo/.

Course assessment

Students will be assessed as follows:

- **Homework (40%)**: There will be four worksheets (Exercise #N) that will walk you through the implementation of different text analysis techniques. At the end of each worksheets, you will find a set of questions. You should partner up with someone else in your class and go through these together.[1] In the next class, I will pick on a pair at random to answer each one of that worksheet's questions, and walk us through your code. This is not a punitive exercise, but rather a space for collaborative learning. More often then not, the obstacles encountered by one person are also encountered by many other. Furthermore, there are many ways to arrive to the same solution, and being exposed to different frameworks is beneficial to all. All that matters to me is that you **try** and, eventually, learn. (Same goes for those who only have to turn in the assignment).
- **Reading Review (10%)**: On the weeks when no assignment is due, I will randomly choose a pair of students to **briefly** explain the research question and the method used in one of the assigned papers for that week. Focus on the following: Did the paper answer the research question? Was the data appropriate for answering that question? Was the method appropriate for answering that question? Are the conclusions arrived at by the author supported by their evidence?
- **Final Essay (50%)**: A 4000-word **max** essay. Further instructions are at the end of the syllabus.

[1] From Barrie: "This is called pair programming and there's a reason we do this. Firstly, coding can be an isolating and difficult thing–it's good to bring a friend along for the ride! Secondly, if there's something you don't know, maybe your partner will. This saves you both time. Thirdly, your partner can check your code as you write it, and vice versa. Again, this means both of you are working together to produce and check something as you go along."

Class Expectations

1. **Always be respectful and mindful of your classmates.**
2. The class starts at 9:00 AM. It is as early for you as it is for me. Please, be on time and awake, or somewhat awake, or faking being awake.
3. I will start the class at 9:00 AM with whoever is in the room. Arriving late? No problem. Just enter discreetly and quietly, take your seat, and we are all good. 9:15 AM is not the time to greet, chat, wave vigorously to your friends in the room. When you do this, you distract those that were on time and you distract me (it is also disrespectful, see point 1).
4. If you are going to be taking notes in your laptop/iPad, close all other tabs that might distract you from the lecture. The secret is to hang to my every word.
5. I cannot make you pay attention and participate. But I can ask you to avoid distracting the rest of the class. Remember: I already know the material. The important part is for you to learn it.
6. If you are going to be watching TikTok during class anyways, at least drop the links to the really funny ones.
7. I do not care if you are wearing pajamas, but please come to class. Worst case scenario, the material presence of your being might allow you to learn through osmosis.

A quick yet important note on cellphones: Our class is 120 minutes long. Most things in life can wait two hours to be resolved/answered/liked/swiped-right/retweeted/watched/poked/high-fived/instagramed/swiped-left/live-streamed. There is no need for you to have your cellphone out and about (yes, I notice when you are in your phone even when you try to hide it under your desk). If, for some reason, you need to have your cellphone out, please let me know before class (you know, as a courtesy).

Children in Class

I applaud all of you who go to school with children! It is difficult to balance academia, work, and family commitments, and I want you to succeed. Here are my policies regarding children in class:

1. All breastfeeding babies are welcome in class as often as needed. If your baby requires your attention, you can step outside and tend to them.
2. Non-nursing babies and older children are welcome as well. As a parent of two school-age children, I understand that babysitters fall through, partners have conflicting schedules, children get sick, and other issues like a global pandemic arise that leave parents with few other options. If you child requires your attention, you can step outside and tend to them.
3. All students are expected to join me in creating a welcoming environment that is respectful of your classmates who bring children to class.

I understand that sleep deprivation and exhaustion are among the most difficult aspects of parenting young children. The struggle of balancing school, work, childcare, and high inflation is tiring, and I will do my best to accommodate any such issues while maintaining the same high expectations for all students enrolled in the class. Please do not hesitate to contact me with any questions or concerns.

Late Work Policy

Legally defined adults are late with things ALL THE TIME (myself included).

That said, deadlines serve their purpose. They can create an external structure to help you plan your workload and prevent everything from piling up on you. Furthermore, we live (and learn) in a community. When I take longer to submit a paper revision to a journal, I make the editor's job more complicated. If many of you turn in your work late, it makes planning the material we need to cover more challenging for me. Finally, there are deadlines that are more absolute than others. If the plane closes its doors at 10:00 AM and you arrive at 10:15 AM, there's no earthly power that can reopen them.

In this class, there are two types of deadlines: 1) the fatal ones, which are deadlines that cannot be postponed, and 2) the non-fatal ones, which are suggestions and planning guides (rather than arbitrary and punitive dates meant to generate anxiety). The fatal deadlines are those that are immovable for practical reasons. For example, any work submitted to me after the deadline I must submit grades will not be considered because, well, I will have already submitted grades. Similarly, due to their nature, the Final Exam must be submitted within the agreed-upon times.

The non-fatal deadlines are more flexible. While I **strongly recommend** that you keep up with the class schedule, I also acknowledge that things happen (e.g., global pandemics, climate crises, life events). Since I don't want your assignments to pile up and I also don't want you to feel like you must disappear if you submit something late, for the rest of the deadlines (e.g., homework), I have adopted a more "liberal" policy with extensions. The only thing I ask is that you proactively communicate with me to find solutions for any delays that will allow you to successfully complete the course. Note that, even if there is no penalty for late submission, if you submit an assignment late, you might also get late feedback, which might lead to knowledge gaps during lectures.

Finally, remember that I also have a life outside the classroom, and it is partly scheduled around important course dates. If you submit an assignment late, there's a good chance it will take us longer to return it corrected.

Course Structure

IMPORTANT NOTE: All Exercises are due on the following week from when they are posted. For example, Exercises #1 is due on Week #4.

Week #1 (January 8): Course Introduction / Why (Computational) Text Analysis?

Topics: Review of syllabus and class organization. Introduction to computational text analysis and natural language processing (NLP).

READINGS:

1. Grimmer, Roberts, and Stewart - Ch. 2.

CODE:

- No Code.

Week #2 (January 15): Tokenization and Word Frequency

Topics: What is a Bag of Words? What are tokens? Why should we care about tokens?

READINGS:

1. Grimmer, Roberts, and Stewart - Ch. 5;
2. Ban, P., Fouirnaies, A., Hall, A. B., & Snyder, J. M. (2019). How newspapers reveal political power. Political Science Research and Methods, 7(4), 661-678;
3. Michel, J.B., et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." Science 331 (6014): 176–82. `https://doi.org/10.1126/science.1199644`;
4. Bollen, J., et al. (2021). Historical language records reveal a surge of cognitive distortions in recent decades. Proceedings of the National Academy of Sciences, 118(30), e2102061118.

CODE:

- Code #1.

Week #3 (January 22): Dictionary-Based Techniques

Topics: What are dictionaries? Why/when are they useful? What are their limitations?

READINGS:

1. Grimmer, Roberts, and Stewart - Ch. 15-16;
2. Young, L., and Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. Political Communication, 29(2), 205-231;

3. Martins, M. D. J. D., & Baumard, N. (2020). The rise of prosociality in fiction preceded democratic revolutions in Early Modern Europe. Proceedings of the National Academy of Sciences, 117(46), 28684-28691;

4. Ventura, T., Munger, K., McCabe, K., & Chang, K. C. (2021). Connective effervescence and streaming chat during political debates. Journal of Quantitative Description: Digital Media, 1.

CODE:

- Code #2 + Exercise #1.

Week #4 (January 29): Natural Language, Complexity, and Similarity

Topics: How do we evaluate complexity in text? Why should we care about complexity in text? How do we evaluate similarity in text? Why is this useful?

READINGS:

1. Grimmer, Roberts, and Stewart - Ch. 7;

2. Spirling, A. (2016). Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832–1915. The Journal of Politics, 78(1), 120-136.

3. Urman, A., Makhortykh, M., & Ulloa, R. (2022). The matter of chance: Auditing web search results related to the 2020 US presidential primary elections across six search engines. Social science computer review, 40(5), 1323-1339;

4. Schoonvelde, M., Brosius, A., Schumacher, G., & Bakker, B. N. (2019). Liberals lecture, conservatives communicate: Analyzing complexity and ideology in 381,609 political speeches. PloS one, 14(2), e0208450.

CODE:

- Code #3.

Week #5 (February 5): Scaling Techniques and Topic Modeling (Unsupervised Learning I)

Topics: What are scaling models and what can they tell us? What is unsupervised learning? What is topic modeling and what can it tell us?

READINGS:

1. Grimmer, Roberts, and Stewart - Ch. 12-13;

2. Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. American political science review, 97(2), 311-331.

3. Slapin, J. B., & Proksch, S. O. (2008). A scaling model for estimating time-series party positions from texts. American Journal of Political Science, 52(3), 705-722.

4. Roberts, M. E., et al. (2014). Structural topic models for open-ended survey responses. American journal of political science, 58(4), 1064-1082.

5. 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.

6. Motolinia, L. (2021). Electoral accountability and particularistic legislation: evidence from an electoral reform in Mexico. American Political Science Review, 115(1), 97-113.

CODE:

- Code #4 + Exercise #2.

Week #6 (February 12): Word Embeddings (Unsupervised Learning II)

READINGS:
Topics: What are word-embeddings? When and how can we use them? What? Topic models again?

1. Grimmer, Roberts, and Stewart - Ch. 8;

2. Meyer, D. (2016). How exactly does word2vec work?. Uoregon. Edu, Brocade. Com, 1-18;

3. Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. The Journal of Politics, 84(1), 101-115;

4. 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.

CODE:

- Code #5.

(February 19): Spring reading week. Enjoy the break!

Week #7 (February 26): General Review and Getting Data

Topics: We will review all the material covered so far. We will learn (creative) ways to obtain text (and the ethical and legal implications of doing it).

READINGS:

1. Wilkerson, J., & Casas, A. (2017). Large-scale computerized text analysis in political science: Opportunities and challenges. Annual Review of Political Science, 20, 529-544;

2. Macanovic, A. (2022). Text mining for social science–The state and the future of computational text analysis in sociology. Social Science Research, 108, 102784;

3. Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. Social Science Computer Review, 33(6), 712-729;

4. Michalopoulos, S., & Xue, M. M. (2021). Folklore. The Quarterly Journal of Economics, 136(4), 1993-2046.

CODE:

- Code #6 + Exercise #2.5 (Optional).

Week #8 (March 4): Supervised Learning I

Topics: We will study the framework to train supervised models, and when to use them. We will learn how Support Vector Machine (SVM) and Bidirectional Long-Short Term Memory (Bi-LSTM) models work.

READINGS:

1. Grimmer, Roberts, and Stewart - Ch. 17-20;
2. Siegel, A. A., et al. (2021). Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. Quarterly Journal of Political Science, 16(1), 71-104.
3. Barberá, P., et al. (2021). Automated text classification of news articles: A practical guide. Political Analysis, 29(1), 19-42.

CODE:

- Code #7 + Exercise #3.

Week #9 (March 11): Supervised Learning II: Transformers Architecture

Topics: We will learn about the Transformers architecture, attention, and the encoder-coder infrastructure.

READINGS:

1. Jay Alammar. 2018. "The Illustrated Transformer";
2. Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30;
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805;
4. Timoneda, J.C., and S. Vallejo Vera. BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformer Models in Political Science Text. Forthcoming Journal of Politics.

CODE:

- Code #8.

Week #10 (March 18): Supervised Learning III: Do We Even Need Theory Anymore?

Topics: Given the computational advances of text analysis, do we even need social scientists anymore? Theory? What is our role as social scientists in an increasingly computational field?

READINGS:

1. Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. Journalism quarterly, 30(4), 415-433;
2. Dávila Gordillo, D., J.C. Timoneda, and S. Vallejo Vera. Machines Do See Color: A Guideline to Classify Different Forms of Racist Discourse in Large Corpora. Working Paper.

CODE:

- Code #9 + Exercise #4.

Week #11 (March 25): Llama 2, ChatGPT, and Generative Modeling

Topics: How do ChatGPT, Llama2, and other generative models work? Other than five-second essays, what else can we do with them?

READINGS:

1. Touvron, H., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.;
2. Kadous, W. (2023). Llama 2 is about as factually accurate as GPT-4 for summaries and is 30X cheaper. Read here;
3. Ramlochan, S. (2023). How Does Llama-2 Compare to GPT-4/3.5 and Other AI Language Models. Read here.

CODE:

- Code #10

Week #12 (April 1): Catch-up Week and Concluding Remarks

Topics: We will probably be somewhat behind at this point. We will use this week to catch-up. I will also answer questions related to the final assignment. We will close the course with some concluding remarks.

READINGS:

1. Time to catch-up on all the readings.

CODE:

- No Code.

Final Essay Instructions

The objective of this activity is for you to write a 4000-word **max** essay on a subject previously approved by me. Think of it as a research note for a journal (see, for example, the Letters at APSR). I will provide a range of data sources to choose from, but you are welcome (encouraged) to suggest your own.

The essay should include *all* of the following:

1. a research question;
2. at least one computational text analysis technique that we have studied;
3. an analysis of the data source you have provided;
4. a write up of the initial findings;
5. an outline of potential extensions of your analysis.

You will also need to provide the code you used in reproducible (markdown) format and will be assessed on both the substantive content of your essay contribution (the social science part) and your demonstrated competency in coding and text analysis (the computational part). While you will only be graded on the final submission, I encourage you to submit advances and/or come to office hours to revise your work. I am happy to provide feedback on your essay's substantive and technical aspects. If this applies to you, use this opportunity to work on sections of your thesis/dissertation/working paper that will require or can be further developed by applying computational text analysis techniques.