HSS 510: Natural Language Processing for Humanities and Social Sciences

Spring 2024
Wed 9:00–12:00pm
N4 1309, School of Digital Humanities and Computational Social Sciences

Instructor: Taegyoon Kim, Ph.D. in Political Science and Social Data Analytics

- Email: taegyoon@kaist.ac.kr
- Office hours: Wed 1:30pm–2:30pm & by appointment, N4 1308
- Webpage: link

Teaching assistant: Jaehong Kim

- Email: luke.4.18@kaist.ac.kr
- Webpage: link

Course overview: This course introduces students to the fundamental concepts and techniques of Natural Language Processing (NLP), emphasizing the development of critical insights for its application in humanities and social sciences research. The course will blend theoretical understanding with practical, hands-on experience. Students will develop not only a mathematical/statistical intuition for key NLP approaches but also learn how to effectively implement these approaches in their own research. Each class will start with a lecture by the instructor, complemented by guided coding. The latter portion of the class will feature two activities. Students will first engage in a review of applied research. This will be followed by a student's hands-on methods tutorial specifically tailored to the topic of the week. While prior experience in NLP is not required, students should possess basic programming skills in Python (knowledge of both Python and R preferred), along with some familiarity with quantitative analysis. By the end of this course, students will gain a comprehensive understanding of NLP's potential in humanities and social sciences, mastering techniques to apply and refine NLP for their research.

Prerequisites: The course will assume that students have a base facility with Python and some level work in quantitative analysis.

Textbooks: Most of our readings will be articles. We will not adhere strictly to any single textbook but read multiple sections of each of the following books.

- [**GRS**] Grimmer, J., Roberts, M.E. and Stewart, B.M., 2022. Text as data: A new framework for machine learning and the social sciences. Princeton University Press. [link]
- [**JM**] Jurafsky, D. and Martin, J.H., Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. [link] [**This book has just been updated (Feb 2024). The chapter numbers below follow the 2023 version**]

Major tasks: Students are expected to complete the following tasks (the numbers in parentheses are grade values). Students should sign up for one slot both for **Application review discussion** (each discussion will be led by up to two students) and for **Method tutorial** (each tutorial will be presented by two students). Sign up on this Google Sheet.

- Attendance (10)
    - Students must attend all lectures unless exceptional circumstances have been discussed with the instructor beforehand.
    - Two points will be deducted from your attendance score for each absence.
    - Arriving more than twenty minutes late to a class will be counted as an absence.
    - Students are assumed to have completed assigned readings and expected to actively participate in class (e.g., application review discussion).
- Application review discussion (20): (A pair of) students will be responsible for presenting a review of and leading discussion for applied article that employs NLP.
    - The objectives include developing the ability to critically assess applied NLP research and gaining insights into creatively utilizing NLP for one's own research.
    - For 10–12 minutes, the presenting students will provide an in-depth review of an article focused on the NLP skill applied within it. This should include 1) what the research objective(s)/question(s) are, 2) what text data are used and how they were collected, 3) what NLP methods are used to analyze the data and how, 4) pros and cons about the use of the methods, and 5) how to improve it
    - For the next 10–12 minutes, the class will have an open discussion co-led by the presenting students and the instructor.

- Methods tutorial (20): (A pair of) students will be responsible for presenting a methods tutorial.
    - The presentations serve a dual purpose: firstly, to provide non-presenting students with a hands-on implementation of the techniques covered in that week, and secondly, to offer presenting students an opportunity to apply and adapt those techniques in their own research.
    - For 10–15 min, students will showcase the implementation of the NLP techniques covered in that week using actual data (preferably data collected for—or related to—one's own research).
    - Students will guide the entire class through their script (either in Python or R), explaining each code block in detail.
    - Be sure to provide the class with access to your tutorial and data so that the whole class can code as you walk through your tutorial.
    - Efforts will be made to offer related course materials for each week (slides, scripts, etc.) in advance so that presenting students will have sufficient time and resources to prepare their tutorial presentation.
- Research paper (40): The research paper may be either an application of NLP/text-as-data to answer a substantive/theoretical question in your field of research or a methodological contribution relevant to the literature on NLP/text-as-data. The objective is to use this as an opportunity to write (and eventually publish) a substantive research paper (do *not* write a course paper to write a course paper!). Students can either work alone or collaborate in teams of up to two.
    - Students are required to submit a one-page description of their proposed paper by **April 17**, but I strongly encourage you to start formulating your idea as soon as you can. Note that the proposal itself is ungraded.
    - Relatedly, ensure that you schedule an one-on-one meeting regarding your proposal with the instructor at your earliest convenience (by **March 29** at the latest)
    - The paper may *not* be data collection just for the sake of data collection. You are required to engage with analytic NLP skills in your paper.
    - If your research objective requires something beyond the scale of what is possible in this course's time frame, you should modify your objective. For example, consider a pilot study on a smaller sample or on a easily available proxy for the ultimate data of interest.

- Papers must be in a format plausible as a submission to an appropriate peer-reviewed outlet. This means that you a substantive title, an abstract, a main text, references, appendices/supplementary materials (if applicable), and replication materials (data and scripts).
- The expected length of the paper is a short paper (3–4,000 words), but feel free to write a longer paper if that is what you wish to do (e.g., a 10,000-word social science journal article).

- Exercises (10): Multiple take-home exercises will be assigned throughout the semester.

Grading scale: Grade values will not be rounded. That is, any grade value that is greater than or equal to 'Lower' and less than 'Upper' will receive the respective grade.

Grade	Lower	Upper
A+	90	101
A_0	87	90
A-	84	87
B+	81	84
B_0	78	81
B-	75	78

Grade	Lower	Upper
C+	72	75
C_0	69	72
C-	66	69
D+	63	66
D_0	60	63
F	0	60

Weekly schedule

The weekly schedule may be modified as needed to align with the class's overall progress and the varying levels of comprehension. 1) The article marked with a cross (†) is designated for application review. 2) Some courses will be held via Zoom, with students participating either individually from different locations (individual/virtual) or together in the classroom (in-class/virtual). 3) Be sure to bring your laptop for every class.

Week 1. Introduction (Feb 28)

- Course overview (key objectives, class components, weekly themes, etc.)
- Introducing background survey (motivations to take the course, research interests, programming skills, backgrounds in quantitative analysis, previous experiences in NLP/text-as-data)
- Complete the pre-course survey
- Make sure that you have installed Python and R by the next class (consult with the TA if you need help)

Week 2. Selecting and cleaning texts (Mar 6)

- Required reading
  - [**GRS**] Sections 3.1 and 3.2 in Chp. 3 "Principles of Selection and Representation."
  - [**GRS**] Chp. 4 "Selecting Documents."
  - [**JM**] Sections 2.1 and 2.2 in Chp. 2 "Regular Expressions, Text Normalization, Edit Distance"

* The last part of the class will discuss web data collection (e.g., web scraping and API)

Week 3. Representing and comparing texts (Mar 13)

- Required reading
  - [**GRS**] Chp. 5 "Bag of Words."
  - [**GRS**] Chp. 7 "The Vector Space Model and Similarity Metrics."
  - †Denny, M.J. and Spirling, A., 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. Political Analysis, 26(2), pp.168-189.

- Optional reading
  - [**GRS**] Sections 3.3 and 3.4 in Chp. 3 "Principles of Selection and Representation."
  - [**JM**] Section 2.4 in Chp. 2 "Regular Expressions, Text Normalization, Edit Distance"
  - Christopher, D., Raghavan, P. and Schütze, H., 2008. Scoring term weighting and the vector space model. Introduction to information retrieval, 100, pp.2-4.

Week 4. Keyword-based methods (Mar 20)

- Required reading
  - [**GRS**] Chp. 16 "Word Counting."
  - [**JM**] Chp. 19 "Lexicons for Sentiment, Affect, and Connotation."
  - †Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A. and Van Bavel, J.J., 2017. Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, 114(28), pp.7313-7318.
- Optional reading
  - [**GRS**] Chp. 15 "Principles of Measurement."
  - Monroe, B.L., Colaresi, M.P. and Quinn, K.M., 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), pp.372-403.
  - Yin, Y., Gao, J., Jones, B.F. and Wang, D., 2021. Coevolution of policy and science during the pandemic. Science, 371(6525), pp.128-130.
  - Young, L. and Soroka, S., 2012. Affective news: The automated coding of sentiment in political texts. Political Communication, 29(2), pp.205-231.

Week 5. Supervised learning methods I (Mar 27)

- Required reading
  - [**GRS**] Chp. 17 "An Overview of Supervised Classification"
  - [**GRS**] Chp. 18 "Coding a Training Set"
  - [**GRS**] Chp. 19 "Classifying Documents with Supervised Learning"
  - [**GRS**] Chp. 20 "Checking Performance"
  - [**JM**] Chp. 5 "Logistic Regression"
  - †Siegel, A.A., Nikitin, E., Barberá, P., Sterling, J., Pullen, B., Bonneau, R., Nagler, J. and Tucker, J.A., 2021. Trumping hate on Twitter? Online hate speech in the 2016 US election campaign and its aftermath. Quarterly Journal of Political Science, 16(1), pp.71-104.

- Optional reading
  - Piper, A., 2022. Biodiversity is not declining in fiction. Journal of Cultural Analytics, 7(3).
  - [**JM**] Chp. 4 "Naive Bayes and Sentiment Classification"
  - Barberá, P., Boydstun, A.E., Linn, S., McMahon, R. and Nagler, J., 2021. Automated text classification of news articles: A practical guide. Political Analysis, 29(1), pp.19-42.

Week 6. Supervised learning methods II (Apr 3) [**individual/virtual: Zoom link**]

- Required reading (tentative)
  - [**JM**] Sections 7.1–7.4 and 7.6 in Chp. 7 "Neural Networks and Neural Language Models"
  - Miller, B., Linder, F. and Mebane, W.R., 2020. Active learning approaches for labeling text: review and assessment of the performance of active learning approaches. Political Analysis, 28(4), pp.532-551.
  - †Benoit, K., Conway, D., Lauderdale, B.E., Laver, M. and Mikhaylov, S., 2016. Crowd-sourced text analysis: Reproducible and agile production of political data. American Political Science Review, 110(2), pp.278-295.
- Optional reading
  - Van Atteveldt, W., Van der Velden, M.A. and Boukes, M., 2021. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. Communication Methods and Measures, 15(2), pp.121-140.
  - Bestvater, S.E. and Monroe, B.L., 2023. Sentiment is not stance: Target-aware opinion classification for political text analysis. Political Analysis, 31(2), pp.235-256.
  - Arnold, C., Biedebach, L., Küpfer, A. and Neunhoeffer, M., 2023. The role of hyperparameters in machine learning models and how to tune them. Political Science Research and Methods, pp.1-8.

Week 7. Election day (Apr 10)

- No class

Week 8. Mid-term break (Apr 17)

- No class

- Proposal deadline

Week 9. Embeddings (Apr 24)

- Required reading
  - [**GRS**] Chp. 8 "Distributed Representations of Words"
  - [**JM**] Chp. 6 "Vector Semantics and Embeddings."
  - Rodriguez, P.L. and 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), pp.101-115.
  - †Kozlowski, A.C., Taddy, M. and Evans, J.A., 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. American Sociological Review, 84(5), pp.905-949.
- Optional reading
  - Soni, S., Klein, L.F. and Eisenstein, J., 2021. Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. Journal of Cultural Analytics, 6(1).
  - Caliskan, A., Bryson, J.J. and Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), pp.183-186.
  - Osnabrügge, M., Hobolt, S.B. and Rodon, T., 2021. Playing to the gallery: Emotive rhetoric in parliaments. American Political Science Review, 115(3), pp.885-899.
  - Garten, J., Hoover, J., Johnson, K.M., Boghrati, R., Iskiwitch, C. and Dehghani, M., 2018. Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis: Distributed dictionary representation. Behavior research methods, 50, pp.344-361.
  - Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
  - Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26.
  - Pennington, J., Socher, R. and Manning, C.D., 2014, October. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Week 10. NLP with Korean (May 1) [**in-class/virtual: Zoom link**]

- Guest lecture (Byungjoon Kim, Ph.D. in Data Science, KAIST)

Week 11. Topic models (May 8)

- Required reading
  - [**GRS**] Chp. 13 "Topic Models."
  - Blei, D.M., 2012. Probabilistic topic models. Communications of the ACM, 55(4), pp.77-84.
  - †Barberá, P., Casas, A., Nagler, J., Egan, P.J., Bonneau, R., Jost, J.T. and Tucker, J.A., 2019. Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. American Political Science Review, 113(4), pp.883-901.
- Optional reading
  - Ying, L., Montgomery, J.M. and Stewart, B.M., 2022. Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. Political Analysis, 30(4), pp.570-589.
  - Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), pp.993-1022.
  - Roberts, M.E., Stewart, B.M., Tingley, D. and Airoldi, E.M., 2013, December. The structural topic model and applied social science. In Advances in neural information processing systems workshop on topic models: computation, application, and evaluation (Vol. 4, No. 1, pp. 1-20).

Week 12. Buddha's Birthday (May 15)

- No class

Week 13. Neural NLP I (May 22) (tentative)

- Required reading
  - Smith, N.A., 2019. Contextual word representations: A contextual introduction. arXiv preprint arXiv:1902.06006.
  - Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794.
  - †Falkenberg, M., Galeazzi, A., Torricelli, M., Di Marco, N., Larosa, F., Sas, M., Mekacher, A., Pearce, W., Zollo, F., Quattrociocchi, W. and Baronchelli, A., 2022. Growing polarization around climate change on social media. Nature Climate Change, 12(12), pp.1114-1121.

- Optional reading
  - [**JM**] Chp. 9 "RNNs and LSTMs"
  - [**JM**] Chp. 10 "Transformers and Pretrained Language Models"
  - Suissa, O., Elmalech, A. and Zhitomirsky-Geffet, M., 2022. Text analysis using deep neural networks in digital humanities and information science. Journal of the Association for Information Science and Technology, 73(2), pp.268-287.
  - Chatsiou, K. and Mikhaylov, S.J., 2020. Deep learning for political science. arXiv preprint arXiv:2005.06540.
  - Mooijman, M., Hoover, J., Lin, Y., Ji, H. and Dehghani, M., 2018. Moralization in social networks and the emergence of violence during protests. Nature human behaviour, 2(6), pp.389-396.

Week 14. Neural NLP II (May 29) (tentative)

- Required reading
  - [**JM**] Chp. 11 Fine-tuning and Masked Language Models
  - Rogers, A., Kovaleva, O. and Rumshisky, A., 2021. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8, pp.842-866.
  - †Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R. and Jurafsky, D., 2022. Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. Proceedings of the National Academy of Sciences, 119(31), p.e2120510119.
- Optional reading
  - Laurer, M., Van Atteveldt, W., Casas, A. and Welbers, K., 2024. Widmann, T. and Wich, M., 2023. Creating and comparing dictionary, word embedding, and transformer-based models to measure discrete emotions in German political text. Political Analysis, 31(4), pp.626-641.
  - Wang, Y., 2023. On Finetuning Large Language Models. Political Analysis, pp.1-5.
  - Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J.A. and Bonneau, R., 2020. A comparison of methods in political science text classification: Transfer learning language models for politics. Available at SSRN 3724644.

Week 15. Promises and pitfalls of large language models (Jun 5) (tentative)

- Required reading

- Ziems, C., Shaikh, O., Zhang, Z., Held, W., Chen, J. and Yang, D., 2023. Can large language models transform computational social science?. Computational Linguistics, pp.1-53.
- Chae, Y. and Davidson, T., 2023. Large language models for text classification: From zero-shot learning to fine-tuning. Open Science Foundation.
- Laurer, M., Van Atteveldt, W., Casas, A. and Welbers, K., 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. Political Analysis, 32(1), pp.84-100.
- Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A.S., Ceder, G., Persson, K.A. and Jain, A., 2024. Structured information extraction from scientific text with large language models. Nature Communications, 15(1), p.1418.

- Optional reading
  - Pham, C.M., Hoyle, A., Sun, S. and Iyyer, M., 2023. TopicGPT: A Prompt-based Topic Modeling Framework. arXiv preprint arXiv:2311.01449.

Week 16. Presentations (Jun 12)

Instruction mode: The primary mode of instruction is in-person. Any change to the mode of instruction will be announced in advance.

Email policy: I try to respond to emails promptly, typically within two business days. If you have complex questions or need an in-depth discussion, I encourage you to attend my office hours.

Office hours: I welcome all students to attend my office hours for discussions related to course content and learning strategies. If you need to set up a meeting outside my office hours, send me an email with your availability, and we will arrange a mutually convenient time to meet.

Late submission policy: Late submissions will incur a penalty of 10% for each day (rounded up) beyond the due date

Academic integrity: As students at KAIST, you are entrusted with upholding the utmost standards of academic integrity. Academic honesty is paramount, and any form of misconduct is strictly prohibited. In the event of suspected misconduct, our class adheres to the established policy of KAIST. All such incidents are promptly reported to the dean of the Department of Humanities and Social Sciences to ensure a fair and transparent resolution.

Syllabus change policy: This syllabus is a guide, and every attempt will be made to provide an accurate overview of the course. However, circumstances and events may make it necessary for the instructor to modify the syllabus during the semester and may depend, in part, on the progress, needs, and experiences of the students.