Text as Data – Literature

Week 1 – Text Representation

Books

- 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
- Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press
- Jurafsky, D. and Martin, J. H. (2025). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition. Online manuscript released August 24, 2025
- Schütze, H., Manning, C. D., and Raghavan, P. (2008). *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge
- James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). Statistical learning. In *An introduction to statistical learning: With applications in Python*, pages 15–67. Springer
- Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press

Articles

- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., et al. (2011).

Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182

- Beauchamp, N. (2017). Predicting and interpolating state-level polls using twitter textual data. *American Journal of Political Science*, 61(2):490–503

- Gilardi, F., Shipan, C. R., and Wüest, B. (2021). Policy diffusion: The issue-definition stage. *American Journal of Political Science*, 65(1):21–35

- Gilardi, F., Shipan, C. R., and Wüest, B. (2021). Policy diffusion: The issue-definition stage. *American Journal of Political Science*, 65(1):21–35

- Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542

- 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

Week 2 – Preprocessing

Articles

- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21

- 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):372–403

- 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):168–189