Text as Data
George Mason University, Spring 2023
Lectures: In person, Mondays from 4:30pm to 7:10pm
Krug Hall 19

Instructor: Noel Johnson

Email: njohnsoL@gmu.edu

Office Location: Carow 8

Office Hours: Wednesdays 2:00 to 3:00 or by appointment

Course Webpage: Blackboard

Course Description

The availability of text data has exploded in recent times and so has the demand for analysis of that data. The goal of this class is to learn the fundamentals of quantitative analysis of text from a social science perspective. The emphasis is on applications, and while there will be some theoretical treatment of the topics at hand, the primary aim is to help students understand the types of questions we can ask with text and how to go about answering them. We will discuss how texts may be modeled as quantitative entities and how they might be compared. We will then move to supervised and unsupervised methods. There are three main goals of the seminar: (1) learn about techniques to analyze text as data; (2) learn how to apply these techniques in a practical way using the programming language R; (3) allow graduate students to work on a research project that they will, hopefully, be able to use for their dissertations.

Much of the material for this seminar is taken from Arthur Spirling's course on "Text as Data" at New York University (`https://github.com/ArthurSpirling`). I am grateful to Prof. Spirling for sharing all of the code for his slides. I also draw on material from Joe Ornstein's class at the University of Georgia. I was also inspired by the syllabus of Alexandra Siegal at the University of Colorado.

Course Requirements

Eighty percent of success is showing up – Woody Allen

Grading will be based on:

- Two paper presentations worth 20% of final grade (10% each)
- Paper worth 50% of final grade
- Paper Presentation worth 20% of final grade
- Attending class and participating is worth 10% of your final grade

Paper Presentations

Starting in Week 2, there will be a paper presentation from a student on one of the papers listed below (or one of their choosing conditional on my approval). The presentation will last 15 minutes and will give an overview of what the paper tests and how it does this with a particular focus on the methods relevant for this class. You should probably have a few slides.

Papers to Choose From

- Gilchrist, D. S. and Sands, E. G. (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy*, 124(5):1339–1382
- Nekoei, A. and Sinn, F. (2021). Herstory: The rise of self-made women
- Hanlon, W. W., Heblich, S., Monte, F., and Schmitz, M. B. (2021). A penny for your thoughts
- Dittmar, J. and Seabold, S. (2019). New media and competition: printing and europe's transformation after gutenberg
- Chaney, E. (2020). Modern library holdings and historic city growth
- Becker, C. (2021). The rise of jim crow rhetoric in republican economic speeches
- Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294
- Ash, E., Chen, D. L., and Naidu, S. (2019). Ideas have consequences: The impact of law and economics on american justice. *Center for Law & Economics Working Paper Series*, 4
- Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636
- Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, 128(4):1325–1369
- 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):883–901

- Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The production of information in an online world. *The Review of Economic Studies*, 87(5):2126–2164
- Durante, R., Pinotti, P., and Tesei, A. (2019). The political legacy of entertainment tv. *American Economic Review*, 109(7):2497–2530
- Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340
- Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71
- Groseclose, T. and Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237
- Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202
- Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870
- Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–20
- Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168
- Michalopoulos, S. and Xue, M. M. (2021). Folklore. *The quarterly journal of economics*, 136(4):1993–2046

Research Paper

The paper should move your dissertation forward. The final work should be about 15-20 pages in length, double-spaced in Times Roman 12 pt. font.

I am requiring you to write the paper in LaTeX. This is the standard among most researchers in economics today and you might as well learn it now rather than later. You should be able to find many, many, tutorials online for getting started, but here are some I have used:

http://www.maths.tcd.ie/~dwilkins/LaTeXPrimer/

Here is another one...

https://www.tug.org/begin.html

And one more...

https://www.researchgate.net/publication/280050294_Template-based_introductory_guide_to_LaTeX_for_Economics_Instructional_Guide_Version_2

Research Paper Presentations

Twenty percent of your course grade is determined by your paper presentation.

I encourage the use of slides for your presentation, but try not to over-do it. If you have questions on what is over-doing it, please refer to this book:

Tufte, E. R. (1983). The visual display of. *Quantitative Information*

I will grade your presentations on "content" (how well you have framed and answered your research question) and "style" (how well you present the material).

Since you're writing the paper in LaTeX, you should probably also make your presentation slides using the LaTeXpresentation environment known as Beamer. Here are some sample slides:

https://www.dropbox.com/sh/hnccpxpzmqjn55d/AACvkPZ25DxR5hhGrURXR91Ea?dl=0

Here are some cooler/fancier slides:

https://www.kylebutts.com/open-source/templates/

Course Materials

The readings each week are in the Course Outline below. These are the books which will be referenced in the Outline as well as several books which I would like to bring to your attention:

- This is a new book on text as data that looks quite good. We will be relying on it as one of our primary readings.
  - 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
- We will also be referencing this great source for statistical learning techniques. You can download the book for free at this link.
  - James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *Statistical learning*. Springer
- A very good introduction to data science using R and the tidyverse is this book. Also free at this link.
  - Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc."
- We will also use material from this great source.

– Silge, J. and Robinson, D. (2017). *Text mining with R: A tidy approach.* " O'Reilly Media, Inc.". Online for free here: https://www.tidytextmining.com/

We will be using R, a statistical package. You can download and install R for free from here: https://cran.r-project.org/

To write and edit R code, you can use any software with which you are familiar and/or enjoy using. I suggest R Studio, which is free: https://www.rstudio.com/products/RStudio/

We will also often be using an R package designed by a team led by Prof. Ken Benoit, specifically engineered for social scientists working with text. The package is called quanteda. You can install it from the command line via: `install.packages("quanteda")`

You will also need a package specifically for data intake, called readtext. You can install it from the command line via: `install.packages("readtext")`

Some Important Dates

First Day of Classes: 23 January

Spring Break: 13 March to 19 March

Last Day of Classes: 1 May

Course Outline (subject to change)

Week 1: Introduction, Overview of Text as Data and Machine Learning, R Review

- Grimmer et al. (2022) Chapters 1-2
- Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as data. *Journal of Economic Literature*, 57(3):535–74
- Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106
- Nick Huntington-Kline videos on getting started in R: https://nickchk.com/videos.html#rstats

Week 2: Representing Text

- Grimmer et al. (2022) 3-5
- 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

Weeks 3-4: Similarity, Complexity, and Dictionary Methods

- Grimmer et al. (2022) 7, 15-16
- James et al. (2021) Chapter 5.2

Week 5: "Out of the Box" Sentiment Analysis

- Sentiment analysis with tidy data https://www.tidytextmining.com/sentiment.html

Weeks 6-7: Supervised Learning

- Grimmer et al. (2022) 17-20
- James et al. (2021) Chapters 2, 4, 8, and 9
- Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87
- Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725

Weeks 8-9: Unsupervised Learning

- Grimmer et al. (2022) 12-14
- James et al. (2021) Chapters 12.1, 12.2, and 12.4
- Topic Modeling https://www.tidytextmining.com/topicmodeling.html

Week 10: Word Embeddings

- Grimmer et al. (2022) 8

Week 11: Foundation Models

- Alexander, Scott. 2019. "GPT-2 As Step Toward General Intelligence." Slate Star Codex.
- Alexander, Scott. 2020. "The Obligatory GPT-3 Post." Slate Star Codex.
- Sargent, T. J. (2022). Sources of artificial intelligence
- Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*
- Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, 151(2):127–138

Week 12: Web Scraping, API's, and OCR

- Joe Ornstein's guide to webscraping.
- Joe Ornstein's guide to api's.
- Joe Ornstein's guide to optical character recognition.
- Melissa Dell's Layout Parser https://dell-research-harvard.github.io/resources/layout-parser

Week 13: Catch-up and Review

Week 14: Student Paper Presentations

References

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? the information content of internet stock message boards. *The Journal of finance*, 59(3):1259–1294.

Ash, E., Chen, D. L., and Naidu, S. (2019). Ideas have consequences: The impact of law and economics on american justice. *Center for Law & Economics Working Paper Series*, 4.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4):1593–1636.

Bandiera, O., Prat, A., Hansen, S., and Sadun, R. (2020). Ceo behavior and firm performance. *Journal of Political Economy*, 128(4):1325–1369.

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):883–901.

Becker, C. (2021). The rise of jim crow rhetoric in republican economic speeches.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020). The production of information in an online world. *The Review of Economic Studies*, 87(5):2126–2164.

Chaney, E. (2020). Modern library holdings and historic city growth.

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.

Dittmar, J. and Seabold, S. (2019). New media and competition: printing and europe's transformation after gutenberg.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Durante, R., Pinotti, P., and Tesei, A. (2019). The political legacy of entertainment tv. *American Economic Review*, 109(7):2497–2530.

Gentzkow, M., Kelly, B., and Taddy, M. (2019a). Text as data. *Journal of Economic Literature*, 57(3):535–74.

Gentzkow, M. and Shapiro, J. M. (2010). What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019b). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Gilchrist, D. S. and Sands, E. G. (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy*, 124(5):1339–1382.

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.

Groseclose, T. and Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, 120(4):1191–1237.

Hanlon, W. W., Heblich, S., Monte, F., and Schmitz, M. B. (2021). A penny for your

thoughts.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: a computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.

Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *Statistical learning*. Springer.

Kelly, B., Papanikolaou, D., Seru, A., and Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–20.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.

Manning, C. D. (2022). Human language understanding & reasoning. *Daedalus*, 151(2):127–138.

Michalopoulos, S. and Xue, M. M. (2021). Folklore. *The quarterly journal of economics*, 136(4):1993–2046.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nekoei, A. and Sinn, F. (2021). Herstory: The rise of self-made women.

Sargent, T. J. (2022). Sources of artificial intelligence.

Silge, J. and Robinson, D. (2017). *Text mining with R: A tidy approach*. " O'Reilly Media, Inc.".

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Tufte, E. R. (1983). The visual display of. *Quantitative Information*.

Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. " O'Reilly Media, Inc.".