Department of Political Science
University of Toronto

POL 2578 H1F
Computer-Assisted Textual Analysis

Syllabus

Winter 2023

Class Time: **Tuesdays, 10AM–12PM**
Class Location: **RW 109 (Ramsey Wright Computer Lab 109)**

Instructor: Ludovic Rheault
Email: ludovic.rheault@utoronto.ca
Office Hours: See Quercus page. Office Location: Sidney Smith 3005

Course Description

Social actors interact using language. As a result, testing social science theories usually requires analyzing, in one way or another, written language. Thankfully, recent advances in computational linguistics have considerably increased the reach of scholars interested in working with textual data. Moreover, swathes of digitized documents have been made available to researchers in recent years. This includes parliamentary records, committee proceedings, bills, laws, international treaties, news reports, social media discussions, blogs, websites, and so forth. How to process and analyze such large quantities of textual data meaningfully is the central focus of this course.

The course introduces students to the state of the art in the field of computer-assisted textual analysis. It covers the most widely used methods for the empirical analysis of textual data, from the preprocessing stages to the interpretation of findings. The course also includes an introduction to machine learning. By the end of this course, students will have gained expertise with an important branch of computational social science. They will also have developed skills with the Python programming language.

Course Format

The course takes place in person in the Ramsey Wright computer lab RW 109 (the building next to Sidney Smith). Students can use the computers available in the lab or bring their own laptop. A typical class combines an advanced lecture on statistical theory introducing new concepts, followed by interactive exercises. Materials to reproduce class examples will be available on Quercus.

Software

Class exercises and demonstrations will be done using the Python programming language. Python is the most popular programming language in the world, and is especially useful for the analysis of textual data. The course begins with an introduction to the language.

In-class examples will be provided from the Jupyter notebook, a user-friendly environment for interactive computing.

Python is freely available on all operating systems. For a new install, consider the Anaconda distribution of Python, which comes with several useful packages pre-installed. Python should be installed by default on Mac computers. Any version of Python 3.x should work fine, but versions 3.7 and above are preferable.

The required software is available in the RW 109 computer lab.

Requirements

This course may be of interest to graduate students using either qualitative or quantitative methods (or both). Although there are no formal requirements for the course, it will involve some advanced concepts in programming and statistics. A background in statistical analysis and/or computing would be useful, at the level of the sequence of courses POL 2504 and POL 2507. The pedagogical approach is tailored to students who may not have had an extended training in mathematics or computing as undergraduate students, as is often the case in the social sciences.

Marking Scheme

The course relies on three assignments (plus a participation mark). The first two assignments are problems sets. The last assignment is a critical evaluation of written work from the field of text as data.

Written Assignment #1	30%	Due February 14, 2023
Written Assignment #2	30%	Due March 28, 2023
Written Assignment #3	30%	Due at the end of term (April 11, 2023)
Participation	10 %	

Readings

The readings for this course comprise a collection of chapters from the following set of seminal texts in the field. The readings recommended for each class are helpful to supplement the lecture notes that will be made available to students. All of these books are accessible for free online, either from the authors' websites or electronically through the UofT Library.

- Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python.* O'Reilly Media.
    - An accessible introduction to natural language processing in Python. The book is available online for free.
- Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2009. *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.
    - A key reference that covers most of the topics discussed in this course, and more. Online versions are available.

- Jurafsky, Daniel and James H. Martin. 2020. *Speech and Language Processing*. 3rd Edition. New Jersey: Prentice Hall.
    - Another useful reference for exploring some of the topics in more depth. Some chapters are available online for free.
- Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.
    - An older reference that nonetheless covers key basic concepts for this course. The book is available electronically through the UofT Library.
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2nd Edition. Berlin: Springer.
    - A useful reference on the particular topic of machine learning. The book is available electronically through the UofT Library.
- Hovy, Dirk. 2020. Text Analysis in Python for Social Scientists: Discovery and Exploration. Cambridge: Cambridge Elements Series.
    - This new resource is a short book that covers many of the topics we study in this course. It is available electronically through the UofT Library.

Evaluations

The course uses two evaluation formats to help students develop different skills related to scientific research.

Problem Sets

The first two written assignments are problem sets designed to evaluate students' ability to put the methods learned into practice. They involve practicing various types of textual analysis using Python and answering short factual questions about the models and their interpretation.

There is no better way to improve one's skills than practice. Therefore, those exercises are not only useful as evaluations, but also as a way for students to gain concrete expertise with the subject-matter. Assignments are done individually. They are submitted on Quercus at the due date.

Critical Review

The last assignment consists of reading and discussing published work relying on text as data. The list of admissible papers will be posted on Quercus. The goal is to demonstrate that, at the end of the term, the graduate student is able to engage with the literature in a meaningful way, to understand the methods and judge their appropriateness, and to identify the strengths and limitations of published studies involving automated textual analysis.

This exercise will be valuable in preparing students to write their own papers using text as data.

Class Schedule: Summary

Date	Topic	Evaluation
January 10	Computers and language & introduction to Python	
January 17	Introduction to Python (continued)	
January 24	Statistics for textual data I	
January 31	Statistics for textual data II	
February 7	Concepts in computational linguistics	
February 14	Lexicons and dictionaries	Assignment 1 due
February 21	[Reading week - No Classes]	
February 28	Meaning and word embeddings	
March 7	Introduction to machine learning	
March 14	Supervised learning I	
March 21	Supervised learning II	
March 28	Unsupervised learning I	Assignment 2 due
April 4	Unsupervised learning II	
April 11	[Final assessment period]	Assignment 3 due

Note: Topics by date are for information only. The schedule above (and the detailed structure in the following pages) may by adjusted during the term due to unforeseen circumstances or to improve the pedagogical benefits to students.

Class Schedule: Detailed

Topic 1: Computers and Text

January 10: Computers and Language; Introduction to Python

1. Brief history of automated textual analysis.
2. Examples of recent applications.
3. Introduction to Python 3 (beginning).

January 17: Introduction to Python (Continued)

1. Introduction to Python 3 (continued).
2. Data types, lists and dictionaries.
3. Input/Output.
4. Functions and conditional statements.
5. Encoding text.
6. Processing textual data in Python.
7. Exercise: Parsing text in various formats (html, xml, pdf files).

Readings

- Hovy (2020), Ch. 1.
- Bird, Klein, and Loper (2009), Ch. 2–4.

Other Useful References

- Aggarwal and Zhai (2012*b*).
- Manning and Schütze (1999), Ch. 1.
- McKinney (2013), Ch. 1.
- Downey, Elkner, and Meyers (2002), Ch. 1–2.
- D'Orazio et al. (2014).
- Jockers (2014).
- Weiss, Indurkhya, and Zhang (2015).
- Krippendorff (2013), Ch. 4.
- Grimmer and Stewart (2013).
- Gentzkow, Kelly, and Taddy (2019).
- Benoit (2019).
- Watch a 45-minute introductory video on Python.

Topic 2: Statistics for Textual Data

January 24: Statistics for Textual Data I

1. Document retrieval and indexing.
2. Tokenization, sentence splitting.
3. Word counts and word distributions.
4. Vectorization.
5. Visualization techniques.

January 31: Statistics for Textual Data II

1. Term-frequency/inverse document frequency (tf–idf) weighting.
2. Word co-occurrences/collocations.
3. Comparing texts.
4. Statistical properties of texts.

Readings

- Manning, Raghavan, and Schütze (2009), Ch. 1–2.
- Hovy (2020), Ch. 2–4.
- Manning and Schütze (1999), Ch. 5–6.

Other Useful References

- Bird, Klein, and Loper (2009), Ch. 2–4.
- Jiang (2012).
- Nenkova and McKeown (2012).
- Python Online Documentation.

Examples of Applications

- Laver and Garry (2000).
- Laver, Benoit, and Garry (2003).
- Alfini and Chambers (2007).
- Lowe (2008).
- Slapin and Proksch (2008).
- Monroe, Colaresi, and Quinn (2008).
- Gentzkow and Shapiro (2010).
- Proksch and Slapin (2010).
- Black et al. (2011).
- Däubler et al. (2012).
- Acton and Potts (2014).
- Yu (2014).
- Spirling (2016).
- Blaxill and Beelen (2016).
- Benoit, Munger, and Spirling (2019).

Topic 3: Linguistics and Natural Language Processing

February 7: Concepts in computational linguistics

1. Overview of linguistic theory.
2. Unigrams, bi-grams and *n*-grams.
3. Part-of-speech tagging.
4. Stemming and lemmatization.
5. Grammar parsing.
6. Named entity recognition.

February 14: Lexicons and dictionaries

1. Creating and using word lexicons (dictionaries).
2. Summarizing text properties.
3. Political science applications: Word Scores and WordFish.

February 28: Meaning and word embeddings

1. Meaning representation and latent semantic analysis.
2. Word embeddings.
3. Word similarities and word relations.

Readings

- Bird, Klein, and Loper (2009), Ch. 5.
- Hovy (2020), Ch. 5.
- Jurafsky and Martin (2020), Ch. 20.

Other Useful References

- Manning, Raghavan, and Schütze (2009), Ch. 6.
- Miller et al. (1990).
- Turney and Pantel (2010).
- Mikolov et al. (2013).
- Manning et al. (2014).
- Landauer, Foltz, and Laham (1998).
- Python Online Documentation.

Examples of Applications

- Tausczik and Pennebaker (2010).
- Bollen, Mao, and Zeng (2011).
- Bollen, Mao, and Pepe (2011).
- Golder and Macy (2011).
- Michel et al. (2011).
- Young and Soroka (2012).
- Jensen et al. (2012).

- Coviello et al. (2014).
- Gentzkow, Shapiro, and Taddy (2016).
- Rheault et al. (2016).
- Vosoughi et al. (2018).
- Martin and McCrain (2019).
- Gennaro and Ash (2021).

Topic 4: Machine Learning

March 7: Introduction to Machine Learning

1. Machine learning and classification.
2. Annotating texts and intercoder reliability.
3. Development, training and testing.
4. An introductory example: sentiment analysis.

March 14: Supervised Learning I

1. Features and classes.
2. "Bag of words" approach.
3. Feature selection.
4. Naive Bayes classifiers.
5. Nearest Neighbor classifiers.
6. Multi-class problems.

March 21: Supervised Learning II

1. Evaluating classifiers.
2. Accuracy measures.
3. Ridge regression.
4. Support vector machines.
5. Applications in Python.

March 28: Unsupervised Learning I

1. Unsupervised learning.
2. Motivating example: topic classification.
3. Clustering analysis.
4. Principal component analysis.

April 4: Unsupervised Learning II

1. Latent Dirichlet Allocation (LDA).
2. Correlated and dynamic topic models.
3. Non-Negative Matrix Factorization.

Readings

- Hastie, Tibshirani, and Friedman (2009), Ch. 2, 6–7, 12.
- Hovy (2020), Ch. 6–7.

Other Useful References

- Manning, Raghavan, and Schütze (2009), Ch. 15.
- Shawe-Taylor and Cristianini (2000).
- Blei, Ng, and Jordan (2003).
- Blei and Lafferty (2006*a*).
- Blei and Lafferty (2006*b*).
- Blei (2012).
- Bird, Klein, and Loper (2009), Ch. 6.
- Hayes and Krippendorff (2007).
- He and Garcia (2009).
- Steyvers and Griffiths (2011).
- Aggarwal and Zhai (2012*a*).
- Richert and Coelho (2013).
- Lantz (2013).
- James et al. (2013).
- Raschka (2015).
- scikit-learn for Python: Online Documentation.

Examples of Applications

- Mosteller and Wallace (1964).
- Airoldi, Fienberg, and Skinner (2007).
- Yu, Kaufmann, and Diermeier (2008).
- Hopkins and King (2010).
- Grimmer (2010).
- Grimmer, Messing, and Westwood (2012).
- Diermeier et al. (2012).
- Hirst et al. (2014).
- Roberts et al. (2014).
- D'Orazio et al. (2014).
- Lucas et al. (2015).
- Harris (2015).
- Reich et al. (2015).
- Roberts, Stewart, and Airoldi (2016).
- Tingley (2017).
- Greene and Cross (2017).
- Peterson and Spirling (2018).
- Barberá et al. (2019).

References

Acton, Eric K., and Christopher Potts. 2014. "That Straight Talk: Sarah Palin and the Sociolinguistics of Demonstratives." *Journal of Sociolinguistics* 18(1): 3–31.

Aggarwal, Charu C. 2012. "Mining Text Streams." In *Mining Text Data*, ed. Charu C. Aggarwal, and ChengXiang Zhai. New York: Springer pp. 297–322.

Aggarwal, Charu C., and ChengXiang Zhai. 2012*a*. "A Survey of Text Classification Algorithms." In *Mining Text Data*, ed. Charu C. Aggarwal, and ChengXiang Zhai. New York: Springer pp. 163–222.

Aggarwal, Charu C., and ChengXiang Zhai. 2012*b*. "An Introduction to Text Mining." In *Mining Text Data*, ed. Charu C. Aggarwal, and ChengXiang Zhai. New York: Springer pp. 1–10.

Airoldi, Edoardo M., Stephen E. Fienberg, and Kiron K. Skinner. 2007. "Whose Ideas? Whose Words? Authorship of Ronald Reagan's Radio Addresses." *PS: Political Science and Politics* 40(3): 501–506.

Alfini, Naomi, and Robert Chambers. 2007. "Words Count: Taking a Count of the Changing Language of British Aid." *Development in Practice* 17(4/5): 492–504.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 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.

Bengio, Yoshua, Ian Goodfellow, and Aaron Courville. 2016. *Deep Learning*. Cambridge: MIT Press.

Benoit, Kenneth. 2019. "Text as data: An overview." In *SAGE Handbook of Research Methods in Political Science and International Relations*, ed. Luigi Curini, and Robert Franzese. London: pp. 461–497.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling. 2019. "Measuring and explaining political sophistication through textual complexity." *American Journal of Political Science* 63(2): 491–508.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. Sebastopol: O'Reilly Media.

Black, Ryan C., Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. 2011. "Emotions, Oral Arguments, and Supreme Court Decision Making." *The Journal of Politics* 73(2): 572–581.

Blaxill, Luke, and Kaspar Beelen. 2016. "A Feminized Language of Democracy? The Representation of Women at Westminster since 1945." *Twentieth Century British History*. doi: 10.1093/tcbh/hww028

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55(4): 77–84.

Blei, David M., and John D. Lafferty. 2006*a*. Correlated Topic Model. In *Neural Information Processing Systems*.

Blei, David M., and John D. Lafferty. 2006*b*. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning*.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3(Jan): 993–1022.

Bollen, Johan, Huina Mao, and Alberto Pepe. 2011. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. pp. 450–453.

Bollen, Johan, Huina Mao, and Xiaojun Zeng. 2011. "Twitter Mood Predicts the Stock Market." *Journal of Computational Science* 2(1): 1–8.

Coviello, Lorenzo, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. 2014. "Detecting Emotional Contagion in Massive Social Networks." *PLoS ONE* 9(3): e90315.

Däubler, Thomas, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2012. "Natural Sentences as Valid Units for Coded Political Texts." *British Journal of Political Science* 42: 937–951.

Denny, Matthew J, and Arthur Spirling. 2018. "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it." *Political Analysis* 26(2): 168–189.

Diermeier, Daniel, Jean-François Godbout, Bei Yu, and Stefan Kaufmann. 2012. "Language and Ideology in Congress." *British Journal of Political Science* 42(1): 31–55.

D'Orazio, Vito, Steven T. Landis, Glenn Palmer, and Philip Schrodt. 2014. "Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines." *Political Analysis* 22(2): 224–242.

Downey, Allen, Jeffrey Elkner, and Chris Meyers. 2002. *How to Think Like a Computer Scientist: Learning with Python*. Wellesley: Green Tea Press.

Enke, Benjamin. 2020. "Moral Values and Voting." *Journal of Political Economy* 128(10): 3679–3729.

Gennaro, Gloria and Elliott Ash. 2021. "Emotion and Reason in Political Language." Working Paper.

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78(1): 35–71.

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy. 2019. "Text as data." *Journal of Economic Literature* 57(3): 535–74.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2016. "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech." *NBER Working Paper* p. 22423.

Golder, Scott A., and Michael W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 333(6051): 1878–1881.

Greene, Derek, and James P. Cross. 2017. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *Political Analysis* 25(10): 77–94.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18(1): 1–35.

Grimmer, Justin, Solomon Messing, and Sean J Westwood. 2012. "How Words and Money Cultivate a Personal Vote: The Effect of Legislator Credit Claiming on Constituent Credit Allocation." *American Political Science Review* 106(4): 703–719.

Grimmer, Justin, and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3): 267–297.

Harris, J. Andrew. 2015. "What's in a Name? A Method for Extracting Information about Ethnicity from Names." *Political Analysis* 23(2): 212–224.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Berlin: Springer.

Hayes, Andrew F., and Klaus Krippendorff. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures* 1(1): 77–89.

He, Haibo, and Edwardo A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21(9): 1263–1284.

Hirst, Graeme, Yaroslav Riabinin, Jory Graham, Magali Boizot-Roche, and Colin Morris. 2014. "Text to Ideology or Text to Party Status?" In *From Text to Political Positions: Text Analysis across Disciplines*, ed. Bertie Kaal, Isa Maks, and Annemarie van Elfrinkhof. John Benjamins Publishing Company pp. 61–79.

Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1): 229–247.

Hovy, Dirk. 2020. *Text Analysis in Python for Social Scientists: Discovery and Exploration*. Cambridge: Cambridge Elements Series.

Hu, Xia, and Huan Liu. 2012. "Text Analytics in Social Media." In *Mining Text Data*, ed. Charu C. Aggarwal, and ChengXiang Zhai. New York: Springer pp. 385–414.

Huang, Leslie and Patrick O. Perry and Arthur Spirling. 2020. "A General Model of Author 'Style' with Application to the UK House of Commons, 1935–2018." *Political Analysis* 28(3): 412–434.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. 2012. "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech." *Brookings Papers on Economic Activity* Fall: 1–81.

Jiang, Jing. 2012. "Information Extraction From Text." In *Mining Text Data*, ed. Charu C. Aggarwal, and ChengXiang Zhai. New York: Springer pp. 11–42.

Jockers, Matthew L. 2014. *Text Analysis with R for Students of Literature*. New York: Springer.

Jurafsky, Daniel, and James H. Martin. 2020. *Speech and Language Processing*. 3rd Edition. New Jersey: Prentice Hall.

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology.* 3 ed. Thousand Oaks: Sage Publications.

Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. "An Introduction to Latent Semantic Analysis." *Discourse Processes* 25: 259–284.

Lantz, Brett. 2013. *Machine Learning with R.* Birmingham: Packt Publishing.

Laver, Michael, and John Garry. 2000. "Estimating Policy Positions from Political Texts." *American Journal of Political Science* 44(3): 619–634.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2): 311–331.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16(4): 356–371.

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(2): 254–277.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* pp. 55–60.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Martin, Gregory J, and Joshua McCrain. 2019. "Local news and national politics." *American Political Science Review* 113(2): 372–384.

McKinney, Wes. 2013. *Python for Data Analysis.* Sebastopol: O'Reilly Media.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 331(6014): 176–182.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of Workshop at ICLR.*

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. "Introduction to WordNet: An On-line Lexical Database." *International Journal of Lexicography* 3(4): 235–244.

Mitchell, Ryan. 2015. *Web Scraping with Python.* Sebastopol: O'Reilly Media.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16(4): 372–403.

Mosteller, Frederick, and David Lee Wallace. 1964. *Inference and Disputed Authorship: The Federalist.* Boston: Addison-Wesley.

Munzert, Simon, Christian Rubba, Peter Meissner, and Dominic Nyhuis. 2015. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining.* Chichester: John Wiley & Sons.

Nenkova, Ani, and Kathleen McKeown. 2012. "A Survey of Text Summarization Techniques." In *Mining Text Data*, ed. Charu C. Aggarwal, and ChengXiang Zhai. New York: Springer pp. 43–76.

Peterson, Andrew, and Arthur Spirling. 2018. "Classification accuracy as a substantive quantity of interest: Measuring polarization in westminster systems." *Political Analysis* 26(1): 120–128.

Porter, M. F. 1980. "An Algorithm for Suffix Stripping." *Program* 14(3): 130–137.

Proksch, Sven-Oliver, and Jonathan B. Slapin. 2010. "Position Taking in European Parliament Speeches." *British Journal of Political Science* 40(3): 587–611.

Proksch, Sven-Oliver and Wratil, Christopher and Wäckerle, Jens. 2019. "Testing the Validity of Automatic Speech Recognition for Political Text Analysis." *Political Analysis* 27(3): 339–359.

Raschka, Sebastian. 2015. *Python Machine Learning.* Birmingham: Packt Publishing.

Reich, Justin, Dustin Tingley, Jetson Leder-Luis, Margaret E. Roberts, and Brandon M. Stewart. 2015. "Computer-Assisted Reading and Discovery for Student-Generated Text in Massive Open Online Courses." *Journal of Learning Analytics* 2(1): 156–184.

Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLoS ONE* 11 (12): e0168843.

Rheault, Ludovic, Erica Rayment, and Andreea Musulan. 2019. "Politicians in the line of fire: Incivility and the treatment of women on social media." *Research & Politics* 6(1): 2053168018816228.

Richert, Willi, and Luis Pedro Coelho. 2013. *Building Machine Learning Systems with Python.* Birmingham: Packt Publishing.

Roberts, Margaret E., Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association.* Forthcoming.

Roberts, Margaret E, Brandon M Stewart, and Richard A Nielsen. 2020. "Adjusting for confounding with text matching." *American Journal of Political Science* 64(4): 887–903.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4): 1064–1082.

Shawe-Taylor, John, and Nello Cristianini. 2000. *Support Vector Machines.* Cambridge: Cambridge University Press.

Slapin, Jonathan B., and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3): 705–722.

Spirling, Arthur. 2016. "Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915." *Journal of Politics* 78(1): 120–136.

Steyvers, Mark, and Tom Griffiths. 2011. *Handbook of Latent Semantic Analysis.* New York: Routledge chapter Probabilistic Topic Models, pp. 427–448.

Tausczik, Yla R. and James W. Pennebaker. 2010. "The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods." *Journal of Language and Social Psychology* 29(1): 24-54.

Tingley, Dustin. 2017. "Rising Power on the Mind." *International Organization* 71 (S1): S165-S188.

Turney, Peter D., and Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* 37: 141–188.

Vosoughi, Soroush, Deb Roy and Sinan Aral. 2018. "The Spread of True and False News Online." *Science* 359 (6380): 1146–1151.

Weiss, Sholom M., Nitin Indurkhya, and Tong Zhang. 2015. *Fundamentals of Predictive Text Mining.* 2 ed. London: Springer-Verlag.

Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29: 205–231.

Yu, Bei. 2014. "Language and Gender in Congressional Speech." *Literary and Linguistic Computing* 29(1): 118–32.

Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. "Classifying Party Affiliation from Political Speech." *Journal of Information Technology and Politics* 5(1): 33–48.

Zipf, George Kingsley. 1932. *Selected Studies of the Principle of Relative Frequency in Language.* Cambridge: Harvard University Press.