

Analysis of Text-Analysis Syllabi: Building a Text-Analysis Syllabus Using Scaling

Nadjim Fréchet, *Université de Montréal*

Justin Savoie, *University of Toronto*

Yannick Dufresne, *Université Laval*

ABSTRACT

In the last decade, text-analytic methods have become a fundamental element of a political researcher's toolkit. Today, text analysis is taught in most major universities; many have entire courses dedicated to the topic. This article offers a systematic review of 45 syllabi of text-analysis courses around the world. From these syllabi, we extracted data that allowed us to rank canonical sources and discuss the variety of software used in teaching. Furthermore, we argue that our empirical method for building a text-analysis syllabus could easily be extended to syllabi for other courses. For instance, scholars can use our technique to introduce their graduate students to the field of systematic reviews while improving the quality of their syllabi.


The latest methodological developments in quantitative text analysis have made its methods accessible to both political science researchers and students. Software that is both free and convenient has provided promising new research opportunities and facilitated the analysis of large corpuses of text (Grimmer and Stewart 2013; Wilkerson and Casas 2017). The attractiveness of text-analytic methods has led many major universities to introduce them into their regular quantitative-research courses or to create courses entirely dedicated to text analysis. In addition to being in high demand on the job market, quantitative skills in text analysis and natural language processing recently enabled large-scale substantive breakthroughs in the social sciences. However, for both early-career professors and some political science departments, the task of building a text-analysis syllabus from scratch can be daunting. Many syllabi are available online and there are numerous textbooks that can be used to teach particular courses. How do we make sense of all of the available methods and software? What are the relevant academic papers and books that should be included in a text-analysis syllabus to properly teach its methods to graduate students?

This article analyzes 45 graduate political science course syllabi across the most highly ranked universities worldwide. First, it aims to show how the systematic collection of syllabi can build a syllabus. It develops and discusses a method to organize approximately 1,000 unique sources into a database and provides an index that identifies the most important sources, based on various indicators of relevance. Second, using the extracted data, it examines how text-as-data is taught to political science graduate students and offers advice for the creation of a text-analysis syllabus. Third, the article discusses important tradeoffs involved in the choice of software, especially the choice between R and Python. It concludes with a general discussion on the pros and cons of the method.

A SYSTEMATIC METHOD TO SYSTEMATICALLY KNOW WHAT TO TEACH

There is little empirical research on effective syllabi building. For instance, although there are abundant articles describing and implementing different text-analytic methods (Grimmer and Stewart 2013; Wilkerson and Casas 2017), they do not provide guidelines on how these methods may be taught in class. For political science instructors, it can be difficult to know where to start, especially if there is no precedent for teaching text analysis in their department. The fact is that new methods often are developed and published as articles in various venues. Hence, it can be tedious to create a pedagogical curriculum that is in line with the most recent and pertinent material. Are the canonical and the newly produced academic sources relevant enough to be included

Nadjim Fréchet  is a PhD student in political science at the Université de Montréal. He can be reached at nadjim@clessn.com.

Justin Savoie  is a PhD student in political science at the University of Toronto. He can be reached at justin.savoie@mail.utoronto.ca.

Yannick Dufresne is assistant professor in political science at Université Laval. He can be reached at yannick.dufresne@pol.ulaval.ca.

in a text-analysis syllabus? Which software should be used? By quantitatively analyzing data from the text-analysis syllabi of highly ranked universities across the world, this article provides a point of departure for instructors teaching such a course.

A Systematic Data-Gathering Method

The syllabi-collection procedure described in this article is based on two criteria: (1) the systematic syllabi selection and collection procedure are to be meaningfully selective; and (2) it must limit potential selection biases. Indeed, the selection process aims to gather data from syllabi from the best-ranked universities, mainly because of their high academic reputation and their unquestionable influence in the academic world.¹ Because we searched precisely for a specific type of document—text-analysis

To extract data from the syllabi, all of the articles and books listed were compiled into a BIBTEX file³; 992 sources were extracted from the syllabi collected. From these 992 sources, 110 books and articles were identified by the index as canonical.

DESCRIPTIVE ANALYSIS OF THE DATA

Figures 1 and 2 show the importance scores of the sources selected by the index. Figure 1 presents substantive texts; figure 2 presents texts about theory and software. Figure 2 shows that the most significant publication identified by the index is Grimmer and Stewart's (2013) *Text as Data*, listed in 78% of collected syllabi. Grimmer and Stewart's article is often used as an introduction to the field and therefore can be an excellent overview at the beginning of a course semester.

Grimmer and Stewart's article is often used as an introduction to the field and therefore can be an excellent overview at the beginning of a course semester.

syllabi—it was unnecessary to randomly select among the population of syllabi at this point (George and Bennett 2005, 173–77). However, there are no selection biases in the source collection because the articles and books quoted on the syllabi are unknown a priori. The canonical source-ranking index contributes also to the source-collection process, limiting potential biases caused by the self-selection of syllabi or cases (Geddes 1990; King, Keohane, and Verba 1994).² Ultimately, 45 syllabi were collected. Following our main selection criteria, only syllabi for courses in or workshops covering text analysis at these universities were selected.

Source-Ranking Index

The source-ranking index introduced in this article is based on a deductive approach consistent with what is assumed to be teaching objectives of text-analysis professors in political science. A deductive approach is appropriate given a lack of existing empirical studies on how text-analysis methods are taught in major universities (George and Bennett 2005, 111–14). The ranking index overweights recently published texts assuming that they represent more current advances in text analysis. However, because the index also overweights highly cited texts, older works are not penalized if they are well cited. The index also gives greater weight to texts that appear on many syllabi so that older and less-well-cited sources that nevertheless are widely used for pedagogical reasons are not penalized.

The ranking index considers the number of times a book or an article appears on the syllabi ($nSyllabi$). We assume that the number of times a publication is listed in the syllabi indicates the importance that political science professors collectively accord to it. The number of times a source has been cited also is considered by the index ($nCites$). The index takes into account a source's number of citations because it weights the source based on its overall contribution to science. The last variable that the ranking index considers is the text's publication year ($Year$). Considering these different elements, the index-ranking formula is as follows:

$$Rank_{(xi)} = nSyllabi_{(xi)} + nCites_{(xi)} + Year_{(xi)}$$

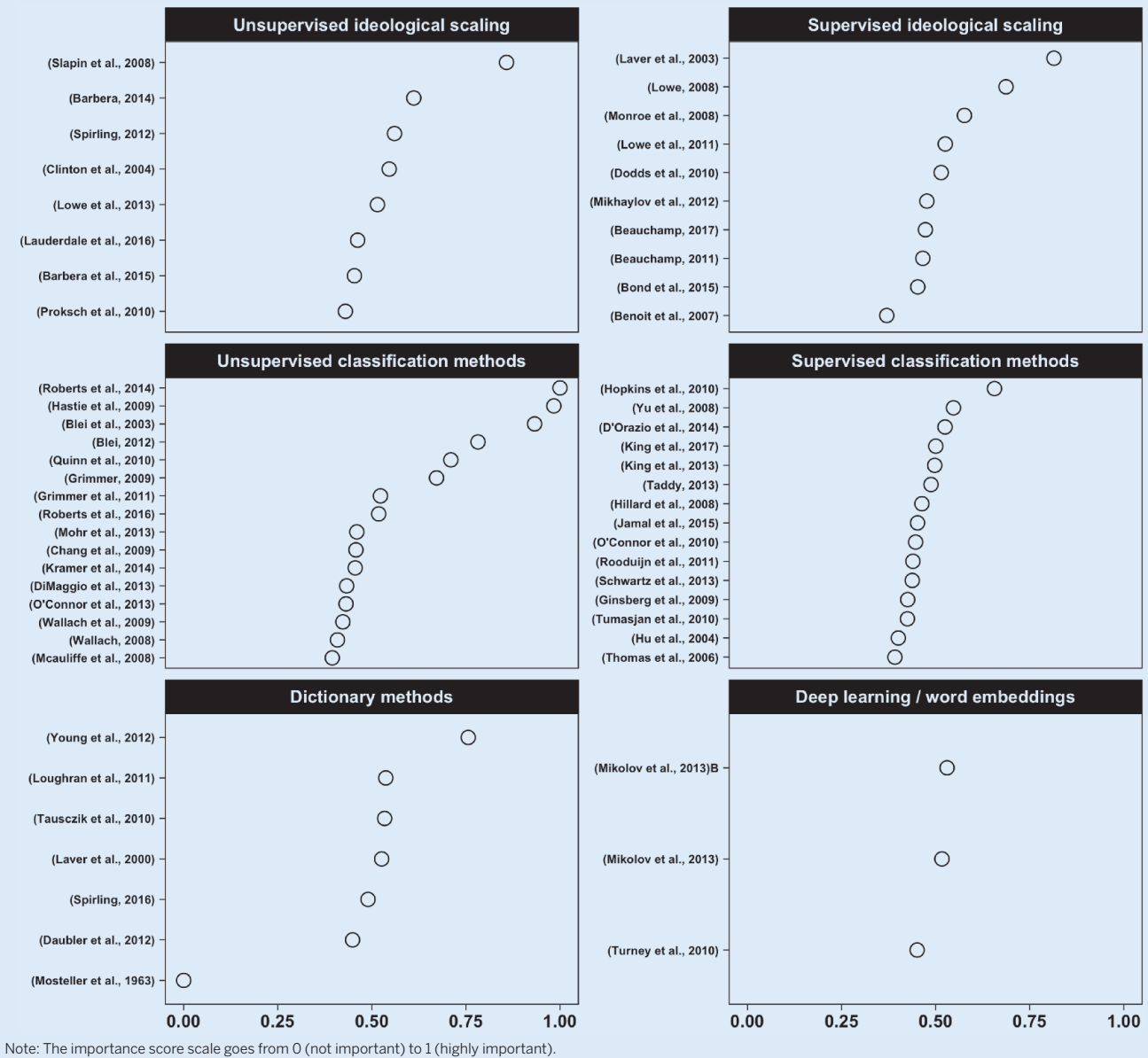
Note: All elements of the equation are standardized and rescaled to range from 0 to 1.

Figure 1 shows six substantive categories of texts found across syllabi.⁴ One striking factor is that classification and scaling are central to the teaching of text analysis. Classification methods originate from computer science and engineering. In these fields, the emphasis often is on “state-of-the-art” machine-learning models for classification and their refinement. By contrast, ideological scaling originates from and is more prevalent in political science. Since Poole and Rosenthal's (1987) work on spatial scaling, political scientists have been developing methods to scale the ideological positions of legislators. More recently, this has been accomplished with text data instead of roll-call vote data. In the syllabi, most scaling texts were written by political scientists. Texts about classification have more diverse origins, including natural and applied sciences.

Together, scaling and classification form the bulk of the text-analysis toolkit. In classes, they typically are introduced mid-semester and discussed for several weeks. The two classic texts for scaling are (1) the use of Wordscores by Laver, Benoit, and Garry (2003) for supervised scaling based on reference texts; and (2) the work of Slapin and Proksch (2008) and Proksch and Slapin (2010) documenting the unsupervised Wordfish approach. They are “classics” in the sense that they represent two approaches central to scaling, are most prevalent across syllabi, and are used widely in research. Lauderdale and Herzog's (2016) Wordshoal is an extension to Wordfish that uses a Bayesian approach to model multiple topics in a single corpus. All of the models proposed in these texts can be estimated using Benoit's (2018) *quanteda* package for R. Our syllabi analysis reveals that more than one week typically is allocated to teaching scaling. This is logical given the influence of spatial models of politics in political science, which often rely on scaling. Recently, text-based scaling using the previously described methods has been implemented using text extracted from audio and video. Other texts included in the two scaling panels of figure 1 include Lowe's (2008) “Understanding Wordscores” and the multiple implementations of scaling methods, such as Barberá's (2015) work on social media.

Regarding classification, we emphasize the difference between supervised and unsupervised approaches. Texts categorized in the unsupervised classification panel of figure 1 typically rely on one or another variant of topic modeling. We note the value of Roberts

Figure 1
Method-Driven Sources



et al.'s (2014) article that introduces the *stm* R package for open-ended survey responses. An advantage of this R package is that it was designed by political scientists with social-scientific applications in mind. The *stm* approach extends Latent Dirichlet topic modeling by allowing for the use of covariates (i.e., metadata) in topic estimation. Both approaches identify (or provide unsupervised classification of) latent topics using a statistical-mixture model. From a statistical point of view, supervised classification is simpler. Most of the work is upstream and involves the creation of a training set (Grimmer and Stewart 2013, 275). When the training set is sufficiently large and reliable, any appropriate classifier can be used, even simple logistic regression. However, this classifier inevitably misclassifies some texts. Hopkins and King (2010) introduced checks to optimize the population-level validity of the results rather than the accuracy of the classifier at the document level. In terms of syllabus building, it means discussing

notions of accuracy, recall, precision, and imbalance in predictive accuracy among courses.

The two remaining categories in figure 1 are dictionary-based methods and word embeddings. Dictionary methods were used mostly before the emergence of supervised regression-based classification. They can be problematic to use outside of their own domains; for instance, a medical dictionary likely should not be used in political science. However, even within political science, it can be contentious to use a dictionary developed in one country to study another country. For example, words relating to "asylum seeking" tend to be very different (e.g., boats versus forest roads) in Australia and in Canada (Grimmer and Stewart 2013, 275). Nevertheless, many syllabi include the dictionary method because it can serve as a starting point to learn more sophisticated methods. Dictionary methods also are relatively easy to implement using *quanteda*'s "dictionary" function. The highest ranked source (0.76 on figure 1) regarding dictionary

methods is Young and Soroka's (2012) article, which is included in approximately one third of the syllabi collected. It introduces the Lexicoder Sentiment Dictionary, which Young and Soroka argue produces sentiment classifications closer to those made by humans.

The remaining categories analyzed group-word embeddings and artificial neural networks. Word embeddings are a promising tool to analyze large corpuses and identify patterns in speech. Given that they are mathematically dense, the articles on this topic by Mikolov et al. (2013a; 2013b) are somewhat arcane for political science graduate students. The users of these methods are currently better served in Python than in R. In the near future, we can expect these methods to be implemented in the R language.⁵

and quanteda) that now exist allowing the user to perform both data wrangling and textual analysis. R is widely used in political science, and most students have been introduced to it by the time they take an upper-level undergraduate or graduate course in text analysis.

Python also has a multitude of text-analytic functionalities. Pandas for data structures and analysis, in combination with NumPy data-structure manipulations, is a near equivalent to tidyverse. Scikit-learn, NLTK, spaCy, and Gensim are extensions for statistical and predictive modeling, text analysis, and natural language processing. Also, Python is used widely in industry. It makes sense for students pursuing a professional master's degree (e.g., a computational social science MA program) to learn

As a general rule, the more we move to large-scale or production applications, the more that Python can be seen as having an advantage.

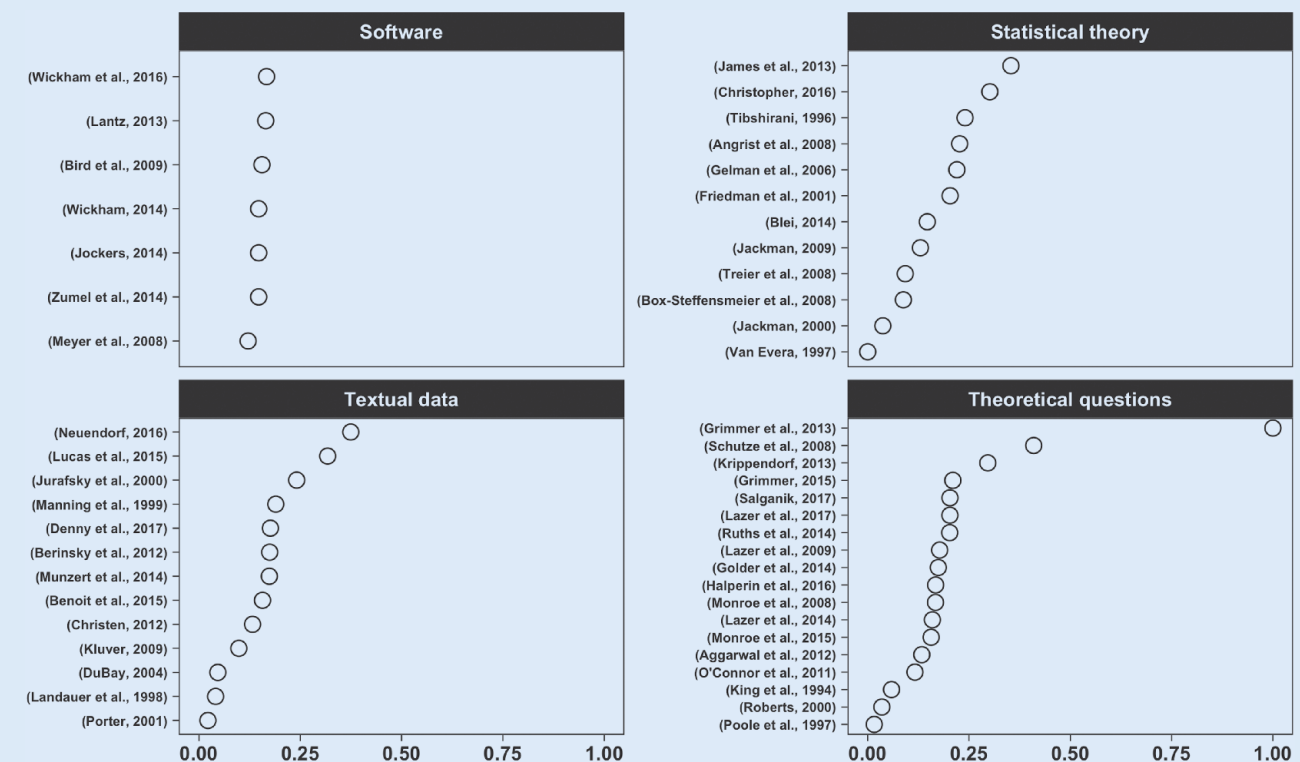
CHOOSING BETWEEN R AND PYTHON

Another concern when developing a syllabus is the choice between R and Python. Among the 45 syllabi studied, 38 mentioned the use of R, which suggests that it is the default choice in political science; 14 mentioned Python (some both R and Python); and 15 mentioned another software (often Stata). In theory, both Python and R can be used for most analyses. In practice, the choice usually depends on the preferred language of the instructor. However, it is necessary to understand the tradeoffs among these choices. R is useful because of the variety of text-analysis packages (e.g., tidyverse, tidytext, stm,

Python in addition to R. A text-analysis course using Python typically covers the language's basic features at McKinney's (2012) level before moving to NLTK and Gensim. In addition to being used in industry, an advantage of Python is that modern word-embedding techniques (e.g., Word2Vec and Doc2Vec) are not easily available in R. Some analyses can be accomplished in R's text2vec package, and Python can be called from R using the reticulate package. As a general rule, the more we move to large-scale or production applications, the more that Python can be seen as having an advantage. Overall, most statistical analyses

Figure 2

Theory and Software-Driven Sources



Note: The importance score scale goes from 0 (not important) to 1 (highly important).

of text can be done in both languages. The choice of software depends on the instructor's skills and the course objectives.

To continue the previous discussion about the pros and cons of R and Python, we review the "software" facet in figure 2. Little variation is observable in the index score for different texts related to software. Wickham and Golemund (2016) is a good introduction to more advanced programming notions in R. Other resources listed in figure 2 include textbooks for machine learning and natural language processing that focus on software implementation. The other three facets of figure 2 list texts classified as discussing textual data, theoretical questions (often broad reflections about

the field evolves. Updating the database of text-analytic resources will allow instructors to create the most current syllabus.

Why be limited to text analysis? The method presented in this article can be applied to other political science subfields. Syllabus building can be arcane, especially for junior instructors. Uniformization of the practice provides a tool that can shed light on what is important to teach, what is taught elsewhere, and what is less central. Therefore, the method presented herein also can be seen as an introduction to the world of systematic reviews.

Perhaps unsurprisingly, a possible critique of such a data-

Syllabus building can be arcane, especially for junior instructors. Uniformization of the practice provides a tool that can shed light on what is important to teach, what is taught elsewhere, and what is less central.

the field), and statistics. General resources on textual data often are covered at the beginning of the semester, including discussion of text acquisition, semantics, and parts of speech. Finally, our analysis of syllabi shows that the extent to which traditional (i.e., advanced) statistics are discussed in a text-analysis course is largely decided by the instructor. For example, Wordfish is a simple ideal point model on word counts. It can be useful for students to hand-code these models for a more in-depth understanding. Our analysis also shows that the extent to which advanced statistics are taught in a text-analysis course depends on the types of other courses offered in the graduate program.

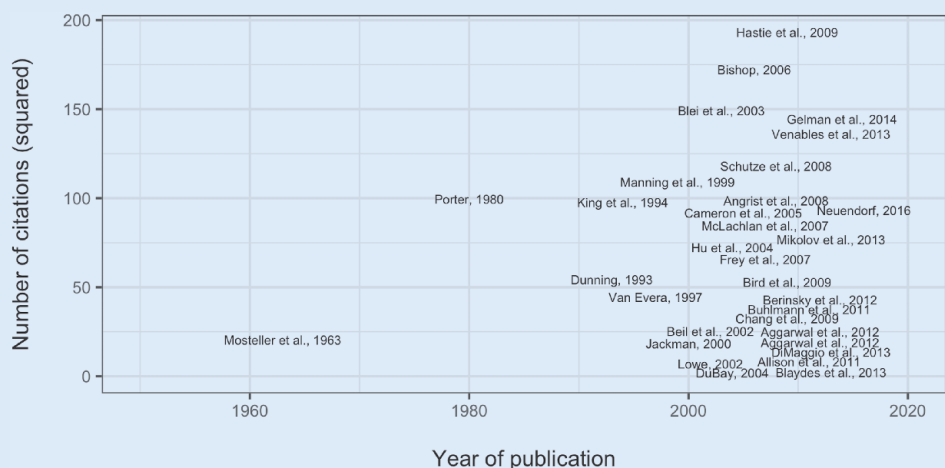
GOING FORWARD: UPDATING THE DATA AND APPLYING THE METHOD TO OTHER TOPICS

This article presents a data-driven method to build a syllabus in political science. It applies this method to the field of text analysis to help new instructors build their own course syllabus. This conclusion presents two possible extensions to the method and addresses one conceivable criticism. The extensions concern the need for periodic updating and the method's application to other political science subfields. The criticism concerns the limitations of such a data-driven approach, especially in relation to the ambiguous nature of syllabi building.

Figure 3 maps the relationship between the year of publication and the number of citations. The key takeaway is the field's novelty. Almost all publications extracted from the syllabi were published after 2000; several were published after 2010. Certainly, the field of textual analysis and natural language processing will continue to evolve. The idea is that by making this article's data freely available, it will be possible to update it as

driven syllabus-creation method is that it leaves little room for the instructor's experience and the ambiguous nature of syllabus building. For instance, Arthur Spirling's undergraduate course at New York University theoretically is less concerned with the technical aspects of text analysis. It is suggested in his syllabus that interested students take courses on web-scraping and natural language processing in other departments that offer these and many other related courses. In contrast, Sven-Oliver Proksch's graduate course at the University of Cologne focuses more on textual data-collection methods such as web scraping. One syllabus covered variational inference, which is useful for testing Bayesian models on large datasets; other syllabi covered advanced use of regular expressions. Whether ethical implications of the use of big data is covered is an important question likely unanswered with a data-driven analysis. Teaching sequences discussed previously also depend on the prior level of students, course objectives, and other courses available. In no case should the data-driven syllabus-creation method entirely replace the judgment of instructors.

Figure 3
Publication Year of the Most-Cited Sources



Preferably, it will be used to complement their knowledge and allow them to present students with the most current material.

ACKNOWLEDGMENTS

The authors thank Marc-Antoine Rancourt for his help on data collection and the members of *Université Laval's Chaire de leadership en enseignement des sciences sociales numériques* for their feedback. ■

NOTES

1. The syllabi are selected from the top 50 universities ranked by the *QS World University Ranking* list in the field of political science. Syllabi covering text analysis were then extracted from their websites.
2. Replication material available at [GitHub.com/justinsavoie/text_syllabi_analysis](https://github.com/justinsavoie/text_syllabi_analysis).
3. More precisely, their Google Scholar BIBTEX citations were compiled on the file. From the BIBTEX file, a database was created using JabRef.
4. The 10 categories were inductively developed based on theory and the author's best knowledge of the various subfields of text analysis. Each source then was classified independently in the 10 categories by two of the coauthors. The two coauthors discussed and agreed on the category of each conflicting source.
5. For a history of word embeddings in relation to previous methods, see Rheault and Cochrane (forthcoming).

REFERENCES

- Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23 (1): 76–91.
- Benoit, Kenneth, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. "quanteda: An R Package for the Quantitative Analysis of Textual Data." *Journal of Open Source Software* 30 (3): 774.
- Geddes, Barbara. 1990. "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics." *Political Analysis* 2: 131–50.
- George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: MIT Press.
- 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): 267–97.
- Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–47.
- King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
- Lauderdale, Benjamin E., and Alexander Herzog. 2016. "Measuring Political Positions from Legislative Speech." *Political Analysis* 24 (3): 374–94.
- Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2): 311–31.
- Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16 (4): 356–71.
- McKinney, Wes. 2012. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. Boston: O'Reilly Media, Inc.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. "Efficient Estimation of Word Representations in Vector Space." Available at arXiv preprint arXiv:1301.3781.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. "Distributed Representations of Words and Phrases and Their Compositionality." San Diego, CA: Advances in Neural Information Processing Systems, 3111–19.
- Poole, Keith T., and Howard Rosenthal. 1987. "Analysis of Congressional Coalition Patterns: A Unidimensional Spatial Model." *Legislative Studies Quarterly* 12 (1): 55–75.
- Proksch, Sven-Oliver, and Jonathan B. Slapin. 2010. "Position Taking in European Parliament Speeches." *British Journal of Political Science* 40 (3): 587–611.
- Rheault, Ludovic, and Christopher Cochrane. Forthcoming. "Word Embeddings for the Estimation of Ideological Placement in Parliamentary Corpora." *Political Analysis*.
- 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): 1064–82.
- 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): 705–22.
- Wickham, Hadley, and Garrett Grolemond. 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Boston: O'Reilly Media, Inc.
- Wilkerson, John, and Andreu Casas. 2017. "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges." *Annual Review of Political Science* 20: 529–44.
- Young, Lori, and Stuart Soroka. 2012. "Affective News: The Automated Coding of Sentiment in Political Texts." *Political Communication* 29 (2): 205–31.