DHQ: Digital Humanities Quarterly
Preview
2023
Volume 17 Number 4

Rich Veins of Ore: A Review of Gabe Ignatow and Rada Mihalcea's *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*

Charlie Harper <crh92_at_case_dot_edu>, Case Western Reserve University

Abstract

An Introduction to Text Mining: Research Design, Data Collection, and Analysis by Gabe Ignatow and Rada Mihalcea is most effective in balancing clear explanations of analytical techniques with sufficient technical details to engage a cross-disciplinary audience. Its process-oriented approach can broaden interest and build greater proficiency in applied text-mining.

In *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*, Ignatow and Mihalcea bring together their distinct areas of expertise to produce a collection of lightly-coupled chapters on digital approaches to text in the social sciences. The collaboration of Ignatow, a sociologist at the University of North Texas, and Mihalcea, a computer scientist at the University of Michigan, is a choice example of the interdisciplinary trends that are foundational to the advancement of the digital humanities and their work addresses the growing need for collaboratively written, practical texts for social sciences and humanities research. In particular, their work is intended to inform the research processes of undergraduate or early graduate students "who want to do research using online tools and data sets" [Ignatow and Mihalcea 2017, xix]. Although, in practice, the tools and datasets to which they are referring are textually-based, the work engages with more widely relevant topics that are applicable to all forms of data, such as ethics, project design, and research dissemination. 1

The book's chapters are succinctly written and, at a typical length of ten to twenty pages, they are easily digestible as standalone readings or supplementary course assignments. Enhancing its purpose as a tool for students, each chapter begins with a set of learning objectives and concludes with a summary, key terms, highlights of main ideas, and questions for review and discussion. Occasionally, these are supplemented by lists of further readings and the text is interspersed with "Spotlight on the Research" info boxes, which summarize a relevant piece of published scholarship. 2

The structure of the work (and much of the content) takes inspiration from the authors' earlier publication, *Text Mining: A Guidebook for the Social Sciences* [Ignatow and Mihalcea 2016]. To address their intended student audience, the authors nicely organize the text into a six-part structure that parallels the research process. Beginning with the first four chapters, which constitute "Part I: Foundations", the authors establish an engaging but appropriately academic tone, and they do not assume prior knowledge of research projects or text analysis. Although, in Chapter 1, they rather quickly dive into theoretical approaches where greater exposition on the applications of text analysis and its research value may have been more appropriate [Ignatow and Mihalcea 2017, 6–12], the early emphasis on theory is notable, since it is all too easy to apply digital methods haphazardly. Their theoretical foundation is further solidified in Chapter 4 with the discussion of philosophical bases for text analysis, such as constructionism and critical realism [Ignatow and Mihalcea 2017, 43–45], and an overview about drawing appropriate inferences from data [Ignatow and Mihalcea 2017, 47–53]. Perhaps most vital for students who may be new to digital research, Chapter 3 is wholly devoted to ethics and it presents essential, high-level takeaways on privacy, informed consent, and the role of the institutional review board in research. 3

In “Part II: Research Design and Basic Tools” (Chapter 5 and 6), the authors attempt to shepherd the reader from foundational theory and concepts into the more practical realm. The decision-making process that goes into research design and issues of data gathering are reviewed in Chapter 5 [Ignatow and Mihalcea 2017, 60–64] and the authors, then, cover data collection using web scraping Chapter 6. One thorny aspect of the text shows through in this section: namely, the inclusion of what I would argue are tangential details and overly specific explanations. While this may be the result of the authors wrestling with their need to address a complex and inter-disciplinary audience, it can divert attention from the main purpose of the work. For example, while the discussion of qualitative, quantitative, and mixed-methods research is relevant [Ignatow and Mihalcea 2017, 64–66], delving into sampling methods, such as stratified and snowball sampling [Ignatow and Mihalcea 2017, 69–71], feels excessive. Additionally, the devotion of an entire chapter to one method for data collection, rather than more broadly addressing best practices for data acquisition and organization, is superfluous. Certainly, web scraping is and will remain important, but it is not the only or even arguably the primary method for new researchers to acquire textual data. This chapter, too, contains some oddities that perhaps further show a struggle with establishing the needs of their audience. For example, the chapter begins by casually mentioning “programming environments like Python or R” [Ignatow and Mihalcea 2017, 75] before devoting a page to explaining what the “web” is [Ignatow and Mihalcea 2017, 76]. Unfortunately, on the whole, the section feels like a missed opportunity to instill solid data collection and organization practices in new researchers. In a pedagogical setting, this section should be supplemented with more thorough works on data management best practices (e.g. [Briney 2015]; [Corti et al. 2019]). 4

“Part III: Text Mining Fundamentals” (Chapters 7-9) is quite distinct from its precursor. Here, the authors constructively lay out the core resources and processes of text analysis, and within the course of my own work, this is the section to which I most frequently refer students and researchers who are new to text analysis. The coverage of lexical resources in Chapter 7 is pragmatic; although constantly expanding, the major resources [Ignatow and Mihalcea 2017, 86–96], such as LIWC (Linguistic Inquiry and Word Count) and WordNet, likely to remain important for some time are mentioned and indicate rich ores of material for text mining. The summaries of each resource are accompanied by effective illustrations and tables, which make clear to the reader each resource’s composition and merit. Chapter 8’s coverage of text processing is detailed, but still approachable for a student. It is, also, here that one first glimpses some of the practical capabilities of text analysis, such as frequency counts, part-of-speech tagging, and named entity recognition [Ignatow and Mihalcea 2017, 107–112]. Those are likely to inspire the reader to begin considering more deeply the applications of text analysis within their own research projects. 5

The authors do well capitalizing on Chapter 8’s productiveness and invigorating the reader to pursue deeper research applications in Parts IV and V, which respectively cover “Text Analysis Methods from the Humanities and Social Sciences” (Chapters 10-12) and “Text Mining Methods from Computer Science” (Chapters 13-16). The parts scale up from the lexical unit and cover conceptually higher techniques such as qualitative and mixed-methods narrative analysis [Ignatow and Mihalcea 2017, 139–142], Naive Bayes text classification [Ignatow and Mihalcea 2017, 180–181], and topic modeling with latent Dirichlet allocation and latent semantic analysis [Ignatow and Mihalcea 2017, 208–212]. They are described well, including clear examples of when and how to use each that are accompanied by ample in-text citations. The text truly hits its stride in these two parts and the authors balance clear explanations with sufficient technical details that remain appropriate for a cross-disciplinary audience. For readers less interested in social sciences theory and basic research practices, and more ready to dig into the meat of text analysis, Parts III to V can be read on their own. 6

The work concludes with a single chapter, “Writing and Reporting Your Research,” that constitutes Part VI. Although the topic of publication and dissemination is an appropriate coda for the structure of the work, which again parallels the research process, the chapter feels somewhat cursory and forced. Approximately seven of the chapter’s seventeen pages are devoted to a list of undergraduate and social sciences journals [Ignatow and Mihalcea 2017, 233–240], of which some are now defunct (e.g. Colorado State’s *Journal of Undergraduate Research and Scholarly Excellence*) and others appear to have ceased publication long before this work was released (e.g. *Lethbridge Undergraduate Research Journal*). The lack of sufficient in-text citations, moreover, sets this chapter apart from others, particularly as the citation and review of exemplary undergraduate research papers would have been welcomed. As earlier suggested for Part II 7 on data management, adding domain-specific examples of model publications or more general works on disseminating digital research (e.g. [Fitzpatrick 2016]; [O’Sullivan et al. 2016]; [van der Weel and Praal 2020]) would strengthen the underlying message.

Despite some problematic areas, Ignatow and Mihalcea’s *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*, remains an excellent resource for students and researchers new to text analysis. The text is punctuated by especially rich and dense veins of technical ore that a reader can readily extract and refine for their own research areas. On its own, an introductory course could easily be structured around the text and any scholar in the humanities or social sciences interested in breaking into the world of digital scholarship would be well rewarded by reading sections of this book. It is worth concluding that the authors have done an exemplary job in taking a process-oriented approach, which emphasizes techniques, theories, and applications without the baggage of relying on a single toolset. The effect is practical content that will last in an otherwise ephemeral world of digital tools. 8

Works Cited

Briney 2015 Briney, K. *Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success*. Pelagic Publishing, United Kingdom (2015).

Corti et al. 2019 Corti, L., Van den Eynden, V., Bishop, L., and Woollard, M. *Managing and Sharing Research Data*. 2nd edition. SAGE Publications, Washington, D.C. (2019).

Fitzpatrick 2016 Fitzpatrick, K. “Peer Review.” In S. Schreibman, R. Siemens, and J. Unsworth (eds.), *A New Companion to Digital Humanities*, West Sussex: Wiley-Blackwell (2016), pp. 439-448.

Ignatow and Mihalcea 2016 Ignatow, G. and Mihalcea, R. *Text Mining: A Guidebook for the Social Sciences*. SAGE Publications, Thousand Oaks, California (2016).

Ignatow and Mihalcea 2017 Ignatow, G. and Mihalcea, R. *An Introduction to Text Mining: Research Design, Data Collection, and Analysis*. SAGE Publications, Thousand Oaks, California (2017).

O’Sullivan et al. 2016 O’Sullivan, J., Long, C.P., and Mattson, M. “Dissemination as Cultivation: Scholarly Communications in a Digital Age.” In C. Crompton, R.J. Lane, and R. Siemens (eds.), *Doing Digital Humanities: Practice, Training, Research*, New York: Routledge (2016), pp. 384-398.

van der Weel and Praal 2020 van der Weel, A. and Praal, F. “Publishing in the Digital Humanities: The Treacle of the Academic Tradition.” In J. Edmond (ed.), *Digital Technology and the Practices of Humanities Research*, Cambridge: Open Book Publishers (2020), pp. 21-48. DOI: 10.11647/OBP.0192.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License.