

Preface

This book provides an introduction to computational text analysis using the open source programming language R. Unlike other very good books on the use of R for the statistical analysis of linguistic data¹ or for conducting quantitative corpus linguistics,² this book is meant for students and scholars of literature and then, more generally, for humanists wishing to extend their methodological toolkit to include quantitative and computational approaches to the study of text. This book is also meant to be short and to the point. R is a complex program that no single textbook can demystify. The focus here is on making the technical palatable and more importantly making the technical useful and immediately rewarding! Here I mean rewarding not in the sense of satisfaction one gets from mastering a programming language, but rewarding specifically in the sense of quick return on your investment. You will begin analyzing and processing text right away and each chapter will walk you through a new technique or process.

Computation provides access to information in texts that we simply cannot gather using our traditionally qualitative methods of close reading and human synthesis. The reward comes in being able to access that information at both the micro and macro scale. If this book succeeds, you finish it with a foundation, with a broad exposure to core techniques and a basic understanding of the possibilities. The real learning will begin when you put this book aside and build a project of your own. My aim is to give you enough background so that you can begin that project comfortably and so that you'll be able to continue to learn and educate yourself.

When discussing my work as a computing humanist, I am frequently asked whether the methods and approaches I advocate succeed in bringing new knowledge to our study of literature. My answer is strong and resounding *yes*. At the same time, that strong *yes* must be qualified a bit; not everything that text analysis reveals is a breakthrough discovery. A good deal of computational work is specifically aimed

¹ Baayen, H. A. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge UP, 2008.

² Gries, Stefan Th. *Quantitative Corpus Linguistics with R: A Practical Introduction*. New York: Routledge, 2009.

at testing, rejecting, or reconfirming the knowledge that we think we already possess. During a lecture about macro-patterns of literary style in the nineteenth century novel, I used an example from *Moby Dick*. I showed how *Moby Dick* is a statistical mutant among a corpus of 1,000 other nineteenth century American novels. A colleague raised his hand and pointed out that literary scholars already know that *Moby Dick* is an aberration, so why, he asked, bother computing an answer to a question we already know?

My colleague's question was good; it was also revealing. The question said much about our scholarly traditions in the humanities. It is, at the same time, an ironic question: as a discipline, we have tended to favor a notion that literary arguments are never closed. Do we really know that *Moby Dick* is an aberration? Maybe *Moby Dick* is only an outlier in comparison with the other 20 or 30 American novels that we have traditionally studied alongside *Moby Dick*. My point in using *Moby Dick* was not to pretend that I had discovered something new about the position of the novel in the American literary tradition, but rather to bring a new type of evidence and a new perspective to the matter and in so doing fortify (in this case) the existing hypothesis.

If a new type of evidence happens to confirm what we have come to believe using far more speculative methods, shouldn't that new evidence be viewed as a good thing? If the latest Mars rover returns more evidence that the planet could have once supported life, that new evidence would be important. Albeit it would not be as shocking or exciting as the first discovery of microbes on Mars, or the first discovery of ice on Mars, but it would be an important evidence nevertheless, and it would add one more piece to a larger puzzle. So, computational approaches to literary study can provide complementary evidence, and I think that is a good thing.

The approaches outlined in this book also have the potential to present contradictory evidence, evidence that challenges our traditional, impressionistic, or anecdotal theories. In this sense, the methods provide us with some opportunity for the kind of falsification that Karl Popper and post-positivism in general offer as a compromise between strict positivism and strict relativism. But just because these methods *can* provide contradiction, we must not get caught up in a numbers game where we only value the testable ideas. Some interpretations lend themselves to computational or quantitative testing; others do not, and I think that is a good thing.

Finally, these methods can lead to genuinely new discoveries. Computational text analysis has a way of bringing into our field of view certain details and qualities of texts that we would miss with just the naked eye.³ Using computational techniques, Patrick Juola recently discovered that J. K. Rowling was the real author of *The Cuckoo's Calling* a book Rowling wrote under the pseudonym Robert Galbraith. Naturally, I think Juola's discovery is a good thing too.

³ See Flanders, Julia. "Detailism, Digital Texts, and the Problem of Pedantry." *TEXT Technology*, 2:2005, 41–70.

This is all I have to say regarding a theory for or justification of text analysis. In my other book, I'm a bit more polemical.⁴ The mission here is not to defend the approaches but to share them.

Lincoln, NE
January 2014

Matthew L. Jockers

⁴ Jockers, Matthew. *Macroanalysis: Digital Methods and Literary History*. UIUC Press, 2013.

Text Analysis with R for Students of Literature

Jockers, M.L.

2014, XVI, 194 p. 40 illus., 10 illus. in color., Hardcover

ISBN: 978-3-319-03163-7