Journal of Statistical Software
January 2008, Volume 29, Book Review 9. http://www.jstatsoft.org/

Reviewer: Ryan Rosario
University of California at Los Angeles

Practical Text Mining with Perl

Roger Bilisoly
John Wiley & Sons, Hoboken, NJ, 2008.
ISBN 978-0-470-17643-6. 296 pp. USD 89.95.
http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470176431.html

Introduction

Bilisoly's *Practical Text Mining with Perl* is the fourth volume in the Wiley series on data mining and discusses the application of Perl to the field of text mining. Text mining is a subdiscipline of data mining that focuses on the extraction of patterns from text. The author divides this volume into four main topics which he denotes as *transitions*: a blend of Perl basics and pattern extraction, probability and statistics applied to information theory, methods for multivariate analysis, and miscellaneous topics. These topics span nine chapters. The author concludes each chapter with a comprehensive set of exercises that use the material introduced in the chapter. Some of the exercises are basic comprehension exercises, while others involve critical thinking and a thirst for programming. Some exercises also provide the beginner with strategies for learning programming languages such as Perl. This book focuses on the use of Perl for mining, cleaning and basic analysis and uses R for statistical analysis and visualization.

Book content

Perl syntax and concepts are cleverly integrated with text mining tasks throughout the book. Perl lessons do not follow the typical sequence of lessons one would expect from a Perl programming text (such as, e.g., Wall, Christiansen, and Orwant 2000). Rather, concepts are introduced and discussed as necessary to the task at hand. Of course, any book cannot cover all of the issues related to text mining, so the author introduces several references in each chapter with a synopsis of what is discussed in the reference. The organization of the book, as well as the numerous references included provides an inspiring experience for the reader. Exercises sometimes involve Perl concepts that have not yet been discussed to keep the advanced reader busy. Chapters also include various code snippets and programs. The author describes the syntax in each line and clearly explains how the syntax is used. Discussion about the various challenges in text mining inspires the reader to think critically about programming and the iterative nature of refining code.

Chapter 1 presents the reader with a breakdown of the content in the book as well as some helpful advice for reading the book for all programming ability levels. Chapter 2, *Text Patterns*, dives into regular expressions. This section contains a lot of material that moves at a quick pace, but there are enough explanations and references for beginners to become acquainted with the concepts. This chapter also explores the various complications of text mining, and how regular expressions can be further refined to minimize false matches and missed matches in text. Topics include pattern matching, match variables, substitutions, greediness, backreferences and lookaround. Perl concepts include basic file input/output, loops, tokenization, sentence segmentation.

Chapter 3, *Quantitative Text Summaries*, introduces Perl data types and the usage of numeric variables and data structures for token counting and text storage. Concepts include scalars, arrays and hashes, as well as predefined functions and subsetting methods for these data types. Complex data structures such as references, pointers, and arrays of arrays are briefly described.

Chapter 4, *Probability and Text Sampling*, introduces the definition and concept of probability. Introductory discussion uses coin flipping as motivation and progresses to using probabilities to describe text. This chapter introduces conditional probability, independence, random variables, mean, variance, the effect of sample size, and the concept of a sampling distribution. Chapter 4 also introduces the bag of words model.

Chapter 5, *Applying Information Retrieval to Text Mining*, discusses the vector space model including the term document matrix, cosine distance, and inverse document frequency. The author begins by reviewing the use of Perl for counting and is introduced to Perl subroutines to create functions of counts. The author also introduces the reader to importing modules. This is the first chapter that introduces the reader to R syntax. The author uses R for basic tasks such as matrix multiplication and calculating the cosine distance.

Chapter 6, *Concordance Lines and Corpus Linguistics*, discusses statistical sampling and the challenges of extracting meaning from text including word morphologies, collocations and phrasal verbs. Readers learn to write more complex subroutines to clean and display text as well as describe it quantitatively.

Chapter 7, *Multivariate Techniques with Text* requires more background in statistics than previous chapters, and immediately mentions principal components analysis (PCA) in the introduction. The second section takes a step backwards and introduces the reader to basic statistics such as the mean and standard deviation as well as z scores, correlations, covariances and cosines. The author provides a quick review of basic linear algebra for correlation matrices. The fourth section is rather short and discusses PCA but does not clearly describe the connection between PCA and text analysis. Chapter 7 concludes with a brief note of factor analysis and the author describes his preference for PCA over factor analysis. Bilisoly uses R throughout the entire chapter.

Chapter 8, *Text Clustering* introduces k-means and other clustering techniques with R. Bilisoly discusses the use of PCA for clustering as well as hierarchical clustering and the dendogram. Classification is discussed througout the chapter, but is not discussed thoroughly due to the lack of availability of public domain training sets.

Chapter 9, *A Sample of Additional Topics*, wraps up this volume in data mining with some other topics the author did not feel fit into other parts of the book. Topics include Perl modules for text mining, tagging, analysis of text in languages other than English, permutation tests and hypothesis testing.

The author also provides reference appendices on Perl and R.

Conclusion

Practical Text Mining with Perl is an excellent book for readers at a variety of different programming skill levels. In particular, readers with some programming background in non-scripting languages, graduate students, and the "ambitious beginner" that is willing to put in the extra time to develop a good foundation, will find this book very useful. This book does not assume any programming ability, however, there are sections that are heavy and detail-oriented that can overwhelm readers that have never seen a scripting language or the concept of regular expressions. To curtail these difficulties, the author provides several references throughout the chapter, which are summarized at the end of the chapter. Readers can refer to any of these references to delve deeper into the content of the chapter, or to clarify or revisit any concepts that may have been difficult to understand. Bilisoly's book would serve as a good text for an introductory text mining course, and could be supplemented with lecture notes for Web mining or data mining courses. On the other hand, the author's choice of Perl strongly dates this volume, as other languages, particularly Python, seem to have gained more traction for text mining recently.

References

Wall L, Christiansen T, Orwant J (2000). *Programming Perl*. 3rd edition. O'Reilly Media, Inc., Sebastopol, CA.

Reviewer:

Ryan Rosario
University of California at Los Angeles
Department of Statistics
Los Angeles, CA 90095, United States of America
E-mail: rosario@stat.ucla.edu
URL: http://www.stat.ucla.edu/~rosario/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 29, Book Review 9 *Published:* 2008-01-14
January 2008