Text Mining with MATLAB®

Rafael E. Banchs

Text Mining with MATLAB®

Second Edition

Springer

Rafael E. Banchs
Mountain View, CA, USA

ISBN 978-3-030-87694-4 ISBN 978-3-030-87695-1 (eBook)
https://doi.org/10.1007/978-3-030-87695-1

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is the result of a multidisciplinary journey during the last few years of my career. As an Electrical Engineer, who did his dissertation on Electromagnetic Field Theory, I found myself surprised about writing a book on Text Mining! It was in 2004, when I first got involved in natural language processing research, at the TALP research centre of Univesitat Politècnica the Catalunya, in Barcelona. There, I participated in the European Project TC-STAR, which focused on the problem of speech-to-speech translation. Later, at Barcelona Media Innovation Centre, I got the opportunity to further explore other natural language applications and problems such as information retrieval and sentiment analysis. Finally, during my years at the Institute for Infocomm Research, in Singapore, I had the opportunity to work on knowledge representation, question answering and dialogue.

In my experience as Electrical Engineer, the MATLAB® programming platform has always been an excellent tool for conducting experimental research and proof of concepts, as well as for implementing prototypes and applications. However, in the natural language processing community, with the exception of a few machine learning practitioners that have entered in the community via text mining applications, there is not a well-established culture of using the MATLAB® platform. As a technical computing software that specializes in operating with matrices and vectors, MATLAB® offers and excellent framework for text mining and natural language processing research and development.

This book has been written with two objectives in mind. It aims at opening the doors of natural language research and applications to MATLAB® users from other disciplines, as well as introducing the new practitioners in the field to the many possibilities offered by the MATLAB® programming platform. The book has been conceived as an introductory book, which should be easy to follow and digest. All examples and figures presented in the book can be reproduced by following the same procedures described for each case.

Finally, I would like to thank all the persons that have encouraged and helped me to make this project come to life. Special thanks to the MATLAB® Book Program and the Springer Editorial teams for all the support provided. Thanks also to my colleagues who have helped reviewing the different chapters of the book. And special thanks to my wife, my children and my parents for their support and their patience!

Rafael E. Banchs
Mountain View, California, August 2021

MATLAB® is a registered trademark of The MathWorks, Inc.

All examples and figures presented in this book were generated with MATLAB® version 9.9.0.1467703 (R2020b) and Text Analytics Toolbox™ version 1.6

All illustrated examples of MATLAB® dialog boxes, figures and graphic interfaces have been reprinted with permission from The MathWorks, Inc.

For MATLAB® and Simulink® product information, please contact:

The MathWorks, Inc

3 Apple Hill Drive

Natick, MA, 01760-2098 USA

Tel: 508-647-7000

Fax: 508-647-7001

E-mail: info@mathworks.com

Web: https://www.mathworks.com

How to buy: https://www.mathworks.com/store

Find your local office: https://www.mathworks.com/company/worlwide

Table of Contents

1 Introduction 1
1.1 About Text Mining and MATLAB® 1
1.2 About this Book 3
1.3 A (very) Brief Introduction to MATLAB® 6
1.4 The Text Analytics Toolbox™ 11
1.5 Further Reading 13
1.6 References 13

PART I: FUNDAMENTALS 15

2 Handling Text Data 17
2.1 Character and Character Arrays 17
2.2 Handling Text with Cell Arrays 20
2.3 Handling Text with Structures 23
2.4 Handling Text with String Arrays 26
2.5 Some Useful Functions 29
2.6 Further Reading 33
2.7 Proposed Exercises 33
2.8 References 35

3 Regular Expressions 37
3.1 Basic Operators for Matching Characters 37
3.2 Matching Sequences of Characters 40
3.3 Conditional Matching 43
3.4 Working with Tokens 45
3.5 Further Reading 48
3.6 Proposed Exercises 48
3.7 References 51

4 Basic Operations with Strings 53
4.1 Searching and Comparing 53
4.2 Replacement and Insertion 61
4.3 Segmentation and Concatenation 64
4.4 Set Operations 71
4.5 Further Reading 77
4.6 Proposed Exercises 77
4.7 References 80

5 Reading and Writing Files 81
5.1 Basic File Formats 81
5.2 Other Useful Formats 90
5.3 Handling Files and Directories 106
5.4 Further Reading 113
5.5 Proposed Exercises 114
5.6 References 117

6 The Structure of Language 119
6.1 Levels of the Linguistic Phenomena 119
6.2 Morphology and Syntax 122
6.3 Semantics and Pragmatics 131
6.4 Further Reading 137
6.5 Proposed Exercises 137
6.6 References 140

PART II: MATHEMATICAL MODELS 143

7 Basic Corpus Statistics 145
7.1 Fundamental Properties 145
7.2 Word Co-occurrences 158
7.3 Accounting for Order 165
7.4 Further Reading 170
7.5 Proposed Exercises 171
7.6 Short Projects 173
7.7 References 175

8 Statistical Models 177
8.1 Basic *n*-gram Models 177
8.2 Discounting 180
8.3 Model Interpolation 188
8.4 Topic Models 192
8.5 Further Reading 204
8.6 Proposed Exercises 205
8.7 Short Projects 207
8.8 References 209

9 Geometrical Models 211
9.1 The Term-Document Matrix 211
9.2 The Vector Space Model 219
9.3 Association Scores and Distances 228
9.4 Further Reading 235
9.5 Proposed Exercises 235
9.6 Short Projects 238
9.7 References 239

10 Dimensionality Reduction 241
10.1 Vocabulary Pruning and Merging 241
10.2 The Linear Transformation Approach 247
10.3 Non-linear Projection Methods 258
10.4 Embeddings 264
10.5 Further Reading 273
10.6 Proposed Exercises 274
10.7 Short Projects 276
10.8 References 277

PART III: METHODS AND APPLICATIONS 279

11 Document Categorization 281
11.1 Data Collection Preparation 281
11.2 Unsupervised Clustering 287
11.3 Supervised Classification in Vector Space 294
11.4 Supervised Classification in Probability Space 308
11.5 Further Reading 317
11.6 Proposed Exercises 318
11.7 Short Projects 322
11.8 References 324

12 Document Search 327
12.1 Binary Search 327
12.2 Vector-based Search 338
12.3 The BM25 Ranking Function 344
12.4 Cross-language Search 347
12.5 Further Reading 358
12.6 Proposed Exercises 359
12.7 Short Projects 361
12.8 References 362

13 Content Analysis 365
13.1 Dimensions of Analysis 365
13.2 Polarity Estimation 370
13.3 Qualifier and Aspect Identification 379
13.4 Entity, Relation and Definition Extraction 389
13.5 Further Reading 398
13.6 Proposed Exercises 399
13.7 Short Projects 403
13.8 References 404

14 Keyword Extraction and Summarization 407
14.1 Keywords and Word Clouds 407
14.2 Text Summarization 420
14.3 Further Reading 429
14.4 Proposed Exercises 430
14.5 Short Projects 432
14.6 References 434

15 Question Answering and Dialogue 435
15.1 Question Answering 435
15.2 Dialogue Systems 450
15.3 Further Reading 466
15.4 Proposed Exercises 466
15.5 Short Projects 472
15.6 References 474