METHODS IN MEDICAL INFORMATICS
Fundamentals of Healthcare Programming in Perl, Python, and Ruby

Jules J. Berman
Chapter 2 Utility Scripts

2.1 Random Numbers
   2.1.1 Script Algorithm
   2.1.2 Analysis

2.2 Converting Non-ASCII to Base64 ASCII
   2.2.1 Script Algorithm
   2.2.2 Analysis

2.3 Creating a Universally Unique Identifier
   2.3.1 Script Algorithm
   2.3.2 Analysis

2.4 Splitting Text into Sentences
   2.4.1 Script Algorithm
   2.4.2 Analysis

2.5 One-Way Hash on a Name
   2.5.1 Script Algorithm
   2.5.2 Analysis

2.6 One-Way Hash on a File
   2.6.1 Script Algorithm
   2.6.2 Analysis

2.7 A Prime Number Generator
   2.7.1 Script Algorithm
   2.7.2 Analysis

Chapter 3 Viewing and Modifying Images

3.1 Viewing a JPEG Image
   3.1.1 Script Algorithm
   3.1.2 Analysis

3.2 Converting between Image Formats
   3.2.1 Script Algorithm
   3.2.2 Analysis

3.3 Batch Conversions
   3.3.1 Script Algorithm
   3.3.2 Analysis

3.4 Drawing a Graph from List Data
   3.4.1 Script Algorithm
   3.4.2 Analysis

3.5 Drawing an Image Mashup
   3.5.1 Script Algorithm
   3.5.2 Analysis

Chapter 4 Indexing Text

4.1 ZIPF Distribution of a Text File
   4.1.1 Script Algorithm
   4.1.2 Analysis

4.2 Preparing a Concordance
   4.2.1 Script Algorithm
   4.2.2 Analysis

4.3 Extracting Phrases
   4.3.1 Script Algorithm
   4.3.2 Analysis

4.4 Preparing an Index
   4.4.1 Script Algorithm
   4.4.2 Analysis
# CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5 Comparing Texts Using Similarity Scores</td>
<td>69</td>
</tr>
<tr>
<td>4.5.1 Script Algorithm</td>
<td>69</td>
</tr>
<tr>
<td>4.5.2 Analysis</td>
<td>76</td>
</tr>
<tr>
<td><strong>Part II Medical Data Resources</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Chapter 5 The National Library of Medicine’s Medical Subject Headings (MeSH)</strong></td>
<td></td>
</tr>
<tr>
<td>5.1 Determining the Hierarchical Lineage for MeSH Terms</td>
<td>81</td>
</tr>
<tr>
<td>5.1.1 Script Algorithm</td>
<td>83</td>
</tr>
<tr>
<td>5.1.2 Analysis</td>
<td>86</td>
</tr>
<tr>
<td>5.2 Creating a MeSH Database</td>
<td>88</td>
</tr>
<tr>
<td>5.2.1 Script Algorithm</td>
<td>88</td>
</tr>
<tr>
<td>5.2.2 Analysis</td>
<td>90</td>
</tr>
<tr>
<td>5.3 Reading the MeSH Database</td>
<td>90</td>
</tr>
<tr>
<td>5.3.1 Script Algorithm</td>
<td>91</td>
</tr>
<tr>
<td>5.3.2 Analysis</td>
<td>92</td>
</tr>
<tr>
<td>5.4 Creating an SQLite Database for MeSH</td>
<td>92</td>
</tr>
<tr>
<td>5.4.1 Script Algorithm</td>
<td>93</td>
</tr>
<tr>
<td>5.4.2 Analysis</td>
<td>96</td>
</tr>
<tr>
<td>5.5 Reading the SQLite MeSH Database</td>
<td>96</td>
</tr>
<tr>
<td>5.5.1 Script Algorithm</td>
<td>96</td>
</tr>
<tr>
<td>5.5.2 Analysis</td>
<td>97</td>
</tr>
<tr>
<td><strong>Chapter 6 The International Classification of Diseases</strong></td>
<td>99</td>
</tr>
<tr>
<td>6.1 Creating the ICD Dictionary</td>
<td>99</td>
</tr>
<tr>
<td>6.1.1 Script Algorithm</td>
<td>100</td>
</tr>
<tr>
<td>6.1.2 Analysis</td>
<td>101</td>
</tr>
<tr>
<td>6.2 Building the ICD-O (Oncology) Dictionary</td>
<td>102</td>
</tr>
<tr>
<td>6.2.1 Script Algorithm</td>
<td>103</td>
</tr>
<tr>
<td>6.2.2 Analysis</td>
<td>104</td>
</tr>
<tr>
<td><strong>Chapter 7 SEER: The Cancer Surveillance, Epidemiology, and End Results Program</strong></td>
<td>107</td>
</tr>
<tr>
<td>7.1 Parsing the SEER Data Files</td>
<td>107</td>
</tr>
<tr>
<td>7.1.1 Script Algorithm</td>
<td>107</td>
</tr>
<tr>
<td>7.1.2 Analysis</td>
<td>109</td>
</tr>
<tr>
<td>7.2 Finding the Occurrences of All Cancers in the SEER Data Files</td>
<td>110</td>
</tr>
<tr>
<td>7.2.1 Script Algorithm</td>
<td>111</td>
</tr>
<tr>
<td>7.2.2 Analysis</td>
<td>114</td>
</tr>
<tr>
<td>7.3 Finding the Age Distributions of the Cancers in the SEER Data Files</td>
<td>115</td>
</tr>
<tr>
<td>7.3.1 Script Algorithm</td>
<td>115</td>
</tr>
<tr>
<td>7.3.2 Analysis</td>
<td>119</td>
</tr>
<tr>
<td><strong>Chapter 8 OMIM: The Online Mendelian Inheritance in Man</strong></td>
<td>123</td>
</tr>
<tr>
<td>8.1 Collecting the OMIM Entry Terms</td>
<td>124</td>
</tr>
<tr>
<td>8.1.1 Script Algorithm</td>
<td>124</td>
</tr>
<tr>
<td>8.1.2 Analysis</td>
<td>125</td>
</tr>
<tr>
<td>8.2 Finding Inherited Cancer Conditions</td>
<td>126</td>
</tr>
<tr>
<td>8.2.1 Script Algorithm</td>
<td>126</td>
</tr>
<tr>
<td>8.2.2 Analysis</td>
<td>128</td>
</tr>
</tbody>
</table>
Chapter 9 PubMed

9.1 Building a Large Text Corpus of Biomedical Information
9.1.1 Script Algorithm
9.1.2 Analysis
9.2 Creating a List of Doublets from a PubMed Corpus
9.2.1 Script Algorithm
9.2.2 Analysis
9.3 Downloading Gene Synonyms from PubMed
9.4 Downloading Protein Synonyms from PubMed

Chapter 10 Taxonomy

10.1 Finding a Taxonomic Hierarchy
10.1.1 Script Algorithm
10.1.2 Analysis
10.2 Finding the Restricted Classes of Human Infectious Pathogens
10.2.1 Script Algorithm
10.2.2 Analysis

Chapter 11 Developmental Lineage Classification and Taxonomy of Neoplasms

11.1 Building the Doublet Hash
11.1.1 Script Algorithm
11.1.2 Analysis
11.2 Scanning the Literature for Candidate Terms
11.2.1 Script Algorithm
11.2.2 Analysis
11.3 Adding Terms to the Neoplasm Classification
11.3.1 Script Algorithm
11.3.2 Analysis
11.4 Determining the Lineage of Every Neoplasm Concept
11.4.1 Script Algorithm
11.4.2 Analysis

Chapter 12 U.S. Census Files

12.1 Total Population of the United States
12.1.1 Script Algorithm
12.1.2 Analysis
12.2 Stratified Distribution for the U.S. Census
12.2.1 Script Algorithm
12.2.2 Analysis
12.3 Adjusting for Age
12.3.1 Script Algorithm
12.3.2 Analysis

Chapter 13 Centers for Disease Control and Prevention Mortality Files

13.1 Death Certificate Data
13.2 Obtaining the CDC Data Files
13.3 How Death Certificates Are Represented in Data Records
13.4 Ranking, by Number of Occurrences, Every Condition in the CDC Mortality Files
13.4.1 Script Algorithm
13.4.2 Analysis
CONTENTS

PART III PRIMARY TASKS OF MEDICAL INFORMATICS

CHAPTER 14 AUTOCODING

14.1 A Neoplasm Autocoder

14.1.1 Script Algorithm
14.1.2 Analysis

14.2 Recoding

CHAPTER 15 TEXT SCRUBBER FOR DEIDENTIFYING CONFIDENTIAL TEXT

15.1 Script Algorithm

15.2 Analysis

CHAPTER 16 WEB PAGES AND CGI SCRIPTS

16.1 Grabbing Web Pages

16.1.1 Script Algorithm
16.1.2 Analysis

16.2 CGI Script for Searching the Neoplasm Classification

16.2.1 Script Algorithm
16.2.2 Analysis

CHAPTER 17 IMAGE ANNOTATION

17.1 Inserting a Header Comment

17.1.1 Script Algorithm
17.1.2 Analysis

17.2 Extracting the Header Comment in a JPEG Image File

17.2.1 Script Algorithm
17.2.2 Analysis

17.3 Inserting IPTC Annotations

17.4 Extracting Comment, EXIF, and IPTC Annotations

17.4.1 Script Algorithm
17.4.2 Analysis

17.5 Dealing with DICOM

17.6 Finding DICOM Images

17.7 DICOM-to-JPEG Conversion

17.7.1 Script Algorithm
17.7.2 Analysis

CHAPTER 18 DESCRIBING DATA WITH DATA, USING XML

18.1 Parsing XML

18.1.1 Script Algorithm
18.1.2 Analysis

18.1.3 Resource Description Framework (RDF)

18.2 Dublin Core Metadata

18.3 Insert an RDF Document into an Image File

18.3.1 Script Algorithm
18.3.2 Analysis

18.4 Insert an Image File into an RDF Document

18.4.1 Script Algorithm
18.4.2 Analysis

18.5 RDF Schema

18.6 Visualizing an RDF Schema with GraphViz

18.7 Obtaining GraphViz
PART IV  MEDICAL DISCOVERY

Chapter 19 Case Study: Emphysema Rates  
19.1 Script Algorithm  
19.2 Analysis  

Chapter 20 Case Study: Cancer Occurrence Rates  
20.1 Script Algorithm  
20.2 Analysis  

Chapter 21 Case Study: Germ Cell Tumor Rates across Ethnicities  
21.1 Script Algorithm  
21.2 Analysis  

Chapter 22 Case Study: Ranking the Death-Certifying Process, by State  
22.1 Script Algorithm  
22.2 Analysis  

Chapter 23 Case Study: Data Mashups for Epidemics  
23.1 Tally of Coccidioidomycosis Cases by State  
23.1.1 Script Algorithm  
23.1.2 Analysis  
23.2 Creating the Map Mashup  
23.2.1 Script Algorithm  
23.2.2 Analysis  

Chapter 24 Case Study: Sickle Cell Rates  
24.1 Script Algorithm  
24.2 Analysis  

Chapter 25 Case Study: Site-Specific Tumor Biology  
25.1 Anatomic Origins of Mesotheliomas  
25.2 Mesothelioma Records in the SEER Data Sets  
25.2.1 Script Algorithm  
25.2.2 Analysis  
25.3 Graphic Representation  
25.3.1 Script Algorithm  
25.3.2 Analysis  

Chapter 26 Case Study: Bimodal Tumors  
26.1 Script Algorithm  
26.2 Analysis  

Chapter 27 Case Study: The Age of Occurrence of Precancers  
27.1 Script Algorithm  
27.2 Analysis
CONTENTS

Epilogue for Healthcare Professionals and Medical Scientists 361
Learn One or More Open Source Programming Languages 361
Don't Agonize Over Which Language You Should Choose 362
Learn Algorithms 362
Unless You Are a Professional Programmer, Relax and Enjoy Being a Newbie 363
Do Not Delegate Simple Programming Tasks to Others 363
Break Complex Tasks into Simple Methods and Algorithms 364
Write Fast Scripts 364
Concentrate on the Questions, Not the Answers 365

Appendix 367
How to Acquire Ruby 367
How to Acquire Perl 367
How to Acquire Python 367
How to Acquire RMagick 368
How to Acquire SQLite 369
How to Acquire the Public Data Files Used in This Book 370
Other Publicly Available Files, Data Sets, and Utilities 376

Index 377