

Table of Contents

Table of Contents	i
Introduction	1
Overview	1
What Is a Bot?	1
What Is a Spider?	2
What Are Agents and Intelligent Agents?	3
What Are Aggregators?	4
The Java Programming Language	4
Wrap Up	5
Chapter 1: Java Socket Programming	6
Overview	6
The World of Sockets	6
Java I/O Programming	14
Proxy Issues	22
Socket Programming in Java	24
Client Sockets	25
Server Sockets	37
Summary	44
Chapter 2: Examining the Hypertext Transfer Protocol	46
Overview	46
Address Formats	46
Using Sockets to Program HTTP	50
Bot Package Classes for HTTP	60
Under the Hood	73
Summary	82
Chapter 3: Accessing Secure Sites with HTTPS	84
Overview	84
HTTP versus HTTPS	84
Using HTTPS with Java	85
HTTP User Authentication	90
Securing Access	96
Under the Hood	105
Summary	115
Chapter 4: HTML Parsing	116
Overview	116
Working with HTML	116
Tags a Bot Cares About	118
HTML That Requires Special Handling	123
Using Bot Classes for HTML Parsing	126
Using Swing Classes for HTML Parsing	128
Bot Package HTML Parsing Examples	133
Under the Hood	153
Summary	163
Chapter 5: Posting Forms	165
Overview	165
Using Forms	165
Bot Classes for a Generic Post	171
Under the Hood	186

Summary	190
Chapter 6: Interpreting Data	191
Overview	191
The Structure of the CSV File	191
The Structure of a QIF File	197
The XML File Format	203
Summary	213
Chapter 7: Exploring Cookies.....	215
Overview	215
Examining Cookies	216
Bot Classes for Cookie Processing.....	230
Under the Hood	232
Summary	238
Chapter 8: Building a Spider	239
Overview	239
Structure of Websites	239
Structure of a Spider.....	242
Constructing a Spider	246
Summary	266
Chapter 9: Building a High-Volume Spider	267
Overview	267
What Is Multithreading?.....	267
Multithreading with Java.....	268
Synchronizing Threads.....	272
Using a Database.....	275
The High-Performance Spider	283
Under the Hood	284
Summary	315
Chapter 10: Building a Bot.....	317
Overview	317
Constructing a Typical Bot	317
Using the CatBot.....	331
An Example CatBot	336
Under the Hood	342
Summary	359
Chapter 11: Building an Aggregator.....	360
Overview	360
Online versus Offline Aggregation	360
Building the Underlying Bot	361
Building the Weather Aggregator	369
Summary	374
Chapter 12: Using Bots Conscientiously	375
Overview	375
Dealing with Websites	375
Webmaster Actions	381
A Conscientious Spider	383
Under the Hood	396
Summary	401
Chapter 13: The Future of Bots	403
Overview	403

Internet Information Transfer.....	403
Understanding XML	404
Transferring XML Data	408
Bots and SOAP.....	412
Summary	412
Appendix A: The Bot Package	414
Utility Classes	414
HTTP Classes.....	416
The Parsing Classes.....	419
Spider Classes	424
Appendix B: Various HTTP Related Charts.....	430
The ASCII Chart	430
HTTP Headers.....	434
HTTP Status Codes	436
HTML Character Constants	439
Appendix C: Troubleshooting.....	441
WIN32 Errors	441
UNIX Errors	441
Cross-Platform Errors	444
How to Use the NOBOT Scripts	446
Appendix D: Installing Tomcat.....	447
Installing and Starting Tomcat	447
A JSP Example	449
Appendix E: How to Compile Examples Under Windows.....	451
Using the JDK	451
Using VisualCafé	456
Appendix F: How to Compile Examples Under UNIX.....	458
Using the JDK	458
Appendix G: Recompiling the Bot Package.....	461
Glossary.....	463