

# Index

---

**Note to the Reader:** Throughout this index **boldfaced** page numbers indicate primary discussions of a topic. *Italicized* page numbers indicate illustrations.

## A

---

- a tags, 251
- absolute URLs
  - defined, 488
  - for hyperlinks, **123–124**
- Accept header, 451
- Accept-Charset header, 451
- Accept-Encoding header, 451
- Accept-Language header, 451
- accept method, 41
- Access, **290–291**, *290–291*
- access methods in bot detection, 395
- action attribute, 194
- actionPerformed method
  - in GetImage, 142
  - in GetSite, 265, 401
  - in SecureGET, 104
  - in SecurePrompt, 101
  - in SendMail, 31
  - in SiteSubmit, 187
  - in ViewURL, 73
  - in ViewURLCookie, 234
  - in WatchBBS, 334
- add method, 67, 435
- addAttribute method, 238, 242, 438
- addAuthHeader method, 113–114, 431
- addCookieHeader method, 431
- addImage method, 162, 437
- addInput method, 178, 193–194, 436
- addNotify method
  - in GetImage, 141
  - in GetSite, 264, 399–400
  - in SecureGET, 103–104
  - in SecurePrompt, 100
  - in SendMail, 28–29
  - in SiteSubmit, 184–185
  - in ViewURL, 71–72
  - in ViewURLCookie, 233
  - in WatchBBS, 332
- addresses
  - in HTTP, **46–47**
  - IP. *See* IP addresses
  - in QIF files, 205
  - of SMTP servers, 33
- addWorkload method
  - in IWorkloadStorable, 309, 441
  - in Spider, 259, 302
  - in SpiderDone, 442
  - in SpiderInternalWorkload, 316, 443
  - in SpiderSQLWorkload, 312, 443
- Advanced tab, 474–475, *475*
- agent headers, 63
- agents, 488
- aggregate hammering, 488
- aggregation, **370**
  - inline, **370–371**
  - offline, **371**
- aggregators
  - defined, 488
  - Weather aggregator, **378–379**, *379*
    - Weather class for, **380–382**
    - weather.jsp page for, **379–380**
  - Weather bot for, **371–372**
    - building, **374–378**
    - extracting data from, **374**
    - planning, **372–374**, *373*
- AI (artificial intelligence), 488
- Allow header, 451
- AlphaCONNECT bot name, 392
- alt attribute, 123, 132
- AltaVista search engine, 179, 392
- ampersands (&)
  - for attributes, 195
  - for forms, 177
  - in URLs, 195
- anchors
  - for links, 251
  - in URLs, 49
- anonymous bot identification, **391–392**
- Apache project, 466–467

Apache web server, 94–97  
 APIs (application programming interfaces), 488  
 Applet class, 130  
 area tags, 251  
 artificial intelligence (AI), 488  
 ASCII data  
   chart for, **446–450**  
   encoding binary to, **108–111**  
 assignWorkload method  
   in IWorkloadStorable, 309, 441  
   in SpiderInternalWorkload, 316, 443  
   in SpiderSQLWorkload, 312, 443  
 asterisks (\*) in exclusion files, 394  
 at signs (@) in URLs, 195  
 Attribute class, **66–74, 435**  
 Attribute method, 435  
 AttributeList class, **67–74, 68, 435**  
 AttributeList method, 435  
 attributes  
   for controls, 173–174  
   in HTMLForm, 194–195  
   for Set-Cookie header, 225  
   in XML, 212, 418  
 attributes method, 150–151, 155–156  
 authentication  
   defined, 488  
   HTTP, **92–94, 93–94**  
     with Bot package, **98**  
     on clients, **96–98**  
     HTTP class for, **98, 113–114**  
     server setup for, **94–96**  
 Authorization header, 451  
 AuthUserFile command, 96  
 autoexec.bat file, 473–474, 474  
 automatic redirection  
   in HTTP, 64–65  
   send for, 77–78  
 available method, 15–16

## B

Babel Fish utility, 146  
 bars (|) in CSV files, 198  
 base tags, 251  
 base URLs, 124, 488  
 base64 encoding, 12  
   defined, 488  
   process of, **108–111**  
 base64Encode method, 108–109, 430  
 Base64OutputStream class, **108–113, 428**  
 Base64OutputStream method, 428

.bash\_profile file, 460, 460, 481–482  
 basic filters, 17  
 beginning tags, **122**  
 bin directory  
   in JDK, **473**  
   in UNIX, **481**  
 binary data, encoding to ASCII, **108–111**  
 blocking threads, 41, 488  
 body in HTTP requests, 54  
 bot exclusion files  
   contents of, **393–394**  
   defined, 488–489  
 bot.jar file  
   in JDK, **472–473**  
   locating, 461–462  
   in UNIX, **480–481**  
   in VisualCafé, 477, 478  
 Bot package, **61, 428**  
   Attribute class, **66–67, 435**  
   AttributeList class, **67–68, 435**  
   Base64OutputStream class, **428**  
   BotExclusion class, **439**  
   CookieParse class, **239, 242–244, 436**  
   defined, 489  
   for extracting images, **139–146, 139**  
   finding, 461  
   HTML.Attribute class, **136–138**  
   HTML.Tag class, **134–136**  
   HTMLForm class, **178–179, 191–195, 436**  
   HTMLForm.FormElement class, **436–437**  
   HTMLPage class, **131–132, 159–166, 177–178, 437**  
   HTMLParse class, **158, 437**  
   HTMLParser class, **438**  
   HTMLTag class, **438**  
   HTTP class, **61–65, 431–433**  
   HTTP user authentication with, **98**  
   for HTTPS, **92**  
   HTTPSocet class, **66, 77–83, 433**  
   ISpiderReportable interface, **440–441**  
   IWorkloadStorable interface, **441**  
   Link class, **132, 166–168, 429**  
   Log class, **74–76, 430**  
   Parse class, **237–242, 438–439**  
   recompiling, **484–485**  
   Spider class, **441–442**  
   SpiderDone class, **442**  
   SpiderInternalWorkload class, **443**  
   SpiderSQLWorkload class, **443**  
   SpiderWorker class, **443–444**  
   URLUtility class, **430**  
   for website translations, **146–156, 157**  
   working with, **68–74, 68**

**500** BotExclusion class—close method

BotExclusion class, 393, **395–396**, 407–412, 439

bots

CatBot. *See* CatBot bot

defined, 488

detection of, **394–395**

identification of

anonymous, **391–392**

combined, **392**

HTTP headers for, **390–391**

unique, **392**

WatchBBS, **328–330**, 330–331

code for, **330–339**

operation of, **339–341**

weaknesses in, **341–342**

Weather, **371–372**

building, **374–378**

extracting data from, **374**

planning, **372–374**, 373

bottlenecks, 279, 489

breaksFlow method, 134

bridge drivers, 289, 293, 489

broadband connections, 489

BufferedInputStream class, 18–19

BufferedOutputStream class, 18

BufferedReader class, 22, 34, 60

buffering

defined, 489

in output streams, **13–14**

build.bat file, 484

buildTag method, 438

bulletin boards, WatchBBS bot, **328–330**, 330–331

code for, **330–339**

operation of, **339–341**

weaknesses in, **341–342**

buttons, **174–175**

byte data type with write, 13

**C**

cacerts file, 90, 489

callback methods, **133–134**

calling method, 35

Cancel\_actionPerformed method

in SecurePrompt, 101

in SendMail, 32–33

canceling spiders, 272

carets (^) in QIF files, 204

case sensitivity in XML, 211, 418

catalina.bat file, 469

CatBot bot

CatBot class for, **353–358**

connecting JSP in, **350**

methods of, **345–346**

properties of, **344–345**

recognizers in, **342–343**, **346–347**, **350–352**,  
**358–368**

running JSP pages in, **348–349**, 348

ShipBot class for, **352–353**

starting, **343**, 344

CatBot Properties, 344–345

categories in QIF files, 205

CGI-BIN resource, 489

chaining filters, 19

character constants, HTML, **455–456**

check boxes, 175

check numbers in QIF files, 205

checkApplet method, 136

child classes, 78–79

child directories in URLs, 124

class libraries for SOAP, 423

class ViewURL, 69

CLASSPATH environmental variable

for bot.jar, 462–463

defined, 489

for JDK, **473–476**, 474–476, **481–482**

in JSSE, 89–90

in Tomcat, 468–469

clear method

in AttributeList, 67, 435

in IWorkloadStorable, 309, 441

in SpiderInternalWorkload, 318, 443

in SpiderSQLWorkload, 315, 443

clearCookies method, 62, 431

clearing

attributes, 67

cookies, 62–63, 431

Client Error status codes, **454–455**

client headers, 63

defined, 489

with lowLevelSend, 80–81

client-side image maps, **124–125**

clients

defined, 489

HTTP user authentication for, **96–98**

in peer-to-peer networks, 4

SMTP sockets, **23–36**, 26

clipping, 489

clone method

in Attribute, 66, 435

in AttributeList, 67, 435

in HTMLTag, 438

close method

in InputStream, 15

in OutputStream, 12, 14

closing  
 input streams, **17**  
 output streams, **14–15**

colons (:)  
 in HTTPS, **92**  
 in URLs, **195**

com.heaton.bot identifier, **61**

combined bot identification, **392**

commas (,)  
 for cookies, **224, 226**  
 in CSV files, **198**

Comment attribute, **225**

comments  
 in exclusion files, **394**  
 in HTML, **121–122**  
 for scripts, **128**  
 in VisualCafé, **36**

communications protocols, **46**

compiling  
 bot package, **484–485**  
 JDK for, **472–476, 474–476**  
 path for, **458–460, 460**  
 under UNIX systems, **480–482**  
 VisualCafé for, **477, 478**

complete queues in spiders, **254**

COMPLETE status code, **308**

completePage method  
 in ISpiderReportable, **257, 307, 441**  
 in Spider, **259, 305**  
 in SpiderDone, **442**  
 in UpdateTarget, **270, 406**

completeWorkload method  
 in IWorkloadStorable, **309, 441**  
 in SpiderInternalWorkload, **317, 443**  
 in SpiderSQLWorkload, **312, 443**

compound names, **362**

Connection class, **293**

connections  
 defined, **489**  
 opening, **58–61**

conscientious spiders, **395–412**

console based logging, **75–76**

constants, HTML, **455–456**

Content-Encoding header, **451**

Content-Length header, **55, 80, 451**

Content-Type header, **451**

Control Panel, **290**

controls, **173–177**

CookieParse class, **239, 242–244, 436**

cookies, **222–223**  
 appearance of, **226–227**  
 clearing, **62–63, 431**

CookieParse class for, **239, 242–244**  
 defined, **489**  
 example, **230–237, 230**  
 exchanging, **227–229**  
 in HTTP, **62, 64–65, 94**  
 in Internet Explorer, **223–224, 223–224**  
 Parse class for, **237–242**  
 receiving, **224–225**  
 returning, **225–226**  
 send for, **77**  
 session and persistent, **229**  
 for Weather bot, **374**

cooperation with websites, **389**

copy method  
 in HTTP, **431**  
 in HTTPSocet, **433**  
 \_country property, **344**

country recognizers, **343**

CREATE TABLE command, **292**

createSocket method, **116**

critical sections, **283, 490**

cross-platform programs  
 defined, **490**  
 troubleshooting, **461–463**

CSV files  
 format of, **198**  
 parsing example, **199–203**

current value of form controls, **173**

## D

daemon threads, **490**

DATA command, **24–25**

data consumers, **12**

data producers, **16**

data scheme, **47**

Data Source Names (DSNs), **290–291**

Database Management Systems (DBMSs), **287, 490**

databases for spiders, **287**  
 JDBC for, **292–295**  
 selecting and configuring, **289–292, 290–291**  
 SQL language for, **287–289**

DataInputStream class, **18, 22**

DataOutputStream class, **18, 22**

Date header, **451**

dates in QIF files, **205**

DBMSs (Database Management Systems), **287, 490**

deadlock situations, **14**

definition tags for tables, **126–127**

DELETE statement, **288**

denial of service (DoS) attacks  
 defined, 490  
 hammering as, 387  
 derived classes, 78–79  
 detection of bots, **394–395**  
 DHCP (Dynamic Host Configuration Protocol)  
 defined, 490  
 for resolving IP addresses to hostnames, **8–9**  
 directories  
 in HTTP user authentication, 95  
 search, 179  
 in URLs, 124  
 Directories tab, 90  
 display method, 216–218  
 distributed hammering, **388**, 490  
 DNS (domain name service)  
 defined, 490  
 for resolving IP addresses to hostnames, **8–11**  
 doIndentation method, 216  
 Domain attribute, 225–226  
 domain name service (DNS)  
 defined, 490  
 for resolving IP addresses to hostnames, **8–11**  
 DoS (denial of service) attacks  
 defined, 490  
 hammering as, 387  
 dots (.)  
 in IP addresses, 5  
 in URLs, 124  
 DSNs (Data Source Names), 290–291  
 Dynamic Host Configuration Protocol (DHCP)  
 defined, 490  
 for resolving IP addresses to hostnames, **8–9**

---

## E

e-commerce, 86  
 e-mail  
 links for, 251  
 SMTP for, 23  
 eatWhiteSpace method, 238, 240, 438  
 8859\_1 character encoding, 488  
 embedded HTML objects, **130**  
 emulation  
 defined, 490  
 forms, **170–171**  
 encoding  
 base64, 12  
 defined, 488  
 process of, **108–111**  
 in SOAP, 420

encryption  
 in Apache server, 96  
 in base64 encoding, 108  
 restrictions on, 89  
 in SSL, 87  
 ending tags  
 in HTML, **122**  
 in XML, 211–212  
 endpoints, 490  
 envelopes in SOAP, 420  
 environmental variables, 475  
 eof method, 238, 240–241, 438  
 equal signs (=)  
 for attributes, 174  
 in URLs, 195  
 in XML, 212  
 error queues in spiders, 254  
 ERROR status code, 308  
 errors and error codes  
 in HTTP headers, 37–38  
 logging, 76  
 with SMTP, 34  
 Etag header, 451  
 evaluating forms, **171**, 172  
 exceptions for URLs, 57, 60–61  
 Excite, bot names for, 392  
 exclamation points (!)  
 in HTML, 121  
 in QIF files, 204  
 exclusion files, **393–395**, 488–489  
 executeQuery method, 294  
 executeUpdate method, 294–295  
 Expires header, 451  
 ExtendThread class, 280  
 Extensible Markup Language. *See* XML (Extensible Markup Language)  
 external links, **250–251**, 490  
 extracting images, **139–146**, 139

---

## F

factorials, 253  
 factories, 116  
 fat clients for aggregators, 371  
 file based logging, 75–76  
 File DSN tab, 291  
 File scheme, 47  
 fileAggregate method, 376, 382  
 FileInputStream class, 16  
 FileOutputStream class, 13  
 FilterInputStream class, 18

FilterOutputStream class, 18

filters

- against bots, 395
- defined, 491
- in I/O programming, **17–19**

findOption method, 346, 361

FindPackage class, 349–350

findPackage method, 349–350

findPrompt method, 347, 361–362

firewalls, 491

firstVowel method, 154

flat files, 416

flush method

- in Base64OutputStream, 112, 428
- in OutputStream, 12, 14

form.html page, 171

form tag, 125, 173

formats, 416–417

FormElement class, 178–179, **436–437**

forms, **170**

- contents of, **173–174**
- controls in, **173–177**
- emulating, **170–171**
- evaluating, **171, 172**
- HTMLForm class for, **178–179, 191–195**
- HTMLPage class for, **177–178**
- listing, 131
- parsing, **125–126, 126**
- posting to, **55–56, 55, 172–173, 191**
- submitting to search engines, **179–191, 180**

foundExternalLink method

- in GetSite, 268–269, 404
- in ISpiderReportable, 257, 307, 441
- in Spider, 259, 303–304
- in SpiderDone, 442

foundInternalLink method

- in GetSite, 268, 397, 404
- in ISpiderReportable, 257, 307, 441
- in Spider, 259, 303
- in SpiderDone, 442

foundOtherLink method

- in GetSite, 405
- in ISpiderReportable, 257–258, 307, 441
- in Spider, 259, 304
- in SpiderDone, 442

frames, **129–130**

frequency of access, detection of, 394

FROM clause, 287

From header, 451

FTP scheme, 47

FtpProxy settings, 21

## G

General tab, 223

GET command in HTTP, 37

get method

- in AttributeList, 67, 435
- in CookieParse, 239, 243–244, 436
- in HTMLParser, 438

get operations in synchronizing threads, 283–284

GET requests, **52–54, 491**

getAction method, 178, 193, 436

getAgent method, 63, 431

getALT method, 132, 167, 429

getAttributeNames method, 156

getAttributeValue method, 438

getBody method, 63, 431

getByName method, 9–10

getClientHeaders method, 63, 431

getConsole method, 76, 430

getCookie method, 63, 431

getCookies method, 63, 237, 240, 431

getCountry method, 345, 356

getDelim method, 435

getExclude method, 395, 410, 439

getFile method, 76, 430

getForms method, 131, 161–162, 178, 437

getHREF method, 132, 167, 429

getHTTP method

- in CatBot, 345, 357
- in HTMLPage, 131, 161, 437
- in SpiderWorker, 444

GetImage.java program, 139–146

getImages method, 131, 144, 161, 437

getInputStream method, 34

getLatestDate method, 336–337, 339

getLevel method, 76, 430

getLinks method, 131–132, 161, 437

getList method, 376, 380–381

getMaxBody method

- in HTTP, 431
- in Spider, 259, 306
- in SpiderDone, 442

getMethod method, 178, 193, 436

getMyInit method, 284–286

getName method

- in Attribute, 66, 435
- in HTMLTag, 438

getOptions method, 436

getOutputStream method, 15, 34

getPackageID method, 353

getParseDelim method, 238, 242, 438

## 504     getParseName method—handleText method

getParseName method, 238, 242, 439  
 getParser method, 158  
 getParseValue method, 238, 242, 439  
 getPassword method, 63, 431  
 getPath method, 76, 430  
 getPerminantCookies method, 63, 431  
 getPrompt method  
     in FormElement, 436  
     in Link, 429  
 getPWD method, 345, 356  
 getRecognizers method, 345, 357  
 getReferrer method, 63, 431  
 getRemoveQuery method  
     in ISpiderReportable, 257, 308, 441  
     in Spider, 260, 304  
     in SpiderDone, 442  
     in UpdateTarget, 270, 406  
 getRobotFile method, 395, 410, 439  
 getSearch method, 367  
 getServerHeaders method, 64, 431  
 GetSite.java program, 261–275, 261, 295, 397–407  
 GetSite\_windowClosed method, 271, 407  
 getSpiderDone method  
     in Spider, 260, 301  
     in SpiderDone, 442  
 getSSLSocket method, 115–117  
 getTag method, 438  
 getTemp method, 374–375, 377–378, 380–381  
 getText method, 33  
 getTimeout method, 64, 431  
 getType method, 179, 436  
 getUID method, 345, 355–356  
 GetURL.java program, 58–61  
 getURL method  
     in HTMLPage, 131, 162, 437  
     in HTTP, 64, 431  
 getURLStatus method  
     in IWorkloadStorable, 309, 441  
     in SpiderInternalWorkload, 317–318, 443  
     in SpiderSQLWorkload, 314, 443  
 getUseCookies method, 64, 431  
 getUser method, 64, 431  
 getValue method, 66, 435  
 getWorkload method  
     in Spider, 260, 299, 302  
     in SpiderDone, 442  
 getWorldSpider method  
     in Spider, 260, 303  
     in SpiderDone, 442  
 GgetParser method, 437

Go\_actionPerformed method  
     in GetImage, 143–144  
     in GetSite, 266, 272–273, 402–403  
     in SecureGET, 104–106  
     in ViewURL, 73–74  
     in ViewURLCookie, 234–235  
     in WatchBBS, 335  
 Google search engine, 179, 392  
 Googlebot bot name, 392  
 gopher scheme, 47  
 gopherProxy settings, 21  
 greater than signs (>)  
     in HTML, 121  
     in XML, 211  
 gsBusy method, 444  
 gzip archives, 466  
 GZIPInputStream class, 18  
 GZIPOutputStream class, 18

## H

halt method  
     in Spider, 260, 272, 305  
     in SpiderDone, 442  
 hammering, **387**  
     avoiding, **388–389**  
     defined, 491  
     types of, **387–388**  
 handleComment method  
     as callback method, 133  
     in Parser, 163  
     in Translate, 150, 155  
 handleEndTag method  
     as callback method, 134  
     in Parser, 163  
     in Translate, 151–152, 155  
 handleError method  
     as callback method, 134  
     in Parser, 163  
 handleSimpleTag method  
     as callback method, 134, 138  
     in Parser, 163–164  
     in Translate, 152, 155  
 handleStartTag method  
     as callback method, 134, 138  
     in Parser, 164–165  
     in Translate, 151, 155  
 handleText method  
     as callback method, 134  
     in Parser, 165  
     in Translate, 152, 155–156

- hard coding
  - defined, 491
  - for forms, 172
- has method, 347, 361
- HEAD requests, 37, 52, **56–57**
- headers
  - in Cookie requests, 225
  - HTTP, 51, 54
    - for bot identification, **390–391**
    - defined, 491
    - list of, **450–452**
  - in POST requests, 55
  - in QIF files, 205
  - Set-Cookie, 224–225
  - table, 126–127
- HELO command, 24
- hexadecimal numbers
  - defined, 491
  - in IP addresses, 6
- hidden controls, **176**
- hierarchical storage in XML, 210, 212, **416–417**
- host headers, 80, 451
- hostnames
  - defined, 491
  - in network programming, **6–8**
  - resolving to IP addresses, **8–11**
  - in URLs, 49
- hosts in URI format, 48
- HotBot, bot names for, 392
- HREF (hypertext references), **249**
  - defined, 491
  - external links, **250–251**
  - internal links, **249–250, 250**
- href attribute
  - calling, 132
  - for hyperlinks, 123
  - for JavaScript, 128
- .htaccess file, 95–97
- HTML (Hypertext Markup Language), 46, **120**
  - callback methods for, **133–134**
  - character constants in, **455–456**
  - comments in, **121–122**
  - defined, 491
  - embedded objects in, **130**
  - extracting images in, **139–146, 139**
  - frames in, **129–130**
  - HTML.Attribute class for, **136–138**
  - HTML.Tag class for, **134–136**
  - HTMLPage class for, **131–132, 159–166**
  - HTMLParse class for, **158**
  - with JavaScript, **127–129**
  - Link class for, **132, 166–168**
  - recognizing, **350–352**
  - tags. *See* tags
  - text in, **120–121**
  - website translation in, **146–156, 157**
  - vs. XML, **211–213, 417–418**
- HTML.Attribute class, **136–138**
- HTML.Tag class, **134–136**
- HTMLEditorKit, 158
- HTMLForm class, 171, 173, **178–179, 191–195, 436**
- HTMLForm.FormElement class, **178–179, 436–437**
- HTMLForm method, 436
- HTMLForm Vector class, 178
- HTMLPage class, **131–133, 159–166, 177–178, 437**
- HTMLPage method, 437
- HTMLParse class, **158, 437**
- HTMLParser class, 340, **438**
- HTMLTag class, **438**
- .htpasswd file, 95–96
- .htpasswd utility, 96
- HTTP (Hypertext Transfer Protocol), 2, **36–43, 40, 46**
  - address formats in, **46–47**
  - Bot package classes for, 92, **430–433, 431**
  - defined, 491
  - for forms, **172–173**
  - headers in, 51, 54
    - for bot identification, **390–391**
    - defined, 491
    - list of, **450–452**
  - vs. HTTPS, **86–87**
  - requests in, **51–53, 53**
    - GET, **53–54**
    - HEAD, **56–57**
    - POST, **54–56, 55**
  - status codes, **452–455**
  - URI format in, **47–48**
  - URL class for, **57–61**
  - URL format in, **49–50**
  - URN format in, **48**
  - user authentication in, **92–94, 93–94**
    - with Bot package, **98**
    - on clients, **96–98**
    - HTTP class for, 98, **113–114**
    - server setup for, **94–96**
- HTTP class, **61, 431–433**
  - interfacing to JSSE, **116–117**
  - methods in, **62–65**
  - vs. URLConnection, **62**
  - URLs in, 124
    - for user authentication, 98, **113–114**
- http parameter, 258
- \_http property, 345
- http.proxy settings, 21

- HTTP scheme, 47
  - HTTPS (Hypertext Transfer Protocol Secure)
    - authentication in. *See* HTTP (Hypertext Transfer Protocol)
    - Base64OutputStream class for, **108–113**
    - defined, 491
    - vs. HTTP, **86–87**
    - HTTP class interfacing to JSSE, **116–117**
    - with Java, **87–92**, 91
    - securing access with, **98–107**, 99
    - SSL class for, **114–116**
  - https.proxy settings, 21
  - HTTPS scheme, 47
  - HTTPSocket class, 61, **66**, **77**, 106, **433**
    - lowLevelSend method in, **78–83**
    - send method in, **77–78**
    - for Weather, 377
  - hyperlinks. *See* links
  - Hypertext Markup Language. *See* HTML (Hypertext Markup Language)
  - hypertext references (HREF), **249**
    - defined, 491
    - external links, **250–251**
    - internal links, **249–250**, 250
  - Hypertext Transfer Protocol. *See* HTTP (Hypertext Transfer Protocol)
- 
- I/O programming, **11**
    - filter streams in, **17–19**
    - input streams in, **15–17**
    - output streams in, **11–15**
  - IAB (Internet Architecture Board), 9
  - id method, 67
  - IDEs, JSSE for, **90–91**, 91
  - IDs for aggregators, 370–371
  - IE (Internet Explorer)
    - cookies in, **223–224**, 223–224
    - with proxy servers, 19–20, 20
    - XML in, 415, 416
  - IETF (Internet Engineering Task Force), 9
  - If-Match header, 451
  - If-Modified-Since header, 451
  - If-None-Match header, 451
  - If-Unmodified-Since header, 451
  - image maps, **124–125**
  - images
    - extracting, **139–146**, 139
    - listing, 131
  - ImplementRunnable class, 281–282
    - importing JDBC, 292
    - InetAddress class, 9–11
    - Informational status codes, **453**
    - Infoseek Sidewinder spider, 392, 491
    - initial value of form controls, 173
    - inline aggregation, **370–371**
    - input streams
      - closing, **17**
      - creating, **15–16**
      - using, **16–17**
    - input tags, 125, 171
    - InputStream class, 15–17, 22, 60
    - INSERT statement, **288**
    - installing
      - JAXP, **213**
      - Tomcat, **466–469**, 468
    - intelligent agents, 491
    - interfaces, 492
    - internal links, **249–250**, 250, 492
    - internalPerform method
      - in Recognize, 347, 358, 360, 362
      - in RecognizeCountry, 364
      - in RecognizeLink, 368
      - in RecognizePackagePage, 350–352
    - Internet addressing for HTTP, 46
    - Internet Architecture Board (IAB), 9
    - Internet Engineering Task Force (IETF), 9
    - Internet Explorer (IE)
      - cookies in, **223–224**, 223–224
      - with proxy servers, 19–20, 20
      - XML in, 415, 416
    - Internet information transfer, **414–415**
    - Internet Options dialog box, 223
    - Internet Protocol (IP), 5, 492
    - Internetwork Packet Exchange (IPX) protocol, 3
    - invokeLater method, 272–273
    - IOException class, 34
    - IP (Internet Protocol), 5, 492
    - IP addresses
      - in bot detection, 394
      - filtering, 395
      - and hostnames, **6–8**
      - in network programming, **5–6**
      - resolving hostnames to, **8–11**
    - IPX (Internetwork Packet Exchange) protocol, 3
      - isBlock method, 135
      - isBusy method, 319–320, 444
      - isEmpty method, 68, 435
      - isExcluded method, 396–397, 409, 439
      - isHalted method
        - in Spider, 305
        - in SpiderDone, 442

isolated hammering, **387–388**, 492  
 ISpiderReportable interface, **256–258**, 297, 299,  
**306–308**, **440–441**  
 isPreformatted method, 135  
 isRecognizable method  
   in Recognize, 347, 358–359  
   in RecognizeCountry, 363–364  
   in RecognizeLink, 367  
   in RecognizePackagePage, 350–352  
 isRecognized method  
   in Recognize, 347, 360  
   in RecognizeCountry, 363  
   in RecognizeLink, 366  
 isWhiteSpace method, 238, 240, 439  
 IWorkloadStorable interface, 298, **308–309**, **441**

## J

Jakarta-tomcat directory, 466  
 Java API for XML Messaging (JAXM), 423  
 Java applets, 130  
 Java Database Connectivity (JDBC), 278, 287,  
**289–290**, **292**  
   Connection object in, **293**  
   defined, 492  
   ResultSet object in, **294–295**  
   Statement object in, **293–294**  
 JAVA\_HOME environmental variable, 466  
 java.io.InputStream class, 15  
 java.io.OutputStream class, 11  
 java.lang package, 280  
 Java language, 492  
 Java Messaging API (JAXM), 423, 492  
 java.net.URL class, 87, 124  
 Java Secure Socket Extension (JSSE)  
   defined, 492  
   for IDEs, **90–91**, 91  
   interfacing to HTTP class, **116–117**  
   for JDK 1.3 and lower, **88–90**  
   for SSL and HTTPS, 87  
   working with, **91–92**  
 Java Server Pages (JSP)  
   defined, 492  
   running, **348–349**, 348  
   support for, 466  
 Java Virtual Machine (JVM), 20  
 javac.exe file, 473, 481  
 Javadoc utility, 492  
 JavaScript language, 120–121  
   defined, 492  
   HTML with, **127–129**

JAXM (Java API for XML Messaging), 423, 492  
 JAXP, installing, **213**  
 JButton1\_actionPerformed method, 188–190  
 JButton2\_actionPerformed method, 187–188  
 JDBC (Java Database Connectivity), 278, 287,  
**289–290**, **292**  
   Connection object in, **293**  
   defined, 492  
   ResultSet object in, **294–295**  
   Statement object in, **293–294**  
 JDBC to ODBC bridges, 289, 293  
 JDK, **472**  
   bin directory for, **473**, **481**  
   bot.jar file for, **472–473**, **480–481**  
   CLASSPATH and system path for, **473–476**,  
   474–476, **481–482**  
   JSSE for, **88–90**  
 JSP (Java Server Pages)  
   defined, 492  
   running, **348–349**, 348  
   support for, 466  
 JSSE (Java Secure Socket Extension)  
   defined, 492  
   for IDEs, **90–91**, 91  
   interfacing to HTTP class, **116–117**  
   for JDK 1.3 and lower, **88–90**  
   for SSL and HTTPS, 87  
   working with, **91–92**  
 JVM (Java Virtual Machine), 20

## L

LANs (local area networks), 414  
 Last-Modified header, 451  
 legal measures against bots, 395  
 length method, 68, 435  
 less than signs (<)  
   in HTML, 121  
   in XML, 211  
 Link class, **132**, **166–168**, **429**  
 Link method, 429  
 links  
   defined, 492  
   external, **250–251**  
   format of, **123**  
   internal, **249–250**, 250  
   listing, 131  
   recognizers for, 343  
   sources of, **251**  
   URLs for, **123–124**

**508** listening—openStream method

listening  
   defined, 492  
   to ports, 41  
 load method, 396, 408–410, 439  
 loading X-Windows, 460–461  
 Local Area Network (LAN) Settings dialog box, 20, 20  
 local area networks (LANs), 414  
 Location header, 451  
 lock symbol, 86  
 locking objects, **283–286**  
 Log class, **74–76, 430**  
 log files for spiders, 274  
 Log.LOG\_LEVEL levels, 75  
 log method  
   in Log, 75–76, 430  
   in SiteSubmit, 188  
 logException method, 76, 430  
 logging levels, 75–76  
 Lookup.java program, 9–10  
 lookup method, 352–353  
 lowLevelSend method  
   in HTTP, 431  
   in HTTPSocket, **78–83**, 116, 433

---

**M**

---

MAIL FROM command, 24–25  
 mailto scheme, 47, 251  
 MalformedURLException class, 57, 60–61  
 manager object, 299  
 manager parameter, 258  
 mark method, 15–17  
 markSupported method, 15  
 Max-Age attribute, 225–226  
 Max-Forwards header, 451  
 MDAC (Microsoft Data Access) package, 290  
 memory-based queues, 295  
 menus, **175–176**  
 method attribute, 194  
 method calls in SOAP, 420  
 Microsoft Access, **290–291**, 290–291  
 Microsoft Data Access (MDAC) package, 290  
 MIME-Version header, 451  
 minus signs (-) in HTML, 121  
 monitoring spider progress, **275**  
 Mozilla code word, 391, 493  
 multitasking, 278, 493  
 multithreading, **278–279**  
   defined, 493  
   and recursion, 252  
   with spiders, 258

threads in  
   creating, **279–281**  
   starting, **282**  
   suspending and resuming, **282–283**  
   synchronizing, **283–286**  
 MutableAttributeSet class, 138  
 MySynchronizedObject class, 284–286

---

**N**

---

name attribute for Set-Cookie header, 225  
 name-value pairs  
   for cookies, 224  
   defined, 493  
   in XML, 212  
 names, 66–67  
 namespaces in XML, 421  
 National Center for Supercomputing Applications (NCSA), 493  
 National Weather Service (NWS), **372**, 390  
 .NET, 423  
 network programming  
   hostnames in, **6–11**  
   IP addresses in, **5–6, 8–11**  
   ports and services in, **4–5**  
 new operator, 14–15  
 news scheme, 47  
 NOBOT scripts, 463  
 nodes in XML, 416  
 notify method, 286  
 notifyAll method, 286  
 NT Service, 291  
 NWS (National Weather Service), **372**, 390

---

**O**

---

object locks  
   defined, 493  
   in synchronizing threads, **283–286**  
 ODBC (Open Database Connectivity), **289–290**, 493  
 ODBC Data Source Administrator, 290, 290  
 ODBC Microsoft Access Setup dialog box, 291, 291  
 offline aggregation, **371**  
 Ok\_actionPerformed method, 101  
 onclick attribute, 175  
 online aggregation, **380–382**  
 Open Database Connectivity (ODBC), **289–290**, 493  
 open method, 131–132, 144, 160, 437  
 open standard, XML as, **419**  
 opening connections, **58–61**  
 openStream method, 60

Optical Character Recognition, 493  
 other links, 493  
 out object, 34  
 output streams, **11**  
   buffering in, **13–14**  
   closing, **14–15**  
   creating, **12–13**  
   using, **13**  
 OutputStream class, 12, 17, 22

## P

packages, 493  
 packets  
   defined, 493  
   in HTTP, 36–37  
   in peer-to-peer networks, 4  
 pages, recognizers for, 343  
 parent directories in URLs, 124  
 Parse class, **237–242, 438–439**  
 parseAttributeName method, 238, 241, 439  
 parseAttributeValue method  
   in CookieParse, 239, 243, 436  
   in Parse, 238, 241, 439  
 parseCookies method, 431  
 ParseCSV.java program, 199–203  
 parseCSVLine method, 200–203  
 parseDate method, 337–338, 341  
 parseHeaders method, 82–83, 431  
 ParseQIF.java program, 206–209  
 parseQIFLine method, 207–208  
 Parser class, 162–165  
 ParserCallback class, 133  
 parsers  
   defined, 493  
   for forms, **125–126, 126**  
   HTML. *See* HTML (Hypertext Markup Language)  
   for URLs, 57  
 parseTag method, 438  
 ParseXML.java program, 214–220  
 parsing classes, **433, 434**  
   Attribute, **435**  
   AttributeList, **435**  
   CookieParse, **436**  
   HTMLForm, **436**  
   HTMLForm.FormElement, **436–437**  
   HTMLPage, **437**  
   HTMLParse, **437**  
   HTMLParser, **438**  
   HTMLTag, **438**  
   Parse, **438–439**

passwords  
   for aggregators, 370–371  
   in Apache server, 95–97  
   in HTTP, 63–65, 93, 98, 113–114  
   prompts for, 99–101, 106–107  
   in URI format, 48  
 patents, 493  
 Path attribute, 225–226  
 path environmental variable, 459  
 paths  
   for compiler, 458–460, *460*  
   in HTTP requests, 53  
   for JDK, **473–476, 481–482**  
   for logging, 76  
   in URI format, 48  
   in URLs, 49  
 pattern recognition, 493  
 payees in QIF files, 205  
 peer-to-peer networks, 3–4, 493  
 perform method, 347, 360  
 periods (.)  
   in IP addresses, 5  
   in URLs, 124  
 persistent cookies, **229, 494**  
 persistent objects, 493  
 Pig Latin translations, 146–147  
 pigLatin method, 152–153, 156  
 Ping command, 7–9  
 PINs for aggregators, 370–371  
 pipes (|) in CSV files, 198  
 polling frequency  
   defined, 494  
   in hammering, **388–389**  
 poolsize parameter, 259  
 ports  
   defined, 494  
   for HTTPS, 92  
   in network programming, **4–5**  
   for Tomcat, 467  
   in URLs, 49  
 post method, 132, 162, 437  
 POST requests, 37, 52, **54–56, 55, 80–81**  
   defined, 494  
   forms for, 170  
 posting to forms, 55–56, 55, **172–173, 191**  
 pound signs (#) in exclusion files, 394  
 Pragma header, 452  
 prepared statements, 293, 494  
 PreparedStatement class, 293  
 prime recognizers, 343, *344*  
   \_primeRecognizer property, 345  
 print method, 24

**510**    println method—reset buttons

println method, 24  
 printlnCommon method, 215–216  
 PrintWriter class, 18, 22, 34  
 processes, 494  
 processFile method, 265–266, 275, 401–402  
 processImage method, 142–145  
 processor time, 494  
 processPage method  
   in HTMLPage, 160–161, 165, 437  
   in ISpiderReportable, 258, 307, 441  
   in Spider, 260, 275, 304  
   in SpiderDone, 442  
   in UpdateTarget, 269–270, 405  
 processResponse method, 431  
 processWorkload method, 320–322, 444  
 progress, spider, 275  
 Project Options dialog box, 90  
 protocols, 2, 46, 494  
 proxies  
   defined, 494  
   issues with, 19–21, 20  
 Proxy-Authenticate header, 452  
 Public header, 451  
 push buttons, 175  
 putMyInit method, 284–286  
 \_pwd property, 345

**Q**

QIF (Quicken Interchange Format) file  
   format of, 204–205  
   parsing example, 206–209  
   structure of, 203–204  
 QIFElement class, 209  
 queries in URI format, 48  
 query strings  
   defined, 494  
   for HTTP user authentication, 94  
 question marks (?)  
   in SQL, 293  
   in URI format, 48  
   in URLs, 195  
 queues  
   defined, 494  
   in spiders, 254–256, 254–255, 295  
 Quicken financial package, 371  
 quotes ("")  
   for attributes, 174  
   in CSV files, 198  
   in XML, 212, 418

**R**

radio buttons, 175  
 raw posting to forms, 191  
 RCPT TO command, 24  
 read method  
   in FilterInputStream, 18  
   in InputStream, 15–16  
 readability of XML, 417–418  
 readers  
   defined, 494  
   in I/O programming, 17–19  
 readLine method, 34–35, 60  
 Recognize class, 346–347, 358–362  
 RecognizeCountry class, 343, 362–364  
 RecognizeLink class, 343, 365–368  
 RecognizePackagePage class, 343, 350–351  
 recognizers, 342–343, 344, 350–352  
   Recognize class, 346–347, 358–362  
   RecognizeCountry class, 362–364  
   RecognizeLink class, 365–368  
 \_recognizers property, 345  
 recompiling bot package, 484–485  
 recursive programming, 252–253  
 RecursiveSpider method, 252  
 redirect command, 494  
 redirection  
   in HTTP, 64–65  
   send for, 77–78  
   status codes for, 454  
 redirection attribute, 64–65  
 Referrer header, 452  
 referrer tag, 61–62  
 referrers  
   defined, 494  
   HTTP, 61–63, 65  
 relative URLs, 50  
   defined, 494  
   for hyperlinks, 123–124  
 requests  
   in Apache server, 97  
   HTTP, 36–37, 51–53, 53  
     GET, 53–54  
     HEAD, 56–57  
     POST, 54–56, 55, 80–81  
 Requests for Comments (RFCs), 4–5  
   for communications protocols, 46  
   defined, 495  
   and socket protocols, 22–23  
 reset buttons, 174

reset method  
 in `InputStream`, 15–17  
 in `SpiderDone`, 325, 442  
`resolveBase` method, 430  
`resolveLink` method, 128  
 resolving IP addresses to hostnames, **8–11**  
 resource leaks, 14, 17  
 resources  
   defined, 495  
   in URI format, 48  
 response code, 452  
 response packets, 36–37  
 result sets, **294–295**, 495  
`ResultSet` class, **294–295**  
`resume` method, 282  
 resuming threads, **282–283**  
 Retry-After header, 452  
 reverse DNS lookup, **10–11**, 495  
 RFCs (Requests for Comments), 4–5  
   for communications protocols, 46  
   defined, 495  
   and socket protocols, **22–23**  
 Rhino language, 129  
 robots.txt file, 393–395, 410–411  
 root documents, 249, 495  
 rows in tables, 127  
 RSA patent, **87–88**  
 run method  
   in `GetSite`, 267, 274  
   in `ImplementRunnable`, 282–283  
   in `Runnable`, 279–281  
   in `Spider`, 260, 301  
   in `SpiderDone`, 442  
   in `SpiderWorker`, 320, 444  
   in `UpdateTarget`, 406  
   in `WatchBBS`, 338–339  
`Runnable` interface, **279–281**  
 running JSP pages, **348–349**, 348  
 running queues, 254  
 RUNNING status code, 308

## S

saving pages with spiders, 275  
 schemes  
   defined, 495  
   in URI format, 47  
   in URLs, 49  
 Scooter spider, 392, 495  
 scripts, 121–122

search engines  
   defined, 495  
   submitting sites to, **179–191**, 180  
 Secure attribute, 225  
 Secure Socket Layer (SSL) protocol  
   defined, 495  
   for HTTPS, 80  
   JSSE for, 87, **114–116**  
`SecureGET.java` program, 102–107  
`SecurePrompt` class, 98–101  
 security  
   for aggregators, 370–371  
   with HTTPS. *See* HTTPS (Hypertext Transfer Protocol Secure)  
 SELECT statement, **287–288**  
 semicolons (;)  
   for cookies, 224, 226  
   in CSV files, 198  
   in URLs, 195  
 send method  
   in HTTP, 64, 113, 431  
   in `HTTPSocket`, **77–78**, 92  
   in `SendMail`, 31–32, 34–35  
`Send_actionPerformed` method, 32  
`SendMail.java` program, 26–33  
 Server Error status codes, **455**  
 server farms, 7  
 server headers, 452  
   defined, 495  
   listing, 64  
 server-side image maps, 124  
 server sockets, **36–43**, 40  
 servers  
   defined, 495  
   for HTTP user authentication, **94–96**  
   in peer-to-peer networks, 4  
`ServerSocket` class, 21–22, 36, 40–41  
 services in network programming, **4–5**  
 servlets, 495  
 session cookies, **229**, 495  
 sessions, 495  
 set classpath command, 474  
 set-cookie command, 224, 226–228  
 set method, 68, 435  
 set operations in synchronizing threads, 283–284  
`setAgent` method, 64, 391, 394, 431  
`setAsciiStream` method, 294  
`setAutoRedirect` method, 64–65, 431  
`setBigDecimal` method, 294  
`setBinaryStream` method, 294  
`setBoolean` method, 294  
`setByte` method, 294

## 512 setBytes method—Socket class

- setBytes method, 294
- setConsole method, 75–76, 430
- setCountry method, 346, 356
- setDate method, 294
- setDelim method, 435
- setDouble method, 294
- setFile method, 75–76, 430
- setFloat method, 294
- setInt method, 294
- setLevel method, 76, 430
- setLong method, 294
- setMaxBody method
  - in HTTP, 431
  - in Spider, 260, 274, 306
  - in SpiderDone, 442
- setName method
  - in Attribute, 67, 435
  - in HTMLTag, 438
- setNextPoll method, 65, 98, 113, 433
- setNull method, 294
- setObject method, 294
- setOptions method, 437
- setParseDelim method, 238, 242, 439
- setParseName method, 238, 242, 439
- setParseValue method, 239, 242, 439
- setPassword method, 65, 98, 113, 433
- setPath method, 76, 430
- setPrompt method
  - in FormElement, 437
  - in Link, 429
- setProperty method, 91–92
- setPWD method, 346, 356
- setReferrer method, 65, 433
- setSearch method, 367
- setShort method, 294
- setStatus method, 313–314, 443
- setString method, 293–294
- setTime method, 294
- setTimeout method, 65, 433
- setTimestamp method, 294
- setType method
  - in FormElement, 179, 437
  - in HTMLForm, 194
- setUID method, 346, 355
- setUnicodeStream method, 294
- setURL method
  - in CatBot, 346, 357
  - in HTTP, 65, 77, 433
- setUseCookie method, 229
- setUseCookies method
  - in HTTP, 65, 433
  - in HTTPSock, 223
- setUser method, 65, 98, 113, 433
- setValue method, 67, 435
- setVisible method
  - in GetImage, 140–141
  - in GetSite, 263, 399
  - in SecureGET, 103
  - in SecurePrompt, 100
  - in SendMail, 28
  - in SiteSubmit, 184
  - in ViewURL, 70
  - in ViewURLCookie, 232
  - in WatchBBS, 332
- setWorldSpider method
  - in Spider, 260, 302–303
  - in SpiderDone, 442
- SGML (Standard Generalized Markup Language)
  - defined, 496
  - as XML basis, **418–419**
- ship.jsp program, 348–349
- ShipBot class, **352–353**
- Shockwave files, 130
- Simple Mail Transfer Protocol (SMTP), **23–36**, 26, 421, 495
- Simple Object Access Protocol. *See* SOAP (Simple Object Access Protocol)
- simple tags in HTML, **122**
- Site class, 185–186
- SiteSubmit.java program, 181–191
- skip method, 15–16
- slashes (/)
  - in URLs, 50, 124, 195
  - in XML, 211–212
- sleep method, 282
- Slurp spider, 392, 495
- SMTP (Simple Mail Transfer Protocol), **23–36**, 26, 421, 495
- \_smtp class, 33–34
- SNA (Systems Network Architecture), 3
- sneakernet, 414
- SOAP (Simple Object Access Protocol), **419–420**
  - for bots, **424**
  - data transfer in, **420–421**
  - defined, 495
  - in Java, **423**
  - structure of, **421–423**
  - WSDL format for, **423**
- SOAP-ENV:Body attribute, 420, 422
- SOAP-ENV:encodingStyle attribute, 420
- SOAP-ENV:Envelope attribute, 420–421
- SOAPRequest.txt file, 421–422
- SOAPResponse.txt file, 422
- Socket class, 21–22, 36, 41

- sockets, 2–3
  - client, 23–36, 26
  - defined, 496
  - in I/O programming, 11–19
  - in network programming, 4–11
  - programming, 21–23
  - proxy issues in, 19–21, 20
  - server, 36–43, 40
  - in TCP/IP networks, 3–4
- spaces in HTTP requests, 53
- Spider class, 258–261, 297–306, 441–442
- Spider classes, 439, 440
  - BotExclusion, 439
  - ISpiderReportable, 440–441
  - IWorkloadStorable, 441
  - Spider, 441–442
  - SpiderDone, 442
  - SpiderInternalWorkload, 443
  - SpiderSQLWorkload, 443
  - SpiderWorker, 443–444
- Spider method, 442
- spiderComplete method
  - in ISpiderReportable, 258, 308, 441
  - in Spider, 261, 305
  - in SpiderDone, 442
  - in UpdateTarget, 270–271, 406
- SpiderDone class, 298–299, 322–325, 442
- SpiderInternalWorkload class, 296, 298, 315–318, 443
- spiders, 248
  - BotExclusion class for, 439
  - canceling, 272
  - conscientious, 395–412
  - databases for, 287
    - JDBC for, 292–295
    - selecting and configuring, 289–292, 290–291
    - SQL language for, 287–289
  - defined, 496
  - example, 261–271, 261
  - high-performance, 295–325, 297
  - ISpiderReportable interface for, 256–258, 297, 306–308, 440–441
  - IWorkloadStorable interface for, 298, 308–309, 441
  - monitoring progress of, 275
  - multithreading in. *See* multithreading
  - non-recursive programming in, 253–256, 254–255
  - queues in, 254–256, 254–255, 295
  - recursive programming in, 252–253
  - setting up, 272
  - Spider class for, 258–261, 297–306, 441–442
  - SpiderDone class for, 298, 322–325, 442
  - SpiderInternalWorkload class for, 298, 315–318, 443
  - SpiderSQLWorkload class for, 298, 310–315, 443
  - SpiderWorker class for, 298, 318–322, 443–444
  - starting, 273–274
  - structure of, 251
  - workload management in, 274–275, 298
- SpiderSQLWorkload class, 298, 310–315, 443
- SpiderSQLWorkload manager, 296
- SpiderSQLWorkload method, 443
- SpiderWorker class, 298–299, 318–322, 443–444
- SpiderWorker method, 444
- SpiderWorkload.mdb database, 290–291
- spy utilities, 51
- SQL (Structured Query Language) language
  - defined, 496
  - statements in, 287–289
- SQL-based queues, 295
- src directory, 484
- SSL (Secure Socket Layer) protocol
  - defined, 495
  - for HTTPS, 80
  - JSSE for, 87, 114–116
- SSL class, 114–116
- SSL.java program, 114–116
- SSLConnectionFactory class, 116
- stacks in recursive programming, 252–253
- Standard Generalized Markup Language (SGML)
  - defined, 496
  - as XML basis, 418–419
- standardRecognition method, 346, 352–353, 357–358
- start method
  - in Spider, 322
  - in Thread, 282
  - in WebServer, 38
- Start\_actionPerformed method, 334
- starting threads, 282
- startup.sh script, 467
- startx command, 461
- stateless connections, 222, 496
- Statement object, 293–294
- states
  - defined, 496
  - URL, 254, 254
- status codes, HTTP, 452–455
- status field, 292
- stop method, 283
- Stop\_actionPerformed method, 335
- streams
  - defined, 496
  - filter, 17–19
  - input, 15–17
  - output, 11–15
- StreamTokenizer class, 199
- String class, 199

## 514 StringBuffer class—Transmission Control Protocol/Internet Protocol (TCP/IP)

- StringBuffer class, 199, 239
  - StringTokenizer class, 199
  - stripAnchor method, 430
  - stripQuery method, 430
  - Structured Query Language (SQL) language
    - defined, 496
    - statements in, **287–289**
  - submit buttons, 174
  - submitting sites to search engines, **179–191, 180**
  - Successful status codes, **453**
  - suspend method, 282
  - suspending threads, **282–283**
  - SymAction class
    - in GetImage, 142
    - in GetSite, 265, 401
    - in SecureGET, 104
    - in SecurePrompt, 101
    - in SendMail, 31
    - in SiteSubmit, 187
    - in WatchBBS, 334
  - symbols, resolving, 462
  - SymWindow class
    - in GetSite, 271, 407
    - in ViewURL, 72–73
    - in ViewURLCookie, 233–234
  - synchronized keyword, 283
  - synchronizing threads, **283–286**
  - syntax
    - defined, 496
    - HTML, 120
  - System DSN tab, 290, 291
  - system path
    - for compiler, 458–460, *460*
    - defined, 496
    - for JDK, **473–476, 474–476, 481–482**
  - System Properties panel, 474–475, *475*
  - Systems Network Architecture (SNA), 3
- T**
- 
- table tag, 126–127
  - tables in HTML, **126–127, 127**
  - tags
    - in HTML, **122**
      - beginning and ending, **122**
      - for forms, **125–126, 126**
      - for hyperlinks, **123–124**
      - for image maps, **124–125**
      - simple, **122**
      - for tables, **126–127, 127**
    - in XML, 211–212, 418
  - tblWorkload table, 292
  - TCP (Transmission Control Protocol), 5
  - TCP/IP (Transmission Control Protocol/Internet Protocol), 2, 496
  - TCP/IP networks, sockets in, **3–4**
  - td tag, 126–127
  - TELNET protocol, 52
  - Telnet scheme, 47
  - terms of service (TOS)
    - defined, 496
    - in websites, **389–390**
  - terms of use, 386
  - testfile.qif file, 206
  - testfile.xml file, 210–211, *211*, 214–215
  - text files for XML, 417
  - text in HTML, **120–121**
  - text input controls, **176**
  - TEXTAREA controls, 174
  - th tag, 126–127
  - thick clients for aggregators, 371
  - thin clients for aggregators, 370–371
  - Thread class, **279–280**
  - thread pools, 322, 496
  - threading, 496
  - threads
    - creating, **279–281**
    - defined, 496
    - starting, **282**
    - suspending and resuming, **282–283**
    - synchronizing, **283–286, 496**
  - timeouts in HTTP, 64–65
  - Tomcat web server
    - installing, **466–469, 468**
    - for JSP pages, 348
  - TomCatClasses directory, 468
  - TOS (terms of service)
    - defined, 496
    - in websites, **389–390**
  - toString method
    - in CookieParse, 239, 243–244, 436
    - in HTML.Tag, 135
    - in HTMLForm, 179, 193–194, 436
    - in Link, 132, 167–168, 429
  - tr tag, 127
  - TracePlus 32/Web Detective utility, 51
  - translate.java program, 148–156, *157*
  - translate.jsp file, 148–149, 469
  - translate method, 149, 155
  - translations for websites, **146–156, 157**
  - Transmission Control Protocol (TCP), 5
  - Transmission Control Protocol/Internet Protocol (TCP/IP), 2, 496

troubleshooting  
 cross-platform errors, **461–463**  
 UNIX errors, **458–461**, *460*  
 WIN32 errors, **458**  
 try blocks, 41

## U

`_uid` property, 345  
 Ultimate Bulletin Board system, 329, 329  
 UML (Uniform Modeling Language) format, 428  
 Uniform Resource Identifiers (URIs)  
   defined, 497  
   in HTTP, **47–48**  
 Uniform Resource Locators. *See* URLs  
 unique bot identification, 392  
 UNIX operating system  
   compiling under, **480–482**  
   with spiders, 292  
   troubleshooting, **458–461**, *460*  
 UNKNOWN status code, 308  
 unsafe characters in URLs, **195**  
 UPDATE statement, **288–289**  
 UpdateTarget class, 269, 405  
 URIs (Uniform Resource Identifiers)  
   defined, 497  
   in HTTP, **47–48**  
 URL class, 61, 87, 124  
   constructor for, **57–58**  
   for opening connections, **58–61**  
 URL encoding, 497  
 URL field, 292  
 url parameter  
   in isExcluded, 396  
   in Spider, 258  
`_url` property, 345  
 URLConnection class, **61–62**  
 URLs (Uniform Resource Locators), **49**  
   components of, **49–50**  
   defined, 497  
   vs. hostnames, 6  
   for hyperlinks, **123–124**  
   relative, **50**  
   retrieving, 131  
   states of, 254, 254  
   unsafe characters in, **195**  
 URLUtility class, **430**  
 URLUtility method, 430  
 URN format, **48**  
 usenet postings for bot defamation, 395  
 UseProxy class, 21

User-agent command line, 394  
 User-Agent header, 390–391, 452  
 user agent names, detection of, 394  
 user authentication. *See* HTTP (Hypertext Transfer Protocol)  
 User DSN tab, 291  
 user IDs  
   in Apache server, 95–97  
   in HTTP, 93, 98, 113–114  
   prompts for, 99–101, 106–107  
 usernames in URI format, 48  
 users in HTTP, 64–65  
 utility classes, **428**, *429*  
   Base64OutputStream, **428**  
   Link, **429**  
   Log, **430**  
   URLUtility, **430**

## V

value attribute, 173–174  
 values  
   of form controls, 173–174  
   in XML, 418  
 Vary header, 452  
 Version attribute, 225  
 vertical bars (|) in CSV files, 198  
 ViewURL.java program, 69–74  
 ViewURL\_windowClosed method, 74  
 ViewURL\_WindowClosing method, 73  
 ViewURLCookie program, 230–237, 230  
 ViewURLCookie\_windowClosed method, 235  
 ViewURLCookie\_WindowClosing method, 234  
 VisualCafé  
   compiling with, **477**, *478*  
   with Java, 90, *91*  
 volatility of site content, 342  
 voyagers, 251  
 .vpj file extension, 477

## W

w parameter, 259  
 W3C (World Wide Web Consortium), 210, 497  
 wait method, 286  
 waitBegin method, 324–325, 442  
 waitDone method, 323–325, 442  
 waiting queues, 254–255, 322  
 WAITING status code, 308  
 Warning header, 452

**516** warnings—ZipOutputStream class

warnings, 461

WatchBBS bot, **328–330**, *330–331*  
 code for, **330–339**  
 operation of, **339–341**  
 weaknesses in, **341–342**

WatchBBS\_windowClosed method, 336

Weather aggregator, **378–379**, *379*  
 Weather class for, **380–382**  
 weather.jsp page for, **379–380**

Weather bot, **371–372**  
 building, **374–378**  
 extracting data from, **374**  
 planning, **372–374**, *373*

Weather class, **380–382**

weather.jsp page, **379–380**

web applications, 497

Web Service Definition Language (WSDL), 421, 497

webfarms, 7, 497

webmaster actions  
 access curtailment, **394–395**  
 bot exclusion files, **393–394**

webmasters, 497

WebServer.java program, 38–43

websites  
 bot identification to, **390–393**  
 conscientious spiders for, **395–412**  
 dealing with, **386**  
 hammering, **387–389**  
 HREF types for, **249–251**, *250*  
 structure of, **248–249**  
 terms of service in, **389–390**  
 translations for, **146–156**, *157*  
 webmaster actions, **393–395**

WebStone bot, 387

WHERE clause  
 in DELETE, 288  
 in SELECT, 287  
 in UPDATE, 288

width attribute, 126

WIN32 errors, **458**

windowClosed method  
 in GetSite, 271, 407  
 in ViewURLCookie, 233  
 in WatchBBS, 335–336

windowClosing method  
 in ViewURL, 72–73  
 in ViewURLCookie, 234

Windows, compiling under  
 JDK for, **472–476**, *474–476*  
 VisualCafé for, **477**, *478*

Windows NT Service, 497

worker threads, 41

workerBegin method, 324–325, 442

workerEnd method, 324–325, 442

workload management in spiders, **274–275**, 298

world spiders, 251

World Wide Web Consortium (W3C), 210, 497

write method  
 in Base64OutputStream, 111–112, 428  
 in FilterOutputStream, 18  
 in OutputStream, 12–13

writers  
 defined, 497  
 in I/O programming, **17–19**

writeString method, 433

WSDL (Web Service Definition Language), 421  
 defined, 497  
 for SOAP, **423**

WWW-Authenticate header, 452

---

**X**

X-Windows, loading, 460–461

Xconfigurator program, 461

Xerces parser, 213

XML (Extensible Markup Language), 414–415  
 for bots, **419**  
 data transfer in. *See* SOAP (Simple Object  
 Access Protocol)  
 defined, 490  
 file structure in, **210–211**, *211*, **415**, *416*  
 hierarchical storage in, 210, 212, **416–417**  
 vs. HTML, **211–213**, **417–418**  
 as open standard, **419**  
 parsing example, **213–220**  
 readability of, **417–418**  
 SGML as basis for, **418–419**

---

**Y**

Yahoo! directory, 179

yield method, 282

Yodlee company, 371

---

**Z**

---

ZipInputStream class, 18

ZipOutputStream class, 18