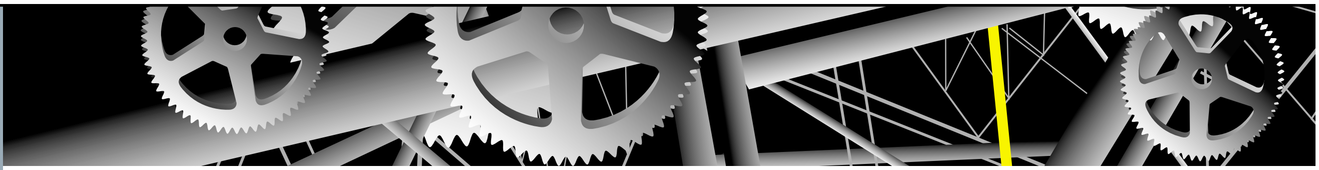
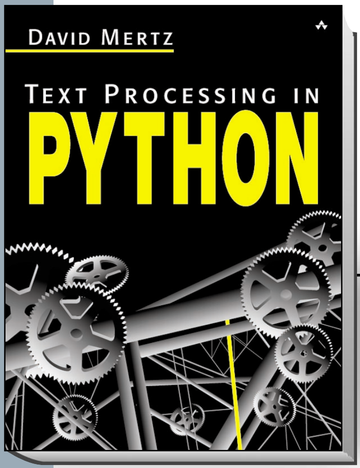




DAVID MERTZ

TEXT PROCESSING IN PYTHON

Text Processing in Python is an example-driven, hands-on tutorial that carefully teaches programmers how to accomplish numerous text processing tasks using the Python language. Filled with concrete examples, this book provides efficient and effective solutions to specific text processing problems and practical strategies for dealing with all types of text processing challenges.

Text Processing in Python begins with an introduction to text processing and contains a quick Python tutorial to get you up to speed. It then delves into essential text processing subject areas, including string operations, regular expressions, parsers and state machines, and Internet tools and techniques. Appendixes cover such important topics as data compression and Unicode. A comprehensive index and plentiful cross-referencing offer easy access to available information. In addition, exercises throughout the book provide readers with further opportunity to hone their skills either on their own or in the classroom. A companion Web site (http://gnosis.cx/TPiP) contains source code and examples from the book.

©2003, PAPER, 544 PAGES,
0-321-11254-7, $49.99

ABOUT THE AUTHOR

DAVID MERTZ came to writing about programming via the unlikely route of first being a humanities professor. Along the way, he was a senior software developer, and now runs his own development company, Gnosis Software ("We know stuff!"). David writes regular columns and articles for IBM developerWorks, Intel Developer Network, O'Reilly ONLamp, and other publications.

When do I use formal parsers to process structured and semistructured data?	**Page 257**
How do I work with full text indexing?	**Page 199**
What patterns in text can be expressed using regular expressions?	**Page 204**
How do I find a URL or an email address in text?	**Page 228**
How do I process a report with a concrete state machine?	**Page 274**
How do I parse, create, and manipulate internet formats?	**Page 345**
How do I handle lossless and lossy compression?	**Page 454**
How do I find codepoints in Unicode?	**Page 465**

Addison
Wesley

FROM ADDISON-WESLEY

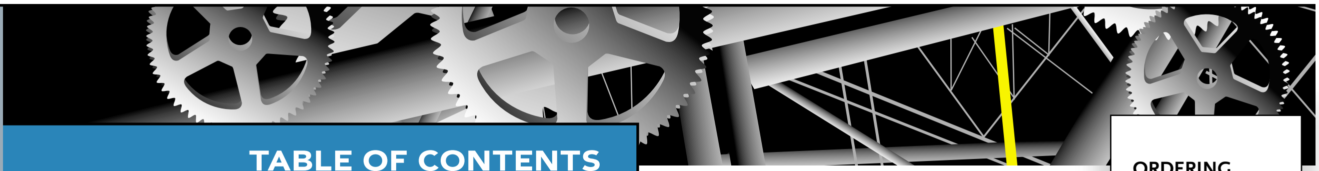
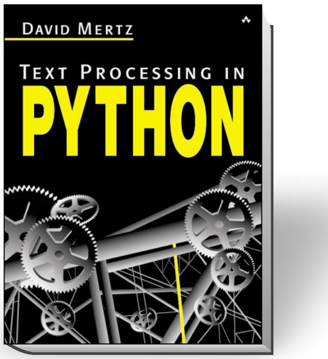


TABLE OF CONTENTS



PREFACE
ACKNOWLEDGMENTS

1 PYTHON BASICS
- Techniques and Patterns
- Utilizing Higher-Order Functions in Text Processing
- Exercise: More on combinatorial functions
- Specializing Python Datatypes
- Base Classes for Datatypes
- Exercise: Filling out the forms (or deciding not to)
- Problem: Working with lines from a large file
- Standard Modules
- Working with the Python Interpreter
- Working with the Local Filesystem
- Running External Commands and Accessing OS Features
- Special Data Values and Formats
- Other Modules in the Standard Library
- Serializing and Storing Python Objects
- Platform-Specific Operations
- Working with Multimedia Formats
- Miscellaneous Other Modules

2 BASIC STRING OPERATIONS
- Some Common Tasks
- Problem: Quickly sorting lines on custom criteria
- Problem: Reformatting paragraphs of text
- Problem: Column statistics for delimited or flat-record files
- Problem: Counting characters, words, lines, and paragraphs
- Problem: Transmitting binary data as ASCII
- Problem: Creating word or letter histograms
- Problem: Reading a file backwards by record, line, or paragraph
- Standard Modules
- Basic String Transformations
- Strings as Files, and Files as Strings
- Converting Between Binary and ASCII
- Cryptography
- Compression
- Unicode
- Solving Problems
- Exercise: Many ways to take out the garbage
- Exercise: Making sure things are what they should be
- Exercise: Finding needles in haystacks (full-text indexing)

3 REGULAR EXPRESSIONS
- A Regular Expression Tutorial
- Just What Is a Regular Expression, Anyway?
- Matching Patterns in Text: The Basics
- Matching Patterns in Text: Intermediate
- Advanced Regular Expression Extensions
- Some Common Tasks
- Problem: Making a text block flush left
- Problem: Summarizing command-line option documentation
- Problem: Detecting duplicate words
- Problem: Checking for server errors
- Problem: Reading lines with continuation characters
- Problem: Identifying URLs and email addresses in texts
- Problem: Pretty-printing numbers
- Standard Modules
- Versions and Optimizations
- Simple Pattern Matching
- Regular Expression Modules

4 PARSERS AND STATE MACHINES
- An Introduction to Parsers
- When Data Becomes Deep and Texts Become Stateful
- What Is a Grammar?
- An EBNF Grammar for IF/THEN/END Structures
- Pencil-and-Paper Parsing
- Exercise: Some variations on the language
- An Introduction to State Machines
- Understanding State Machines
- Text Processing State Machines
- When Not to Use a State Machine
- When to Use a State Machine
- An Abstract State Machine Class
- Processing a Report with a Concrete State Machine
- Subgraphs and State Reuse
- Exercise: Finding other solutions
- Specialized Parsers in the Standard Library
- Low-Level State Machine Parsing
- High-Level EBNF Parsing
- High-Level Programmatic Parsing

5 INTERNET TOOLS AND TECHNIQUES
- Working with Email and Newsgroups
- Manipulating and Creating Message Texts
- Communicating with Mail Servers
- Message Collections Message Parts
- World Wide Web Applications
- Common Gateway Interface
- Parsing, Creating, and Manipulating HTML Documents
- Accessing Internet Resources
- Synopses of Other Internet Modules
- Standard Internet-Related Tools
- Third-Party Internet-Related Tools
- Understanding XML
- Python Standard Library XML Modules
- Third-Party XML-Related Tools

APPENDICES
GLOSSARY
INDEX

ORDERING INFORMATION:

SINGLE COPY SALES:
Visa, Master Card, American Express, Checks, or Money Orders only —
Tel: 515-284-6761
Fax: 515-237-4416
Toll-Free: 800-282-0693

GOVERNMENT AGENCIES:
Prentice Hall Customer Service —
(#GS-02F-8023A)
Toll-Free: 800-922-0579

COLLEGE PROFESSORS:
Desk or Review Copies —
exam@aw.com

CORPORATE ACCOUNTS:
Quantity, Bulk Orders totalling 10 or more books. Purchase orders only —
No credit cards.
Fax: 317-817-7307
Toll-Free: 800-382-3419

INTERNATIONAL ORDERING INFORMATION:

CANADA:
beverley.carkner@pearsoned.com

UK/EMEA:
Europe, Middle East, South Africa
pollysawyer@pearsoned-ema.com

BENELUX:
raymond.ashruf@pearsoned-ema.com

AUSTRALIA:
trade@pearsoned.com.au

SOUTH ASIA:
leehong.tan@pearsoned.com.sg

NORTH ASIA:
cheranL@pearsoned.com.hk

OTHER REGIONS:
sarah.hlibka@pearsoned.com

Addison Wesley

FOR MORE INFORMATION VISIT: www.awprofessional.com/titles/0321112547/