

ББК 32.973.26-018.2я73+81.1я73

К47

Печатается по решению редакционно-издательского совета
Волгоградского государственного университета

Рецензент – кандидат физико-математических наук,
доцент *А.В. Карнов* (Волгоградский государственный университет)

Клячин, В. А.

К47 Автоматизированная обработка текстовой информации [Текст] : учеб. пособие / В. А. Клячин ; Федер. гос. авт. образоват. учреждение высш. проф. образования «Волгогр. гос. ун-т» ; Каф. компьютер. наук и эксперимент. мат. – Волгоград : Изд-во ВолГУ, 2012. – 152 с.

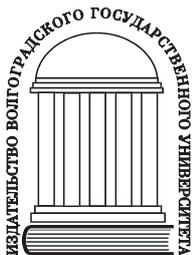
ISBN 978-5-9669-1118-8

Данное пособие посвящено изучению языка программирования Perl, разработанного с целью создания универсального средства автоматической обработки текста. Излагается синтаксис языка, типы переменных, операторы и конструкции; рассматривается программирование классов и разработка модулей, язык регулярных выражений как одно из современных средств поиска и преобразования строк, которое встроено в различные языки программирования и управления данными, а также Perl-диалект регулярных выражений, поддерживаемый такими системами, как PHP, MySQL, JavaScript, Qt и др. Освещаются типовые задачи обработки текста и их решения средствами языка Perl и регулярных выражений.

Предназначено для студентов, обучающихся по направлениям подготовки «Математическое обеспечение и администрирование информационных систем», «Прикладная математика и информатика», и может быть полезно всем, интересующимся регулярными выражениями.

ББК 32.973.26-018.2я73+81.1я73

ISBN 978-5-9669-1118-8



© Клячин В. А., 2012

© ФГАОУ ВПО «Волгоградский
государственный университет», 2012

© Оформление. Издательство Волгоградского
государственного университета, 2012

Оглавление

1	Основы языка Perl	11
1.1	Базовый Perl	12
1.1.1	Введение в Perl	12
1.1.2	Переменные в Perl	13
1.1.3	Выполнение скрипта	23
1.1.4	Система ввода – вывода	24
1.1.5	Скалярный и списочный контексты	26
1.1.6	Операции и операторы Perl	27
1.1.7	Операторы	30
1.1.8	Оператор двойных кавычек и другие по- добные операторы.	34
1.1.9	Встроенный документ	35
1.1.10	Пакеты и пространства имен	36
1.1.11	Подпрограммы	38
1.1.12	Функция eval()	45
1.1.13	Функция sort()	45
1.1.14	Ссылки	46
1.1.15	Организация структур данных и доступа к ним	49
1.1.16	Срезы массивов	51
1.1.17	Хеши массивов	52
1.1.18	Массив хешей	53
1.1.19	Хеш хешей	54
1.1.20	Сложные структуры данных	56
1.2	Объектно-ориентированный Perl	58
1.2.1	Модули и пакеты	58
1.2.2	Модули	60
1.2.3	Создание модулей	61
1.2.4	Объектно-ориентированные модули	62
1.2.5	Создание объектов	64
1.2.6	Наследуемые конструкторы	65
1.2.7	Наследование классов	67
1.2.8	Порождение классов с помощью пакета Class::Struct	69

1.2.9	Перегрузка операторов	70
1.2.10	Генерация HTML-таблиц	74
1.2.11	Некоторые стандартные модули Perl	78
1.2.12	Модуль IO::File	78
1.2.13	Межпроцессное взаимодействие	80
1.2.14	Модули для работы с серверами FTP и NT- ТР	82
1.2.15	Поддержка кодировок	85
2	Язык описания текстовых шаблонов	87
2.1	Синтаксис регулярных выражений	87
2.1.1	Введение в регулярные выражения	87
2.1.2	Оператор поиска	93
2.1.3	Один пример	95
2.1.4	Модификация строки	97
2.1.5	Опережающие и ретроспективные проверки	99
2.1.6	Еще пара слов об квантификаторах	101
2.1.7	Механизм обработки регулярных выражений	102
2.1.8	Возврат в механизме НКА	104
2.1.9	Захватывающие квантификаторы	108
2.1.10	Возврат при позиционной проверке	108
2.2	Методы составления регулярных выражений	109
2.2.1	Метод замещения	109
2.2.2	Метод уточнения окружения	110
2.2.3	Метод объединения компонентов	111
2.3	Обсуждение примеров решений задач	112
2.3.1	Преобразование текстового файла в фор- мат HTML	112
2.3.2	Поиск IP-адресов	113
2.3.3	Поиск парных скобок	115
2.3.4	Поиск текста в ограничителях	117
2.3.5	Поиск слов по динамически генерируемо- му шаблону	119
3	Типовые задачи обработки текста	121
3.1	Организация типовых сценариев обработки текста	121
3.2	Массовая обработка текстовых файлов на приме- ре FB2	123
3.3	Разбиение текста	131
3.4	Фильтрация текста	132
3.5	Форматирование и оформление текста	133
3.6	Удаление лишнего текста	134
3.7	Выделение необходимой части из текста	135
3.8	Преобразование текста в объект программы	135

3.9	Работа со структурированными текстами	139
3.9.1	Организация сообщений на форуме YaBB .	139
3.9.2	Работа с документами XML	142
3.9.3	Пример работы с файлом контактов Google (CSV)	144