Python para Informática

Explorando a Informação

Version 2.7.2

Autor: Charles Severance
Tradução: @victorjabur

Histórico de Publicação:

Maio 2015: Checagem editorial obrigado a Sue Blumenberg.

Outubro 2013: Revisão principal dos Capítulos 13 e 14 para mudar para JSON e usar OAuth. Novo capítulo adicionado na Visualização.

Setembro 2013: Livro publicado na Amazon CreateSpace

Janeiro 2010: Livro publicado usando a máquina da Universidade de Michigan Espresso Book.

Dezembro 2009: Revisão principal dos capítulos 2-10 de *Think Python: How to Think Like a Computer Scientist* e escrita dos capítulos 1 e 11-15 para produzir *Python for Informatics: Exploring Information*

Junho 2008: Revisão principal, título alterado para *Think Python: How to Think Like a Computer Scientist*.

Agosto 2007: Revisão principal, título alterado para *How to Think Like a (Python) Programmer*.

Abril 2002: Primeira edição de *How to Think Like a Computer Scientist*.

O código fonte LaTeX para a *Think Python: How to Think Like a Computer Scientist* versão deste livro está disponível em `http://www.thinkpython.com`.

Prefácio

Python para Informática: Adaptação de um livro aberto

É muito comum que acadêmicos, em sua profissão, necessitem publicar continuamente materiais ou artigos quando querem criar algo do zero. Este livro é um experimento em não partir da estaca zero, mas sim "remixar" o livro entitulado *Think Python: How to Think Like a Computer Scientist* escrito por Allen B. Downey, Jeff Elkner e outros.

Em dezembro de 2009, quando estava me preparando para ministrar a disciplina **SI502 - Programação para Redes** na Universidade de Michigan para o quinto semestre e decidi que era hora de escrever um livro de Python focado em explorar dados ao invés de entender algoritmos e abstrações. Minha meta em SI502 é ensinar pessoas a terem habilidades na manipulação de dados para a vida usando Python. Alguns dos meus estudantes planejavam se tornarem profissionais em programação de computadores. Ao invés disso, eles escolheram ser bibliotecários, gerentes, advogados, biólogos, economistas, etc., e preferiram utilizar habilmente a tecnologia nas áreas de suas escolhas.

Eu nunca consegui encontrar o livro perfeito sobre Python que fosse orientado a dados para utilizar no meu curso, então eu comecei a escrever o meu próprio. Com muita sorte, em uma reunião eventual três semanas antes de eu começar a escrever o meu novo livro do zero, em um descanso no feriado, Dr. Atul Prakash me mostrou o *Think Python* livro que ele tinha usado para ministrar seu curso de Python naquele semestre. Era um texto muito bem escrito sobre Ciência da Computação com foco em explicações diretas e simples de se aprender.

Toda a estrutura do livro foi alterada, visando a resolução de problemas de análise de dados de um modo tão simples e rápido quanto possível, acrescido de uma série de exemplos práticos e exercícios sobre análise de dados desde o início.

Os capítulos 2–10 são similares ao livro *Think Python* mas precisaram de muitas alterações. Exemplos com numeração e exercícios foram substituídos por exercícios orientados a dados. Tópicos foram apresentados na ordem necessária para construir soluções sofisticadas em análise de dados. Alguns tópicos tais como `try` e `except` foram movidos mais para o final e apresentados como parte do capítulo de condicionais. Funções foram necessárias para simplificar a complexidade na manipulação dos programas introduzidos anteriormente nas primeiras

lições em abstração. Quase todas as funções definidas pelo usuário foram removidas dos exemplos do código e exercícios, com exceção do Capítulo 4. A palavra "recursão"[1] não aparece no livro inteiro.

Nos capítulos 1 e 11–16, todo o material é novo, focado em exemplos reais de uso e exemplos simples de Python para análise de dados incluindo expressões regulares para busca e transformação, automação de tarefas no seu computador, recuperação de dados na internet, extração de dados de páginas web, utilização de *web services*, transformação de dados em XML para JSON, e a criação e utilização de bancos de dados utilizando SQL (Linguagem estruturada de consulta em bancos de dados).

O último objetivo de todas estas alterações é a mudança de foco, de Ciência da Computação para uma Informática que inclui somente tópicos que podem ser utilizados em uma turma de primeira viagem (iniciantes) que podem ser úteis mesmo se a escolha deles não for seguir uma carreira profissional em programação de computadores.

Estudantes que acharem este livro interessante e quiserem se aprofundar devem olhar o livro de Allen B. Downey's *Think Python*. Porque há muita sinergia entre os dois livros, estudantes irão rapidamente desenvolver habilidades na área com a técnica de programação e o pensamento em algoritmos, que são cobertos em *Think Python*. Os dois livros possuem um estilo de escrita similar, é possível mover-se para o livro *Think Python* com o mínimo de esforço.

Com os direitos autorais de *Think Python*, Allen me deu permissão para trocar a licença do livro em relação ao livro no qual este material é baseado de GNU Licença Livre de Documentação para a mais recente Creative Commons Attribution — Licença de compartilhamento sem ciência do autor. Esta baseia-se na documentação aberta de licenças mudando da GFDL para a CC-BY-SA (i.e., Wikipedia). Usando a licença CC-BY-SA, os mantenedores deste livro recomendam fortemente a tradição "copyleft" que incentiva os novos autores a reutilizarem este material da forma como considerarem adequada.

Eu sinto que este livro serve de exemplo sobre como materiais abertos (gratuito) são importantes para o futuro da educação, e quero agradecer ao Allen B. Downey e à editora da Universidade de Cambridge por sua decisão de tornar este livro disponível sob uma licença aberta de direitos autorais. Eu espero que eles fiquem satisfeitos com os resultados dos meus esforços e eu desejo que você leitor esteja satisfeito com *nosso* esforço coletivo.

Eu quero fazer um agradecimento ao Allen B. Downey e Lauren Cowles por sua ajuda, paciência, e instrução em lidar com este trabalho e resolver os problemas de direitos autorais que cercam este livro.

Charles Severance
www.dr-chuck.com

[1] Com exceção, naturalmente, desta linha.

Ann Arbor, MI, USA
9 de Setembro de 2013

Charles Severance é um Professor Associado à Escola de Informação da Universidade de Michigan.

Tradução:
@victorjabur

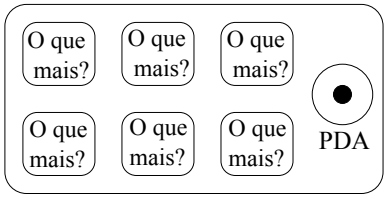
Sumário

Capítulo 1

Por que você deve aprender a escrever programas ?

Escrever programas (ou programação) é uma atividade muito criativa e recompensadora. Você pode escrever programas por muitas razões, que vão desde resolver um difícil problema de análise de dados a se divertir ajudando alguém a resolver um problema. Este livro assume que *qualquer pessoa* precisa saber como programar, e uma vez que você sabe como programar, você irá imaginar o que você quer fazer com suas novas habilidades.

Nós estamos cercados no nosso dia a dia por computadores, desde notebooks até celulares. Nós podemos achar que estes computadores são nossos "assistentes pessoais" que podem cuidar de muitas coisas a nosso favor. O hardware desses computadores no nosso dia a dia é essencialmente construído para nos responder a uma pergunta, "O que você quer que eu faça agora ?"



Programadores adicionam um sistema operacional e um conjunto de aplicações ao hardware e nós terminamos com um Assistente Pessoal Digital que é muito útil e capaz de nos ajudar a fazer diversas coisas.

Nossos computadores são rápidos, tem vasta quantidade de memória e podem ser muito úteis para nós, somente se conhecermos a linguagem falada para explicar para um computador o que nós gostaríamos de fazer "em seguida". Se nós conhecemos esta linguagem, nós podemos pedir ao computador para fazer tarefas repetitivas a nosso favor. Curiosamente, as coisas que os computadores podem fazer melhor são frequentemente aquelas coisas que humanos acham chatas e entediantes.

Por exemplo, olhe para os três primeiros parágrafos deste capítulo e me diga qual é a palavra mais usada e quantas vezes. Contá-las é muito doloroso porque não é o tipo de problema que mentes humanas foram feitas para resolver. Para um computador o oposto é verdade, ler e entender o texto de um pedaço de papel é difícil, mas contar palavras dizendo a você quantas vezes ela aparece é muito fácil:

```
python palavras.py
Digite o nome do arquivo: palavras.txt
para 16
```

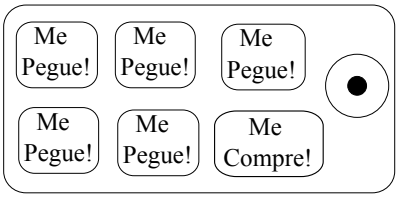
Nosso "assistente de análise pessoal de informações" rapidamente conta para nós que a palavra "para" foi utilizada dezesseis vezes nos primeiros três parágrafos deste capítulo.

Este fato de que os computadores são bons em coisas que humanos não são é a razão pela qual você precisa tornar-se qualificado em falar a "linguagem do computador". Uma vez que você aprende esta nova linguagem, pode delegar tarefas mundanas para o seu parceiro (o computador), ganhando mais tempo para fazer coisas que você foi especialmente adaptado para fazer. Você agrega criatividade, intuição e originalidade para o seu parceiro.

1.1 Criatividade e motivação

Embora este livro não se destine a programadores profissionais, programação profissional pode ser um trabalho muito gratificante, tanto financeiramente quanto pessoalmente. Construir programas úteis, elegantes, inteligentes para que outros utilizem é uma atividade criativa. Seu computador ou assistente pessoal digital (PDA) geralmente contém muitos programas diferentes feitos por diversos grupos de programadores, todos competindo por sua atenção e seu interesse. Eles tentam dar o seu melhor para atender suas necessidades e dar a você uma boa experiência de usabilidade no processo. Em algumas situações, quando você executa um trecho de software, os programadores são diretamente recompensados por sua escolha.

Se nós pensarmos em programas como resultado criativo de grupos de programadores, então talvez a figura a seguir seja uma versão mais sensata de nosso PDA:

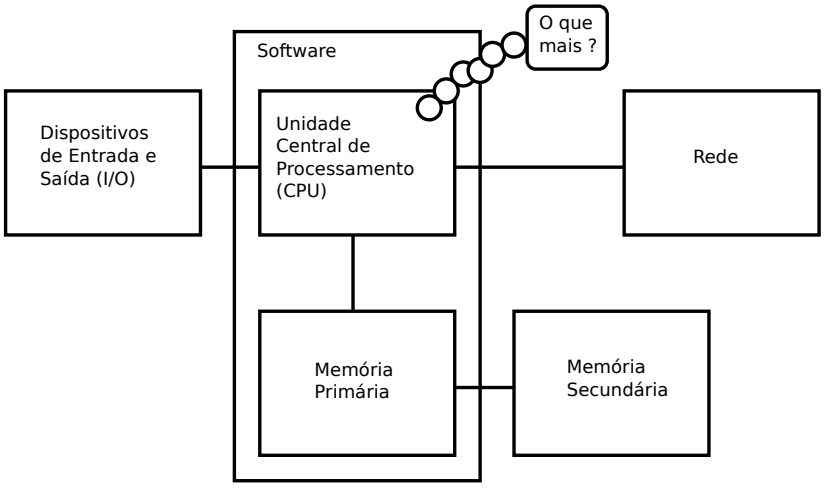


Por enquanto, nossa motivação primária não é ganhar dinheiro ou agradar usuários finais, mas sermos mais produtivos na manipulação de dados e informações que nós encontraremos em nossas vidas. Quando você começar, você será tanto o

programador quanto o usuário final de seus programas. Conforme você ganhar habilidades como programador e melhorar a criatividade em seus próprios programas, mais você pode pensar em programar para os outros.

1.2 Arquitetura física do Computador - Hardware

Antes de começar a estudar a linguagem, nós falamos em dar instruções aos computadores para desenvolver software, nós precisamos aprender um pouco mais sobre como os computadores são construídos. Se você desmontar seu computador ou celular e olhar por dentro, você encontrará as seguintes partes:



As definições resumidas destas partes são:

- A **Unidade Central de Processamento** (ou CPU) é a parte do computador que é feita para sempre te perguntar: "O que mais ?" Se seu computador possui uma frequência de 3.0 Gigahertz, significa que a CPU irá te perguntar "O que mais ?" três bilhões de vezes por segundo. Você irá aprender como conversar tão rápido com a CPU.

- A **Memória Principal** é utilizada para armazenar informação que a CPU precisa com muita pressa. A memória principal é aproximadamente tão rápida quanto a CPU. Mas a informação armazenada na memória principal se perde quando o computador é desligado (volátil).

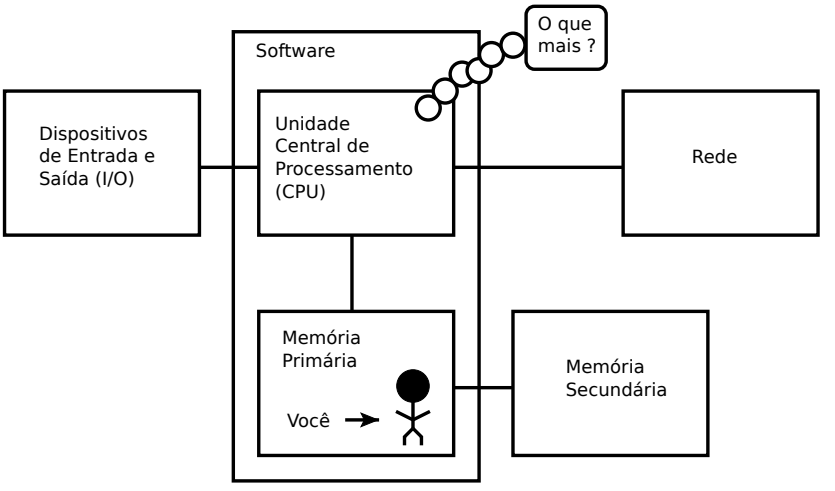
- A **Memória Secundária** é também utilizada para armazenar informação, mas ela é muito mais lenta que a memória principal. A vantagem da memória secundária é que ela pode armazenar informação que não se perde quando o computador é desligado. Exemplos de memória secundária são discos rígidos (HD), pen drives, cartões de memória (sd card) (tipicamente) encontradas no formato de USB e portáteis.

- Os **Dispositivos de Entrada e Saídas** são simplesmente nosso monitor (tela), teclado, mouse, microfone, caixa de som, touchpad, etc. Eles são todas as formas com as quais interagimos com o computador.

- Atualmente, a maioria dos computadores tem uma **Conexão de Rede** para buscar informação em uma rede. Nós podemos pensar a rede como um lugar muito lento para armazenar e buscar dados que podem não estar "disponíveis". Em essência, a rede é mais lenta e às vezes parece uma forma não confiável de **Memória Secundária**.

É melhor deixar a maior parte dos detalhes de como estes componentes funcionam para os construtores dos computadores. Isso nos ajuda a ter alguma terminologia que podemos utilizar para conversar sobre essas partes conforme escrevemos nossos programas.

Como um programador, seu trabalho é usar e orquestrar cada um destes recursos para resolver um problema que você precisa resolver e analisar os dados que você obtém da solução. Como um programador você irá "conversar" com a CPU e contar a ela o que fazer em um próximo passo. Algumas vezes você irá dizer à CPU para usar a memória principal, a memória secundária, a rede ou os dispositivos de entrada e saída.



Você precisa ser a pessoa que responde à pergunta "O que mais ?" para a CPU. Mas seria muito desconfortável se você fosse encolhido para uma altura de apenas 5 mm e inserido dentro de um computador e ainda ter que responder uma pergunta três bilhões de vezes por segundo. Então, ao invés disso, você deve escrever suas instruções previamente. Nós chamamos essas instruções armazenadas de **programa** e o ato de escrever essas instruções e garantir que essas estejam corretas de **programação**.

1.3 Entendendo programação

No restante deste livro, nós iremos tentar fazer de você uma pessoa com habilidades na arte da programação. No final você será um **programador**, no entanto não um programador profissional, mas pelo menos você terá os conhecimentos para analisar os problemas de dados/informações e desenvolver um programa para resolver tais problemas.

Resumidamente, você precisa de duas qualidades para ser um programador:

- Primeiramente, você precisa conhecer uma linguagem de programação (Python) - você precisa conhecer o vocabulário e a gramática. Você precisa saber pronunciar as palavras desta nova linguagem corretamente e conhecer como construir "sentenças" bem formadas nesta linguagem.
- Segundo, você precisa "contar uma estória". Na escrita da estória, você combina palavras e sentenças para convencer o leitor. É necessário qualidade e arte na construção da estória, adquiri-se isso através da prática de contar estórias e obter um feedback. Na programação, nosso programa é a "estória" e o problema que você quer resolver é a "idéia".

Uma vez que você aprende uma linguagem de programação, como o Python, você irá achar muito mais fácil aprender uma segunda linguagem de programação, tal como JavaScript ou C++. A nova linguagem de programação possuirá um vocabulário e gramática bastante diferente, mas as habilidades na resolução do problemas serão as mesmas em qualquer linguagem.

Você aprenderá o "vocabulário" e "sentenças" do Python rapidamente. Levará muito tempo para você tornar-se hábil em escrever programas coerentes para resolver um novo problema. Nós ensinamos programação assim como ensinamos a escrever. Nós leremos e explicaremos programas, nós escreveremos programas simples, e então nós aumentaremos a complexidade dos programas ao longo do tempo. Em algum momento, você "deslancha" e vê os padrões por si próprio e pode visualizar com maior naturalidade como escrever um programa para resolver o problema. Uma vez que você chega neste ponto, programar torna-se um processo muito agradável e criativo.

Nós iniciamos com o vocabulário e a estrutura de programas em Python. Seja paciente com os exemplos simples, relembre quando você iniciou a leitura pela primeira vez.

1.4 Palavras e Sentenças

Diferentemente dos idiomas humanos, o vocabulário do Python é atualmente muito pequeno. Nós chamamos esse "vocabulário" de "palavras reservadas". Estas palavras tem um significado especial no Python. Quando o Python encontra

estas palavras em um programa, elas possuem um e somente um significado para o Python. Quando você escrever seus programas você irá definir suas próprias palavras com significado, são chamadas **variáveis**. Você pode escolher muitos nomes diferentes para as suas variáveis, mas você não pode usar qualquer palavra reservada do Python como o nome de uma variável.

Quando nós treinamos um cachorro, nós usamos palavras especiais, tais como: "sentado", "fique" e "traga". Quando você conversar com cachorros e não usar qualquer uma dessas palavras reservadas, eles ficarão olhando para você com um olhar curioso até que você diga uma palavra reservada. Por exemplo, se você disser: "Eu desejo que mais pessoas possam caminhar para melhorar a sua saúde", o que os cachorros vão ouvir será: "blah blah blah **caminhar** blah blah blah blah." Isto porque "caminhar" é uma palavra reservada na linguagem dos cachorros. Muitos podem sugerir que a linguagem entre humanos e gatos não tem palavras reservadas[1].

As palavras reservadas na linguagem pelas quais os humanos conversam com o Python, incluem as seguintes:

```
and       del       from      not       while
as        elif      global    or        with
assert    else      if        pass      yield
break     except    import    print
class     exec      in        raise
continue  finally   is        return
def       for       lambda    try
```

É isso, e ao contrário do cachorro, o Python é completamente treinado. Quando você diz "try", o Python irá tentar todas as vezes que você pedir sem desobedecer.

Nós aprenderemos as palavras reservadas e como elas são usadas mais adiante, por enquanto nós iremos focar no equivalente ao Python de "falar" (na linguagem humano-para-cachorro). Uma coisa legal sobre pedir ao Python para falar é que nós podemos até mesmo pedir o que nós queremos através de uma mensagem entre aspas:

```
print 'Hello world!'
```

E finalmente nós escrevemos a nossa primeira sentença sintaticamente correta em Python. Nossa sentença inicia com uma palavra reservada **print** seguida por uma cadeia de caracteres textuais de nossa escolha entre aspas simples.

1.5 Conversando com Python

Agora que você tem uma palavra e uma simples sentença que nós conhecemos em Python, nós precisamos saber como iniciar uma conversação com Python para testar nossas habilidades na nova linguagem.

[1] http://xkcd.com/231/

Antes de você conversar com o Python, você deve primeiramente instalar o programa Python em seu computador e aprender como inicializá-lo. Isto é muita informação para este capítulo, então eu sugiro que você consulte `www.pythonlearn.com` onde se encontra instruções e screencasts de preparação e inicialização do Python em sistemas Windows e Macintosh. Em algum momento, você estará no interpretador Python, executando o modo interativo e aparecerá algo assim:

```
Python 2.6.1 (r261:67515, Jun 24 2010, 21:47:49)
[GCC 4.2.1 (Apple Inc. build 5646)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

O prompt >>> é a forma do interpretador Python perguntar o que você deseja: "O que você quer que eu faça agora?" Python está pronto para ter uma conversa com você. Tudo o que você deve conhecer é como falar a linguagem Python.

Digamos, por exemplo, que você não conhece nem mesmo as mais simples palavras ou sentenças da linguagem Python. Você pode querer usar a linha padrão que os astronautas usam quando eles estão em uma terra distante do planeta e tentam falar com os habitantes do planeta:

```
>>> Eu venho em paz, por favor me leve para o seu líder
  File "<stdin>", line 1
    Eu venho em paz, por favor me leve para o seu líder
          ^
SyntaxError: invalid syntax
>>>
```

Isto não deu muito certo. A menos que você pense algo rapidamente, os habitantes do planeta provavelmente irão apunhalá-lo com uma lança, colocá-lo em um espeto, assá-lo no fogo e comê-lo no jantar.

A sorte é que você trouxe uma cópia deste livro em sua viagem, e caiu exatamente nesta página, tente novamente:

```
>>> print 'Ola Mundo!'
Ola Mundo!
```

Isso parece bem melhor, então você tenta se comunicar um pouco mais:

```
>>> print 'Voce deve ser um Deus lendario que veio do ceu'
Voce deve ser um Deus lendario que veio do ceu
>>> print 'Nos estivemos esperando voce por um longo tempo'
Nos estivemos esperando voce por um longo tempo
>>> print 'Nossa lenda nos conta que voce seria muito apetitoso com mostarda'
Nossa lenda nos conta que voce seria muito apetitoso com mostarda
>>> print 'Nos teremos uma festa hoje a noite a menos que voce diga
  File "<stdin>", line 1
    print 'Nos teremos uma festa hoje a noite a menos que voce diga
                                                                   ^
SyntaxError: EOL while scanning string literal
>>>
```

A conversa foi bem por um momento, até que você cometeu o pequeno erro no uso da linguagem e o Python trouxe a lança de volta.

Até o momento, você deve ter percebido que o Python é incrivelmente complexo, poderoso e muito exigente em relação à sintaxe que você utiliza para se comunicar com ele, Python *não* é inteligente. Você está na verdade tendo uma conversa com você mesmo, mas usando uma sintaxe apropriada.

De certa forma, quando você usa um programa escrito por alguém, a conversa ocorre entre você e os programadores, neste caso o Python atuou como um intermediário. Python é uma forma para que os criadores de programas se expressem sobre como uma conversa deve proceder. E em poucos capítulos, você será um dos programadores usando Python para conversar com os usuários de seus programas.

Antes de sairmos da nossa primeira conversa com o interpretador do Python, você deve conhecer o modo correto de dizer "ate-logo" quando interagir com os habitantes do Planeta Python.

```
>>> ate-logo
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
NameError: name 'ate' is not defined

>>> se voce nao se importa, eu preciso ir embora
  File "<stdin>", line 1
    se voce nao se importa, eu preciso ir embora
          ^
SyntaxError: invalid syntax

>>> quit()
```

Você pode perceber que o erro é diferente nas duas primeiras tentativas incorretas. No primeiro erro, por tratar-se de uma palavra simples, o Python não pode encontrar nenhuma função ou variável com este nome. No segundo erro, existe um erro de sintaxe, não sendo reconhecida a frase como válida.

O jeito correto de se dizer "ate-logo" para o Python é digitar **quit()** no prompt do interpretador interativo. É provável que você tenha perdido certo tempo tentado fazer isso, ter um livro em mãos irá tornar as coisas mais fáceis e pode ser bastante útil.

1.6 Terminologia: interpretador e compilador

Python é uma linguagem de **alto nível** cujo objetivo é ser relativamente fácil para humanos lerem e escreverem e para computadores lerem e processarem. Outras linguagens de alto nível incluem Java, C++, PHP, Ruby, Basic, Perl, JavaScript, e muito mais. O atual hardware dentro da Unidade Central de Processamento (CPU) não é capaz de entender nenhum destes comando em alto nível.

A CPU entende a linguagem que chamamos de **linguagem de máquina**. Linguagem de máquina é muito simples e francamente cansativa de se escrever porque ela é representada em zeros e uns:

```
01010001110100100101010000001111
11100110000011101010010101101101
...
```

Linguagem de máquina parece simples olhando-se de um modo superficial, dado que são apenas zeros e uns, mas sua sintaxe é muito mais complexa e mais intrincada que o Python. Poucos programadores escrevem em linguagem de máquina. Ao invés disso, nós usamos vários tradutores para permitir que os programadores escrevam em linguagem de máquina a partir de linguagens de alto nível como o Python ou o JavaScript. Essas linguagens convertem os programas para linguagem de máquina que, desse modo, são executados pela CPU.

Visto que linguagem de máquina é vinculada ao hardware do computador, linguagem de máquina não é **portável** entre diferentes tipos de hardware. Programas que foram escritos em linguagens de alto nível podem mover-se entre diferentes computadores usando um interpretador diferente em cada máquina ou então recompilando o código para criar uma versão de linguagem de máquina do programa para a nova máquina.

Os tradutores das linguagens de programação se enquadram em duas características gerais: (1) interpretadores e (2) compiladores

Um **interpretador** lê o código fonte de um programa da forma como foi escrito pelo programador, analisa, e interpreta as instruções em tempo de execução. Python é um interpretador e quando ele está rodando Python no modo interativo, nós podemos digitar uma linha de Python (uma sentença) e o Python a processa imediatamente e está pronto para receber outra linha de Python.

Algumas das linhas de Python diz a ele que você quer armazenar algum valor para resgatar depois. Nós precisamos dar um nome para um valor de forma que possa ser armazenado e resgatado através deste nome simbólico. Nós usamos o termo **variável** para se referir aos apelidos que nós demos ao dado que foi armazenado.

```
>>> x = 6
>>> print x
6
>>> y = x * 7
>>> print y
42
>>>
```

Neste exemplo, nós pedimos ao Python para armazenar o valor seis e usar um apelido **x**, de modo a nós podermos resgatar o valor mais tarde. Nós verificamos que o Python realmente lembrou dos valores quando usamos a função **print**. Então nós perguntamos ao Python para resgatar **x**, multiplicá-lo por sete e armazenar de

novo em uma variável **y**. Então nós pedimos ao Python para exibir o valor corrente em **y**.

Mesmo que nós digitemos estes comandos em uma única linha de Python por vez, o Python está processando elas em sequência, mantendo a ordem, de forma que as instruções seguintes consigam recuperar os dados criados pelas anteriores. Nós escrevemos nosso primeiro parágrafo simples com quatro sentenças com uma ordem lógica e com um significado.

É da natureza de um **interpretador** ser capaz de ter uma conversa interativa, como foi mostrado acima. Um **compilador** precisa ter em mãos o programa completo em um arquivo, então ele roda um processo para traduzir um código fonte em alto nível para uma linguagem de máquina e, em seguida, o compilador coloca o resultado deste processo em um outro arquivo para posterior execução.

Se você está em um sistema Windows, frequentemente este programa executável em código de máquina tem o sufixo ".exe" ou ".dll", os quais são chamados de "executável" ou "biblioteca de link dinâmico", respectivamente. Em Linux e Macintosh, não há um sufixo que marca unicamente um arquivo como executável.

Se você abrir um arquivo executável em um editor de texto, verá algo completamente doido e ilegível.

```
^?ELF^A^A^A^@^@^@^@^@^@^@^@^B^@^C^@^A^@^@^@\xa0\x82
^D^H4^@^@^@\x90^]^@^@^@^@^@^@4^@ ^@^G^@(^@$^@!^@^F^@
^@^@4^@^@^@4\x80^D^H4\x80^D^H\xe0^@^@^@\xe0^@^@^@^E
^@^@^@^D^@^@^@^C^@^@^@^T^A^@^@^T\x81^D^H^T\x81^D^H^S
^@^@^@^S^@^@^@^D^@^@^@^A^@^@^@^A\^D^HQVhT\x83^D^H\xe8
....
```

Não é nada fácil ler ou escrever código de máquina, assim é bom que tenhamos **interpretadores** e **compiladores** que nos permitam escrever em linguagens de alto nível assim como o Python ou o C.

Agora neste ponto em nossa discussão de compiladores e interpretadores, você deve estar com algumas dúvidas sobre o funcionamento do interpretador Python. Em que linguagem é escrito? É escrito em uma linguagem compilada? Quando nós digitamos "python", o que exatamente acontece?

O interpretador Python é escrito em uma linguagem de alto nível chamada "C". Você pode dar uma olhada no código fonte do interpretador através do endereço `www.python.org` e trabalhar como você quiser com o código fonte. Então Python é um programa compilado em uma linguagem de máquina. Quando você instalou Python em seu computador (ou o fornecedor instalou), você copiou um código de máquina do programa Python traduzido para o seu sistema. Em Windows, o código de máquina executável para o Python encontra-se em um arquivo com um nome a seguir:

```
C:\Python27\python.exe
```

Isso é mais do que você realmente precisa conhecer para ser um programador Python, mas às vezes, isso ajuda a entender questões que intrigam justamente no início.

1.7 Escrevendo um programa

Digitar comandos em um Interpretador Python é uma boa maneira de experimentar as características da linguagem, mas isto não é recomendado para resolver problemas mais complexos.

Quando nós queremos escrever um programa, usamos um editor de texto para escrever as instruções Python em um arquivo, o qual chamamos de **script**. Por convenção, scripts Python tem nomes que terminam com `.py`.

Para executar o script, você tem que dizer ao interpretador do Python o nome do arquivo. Em uma janela de comandos Unix ou Windows, você digita `python hello.py` como a seguir:

```
csev$ cat hello.py
print 'Ola Mundo!'
csev$ python hello.py
Ola Mundo!
csev$
```

O "csev$" é o prompt do sistema operacional, e o "cat hello.py" é para nos mostrar que o arquivo "hello.py" tem uma linha de programa Python para imprimir uma string.

Nós chamamos o interpretador Python e pedimos a ele para ler o código fonte do arquivo "hello.py" ao invés dele nos perguntar quais são as próximas linhas de modo interativo.

Você notará que não é preciso ter o **quit()** no fim do programa Python no arquivo. Quando o Python está lendo o seu código fonte de um arquivo, ele sabe que deve parar quando chegar ao fim do arquivo.

1.8 O que é um programa ?

A definição de um **programa** em sua forma mais básica é uma sequência de comandos Python que foram criados para fazer algo. Mesmo o nosso simples script **hello.py** é um programa. É um programa de uma linha e não é particularmente útil, mas na estrita definição, é um programa Python.

Pode ser mais fácil entender o que é um programa, imaginando qual problema ele foi construído para resolver, e então olhar para o programa que resolve um problema.

Vamos dizer que você está fazendo uma pesquisa de computação social em posts do Facebook e está interessado nas palavras mais frequentes em uma série de posts. Você pode imprimir o stream de posts do Facebook e debruçar-se sobre o texto procurando pela palavra mais comum, mas pode levar um tempo longo e ser muito propenso a erros. Você pode ser inteligente para escrever um programa Python para tratar disso rapidamente a com acurácia, então você pode passar seu final de semana fazendo algo divertido.

Por exemplo, olhe para o seguinte texto sobre o palhaço e o carro. Olhe para o texto e imagine qual é a palavra mais comum e quantas vezes ela aparece:

```
O palhaço correu atrás do carro e o carro correu para a tenda
e a tenda caiu em cima do palhaço e do carro
```

Então imagine que você está fazendo esta tarefa olhando para milhões de linhas de texto. Francamente será mais rápido para você aprender Python e escrever um programa Python para contar as palavras do que você manualmente escanear as palavras.

A notícia ainda melhor é que eu já fiz para você um programa simples para encontrar a palavra mais comum em um arquivo texto. Eu escrevi, testei e agora eu estou dando isso para que você use e economize algum tempo.

```
name = raw_input('Enter file:')
handle = open(name, 'r')
text = handle.read()
words = text.split()
counts = dict()

for word in words:
   counts[word] = counts.get(word,0) + 1

bigcount = None
bigword = None
for word,count in counts.items():
    if bigcount is None or count > bigcount:
        bigword = word
        bigcount = count

print bigword, bigcount
```

Você nem precisa conhecer Python para usar este programa. Você precisará chegar até o capítulo 10 deste livro para entender completamente as impressionantes técnicas Python que foram utilizadas para fazer o programa. Você é o usuário final, você simplesmente usa o programa e admira-se com a inteligência e em como ela poupou seus esforços manuais. Você simplesmente digitou o código em um arquivo chamado **words.py** e executou ou então fez o download do código fonte no site `http://www.pythonlearn.com/code/` e executou.

Este é um bom exemplo de como o Python e sua linguagem podem atuar como um intermediário entre você (o usuário final) e eu (o programador). Python é uma

forma para trocarmos sequências úteis de instruções (i.e., programas) em uma linguagem comum que pode ser usada por qualquer um que instalar Python em seu computador. Então nenhum de nós está conversando *com o Python* mas sim nos comunicando uns com os outros *através* de Python.

1.9 A construção de blocos de programas

Em poucos capítulos, nós iremos aprender mais sobre o vocabulário, estrutura das sentenças, dos parágrafos e da história do Python. Nós iremos aprender sobre as capacidades poderosas do Python e como compor estas capacidades juntas para criar programas úteis.

Há alguns padrões conceituais de baixo nível que nós usamos para construir programas. Estas construções não são apenas para programas Python, elas são parte de todas as linguagens de programação desde linguagens de baixo nível até as de alto nível.

input: Obter dados do "mundo externo". Estes dados podem ser lidos de um arquivo ou mesmo de algum tipo de sensor como um microfone ou um GPS. Em nossos primeiros programas, nosso input virá de um usuário que digita dados no teclado.

output: Exibe os resultados do programa em uma tela ou armazena-os em um arquivo ou talvez os escreve em algum dispositivo tal como um alto falante para tocar música ou falar o texto.

execução sequencial: Executa instruções uma após a outra respeitando a sequência encontrada no script.

execução condicional: Avalia certas condições e as executa ou pula a sequência de instruções.

execução repetitiva: Executa algumas instruções repetitivamente, geralmente com alguma variação.

reúso: Escrever um conjunto de instruções uma única vez, dar um nome a elas e reusar estas instruções em várias partes de um programa.

Parece simples demais para ser verdade, e naturalmente que isto nunca é tão simples. É como dizer que caminhar é simplesmente "colocar um pé na frente do outro". A "arte" de escrever um programa é compor e costurar estes elementos básicos muitas vezes para produzir algo que seja útil aos usuários.

O programa de contar palavras acima usa todos estes padrões exceto um.

1.10 O que pode dar errado?

Como vimos em nossa última conversa com o Python, devemos nos comunicar de modo preciso quando escrevemos código Python. O mínimo desvio ou erro fará com que o Python pare de executar o seu programa.

Programadores iniciantes muitas vezes tomam o fato de que o Python não deixa espaço para erros como prova de que ele é malvado e cruel. Enquanto o Python parece gostar de todo mundo, ele os conhece pessoalmente e guarda um ressentimento contra eles. Devido a este ressentimento, o Python avalia nossos programas perfeitamente escritos e os rejeita como "incorretos" apenas para nos atormentar.

```
>>> primt 'Ola mundo!'
  File "<stdin>", line 1
    primt 'Ola mundo!'
                  ^
SyntaxError: invalid syntax
>>> primt 'Ola mundo'
  File "<stdin>", line 1
    primt 'Ola mundo'
                 ^
SyntaxError: invalid syntax
>>> Eu te odeio Python!
  File "<stdin>", line 1
    Eu te odeio Python!
        ^
SyntaxError: invalid syntax
>>> se você vier aqui fora, vou te dar uma lição
  File "<stdin>", line 1
    se você vier aqui fora, vou te dar uma lição
             ^
SyntaxError: invalid syntax
>>>
```

Não se ganha muita coisa discutindo com o Python. Ele é somente uma ferramenta. Ele não tem emoções e fica feliz e pronto para te servir quando você precisar dele. Suas mensagens de erro parecem ásperas, mas elas apenas tentam nos ajudar. Ele recebeu o seu comando e simplesmente não conseguiu entender o que você digitou.

Python se parece muito com um cachorro, te ama incondicionalmente, consegue entender apenas algumas poucas palavras, olha para você com um olhar doce na face (>>>), e fica esperando você dizer algo que ele entenda. Quando o Python diz "SyntaxError: invalid syntax", está simplesmente abanando o rabo e dizendo, "Parece que você disse algo que eu não consegui entender, por favor, continue conversando comigo (>>>)."

Conforme seu programa vai se tornando mais sofisticado, você encontrará três tipos genéricos de erro:

Erros de Sintaxe: Estes são os primeiros erros que você cometerá e os mais fáceis de se consertar. Um erro de sintaxe significa que você violou as "re-

gras gramaticais" do Python. Python dá o seu melhor para apontar a linha correta e o caractere que o confundiu. A única parte complicada dos erros de sintaxe é que às vezes os erros que precisam de conserto na verdade ocorrem um pouco antes de onde o Python *indica* e isso confunde um pouco. Desta forma, a linha e caractere que o Python indica no erro de sintaxe pode ser que seja apenas um ponto de início para sua investigação.

Erros de Lógica: Um erro de lógica é quando o seu programa tem uma boa sintaxe mas há um erro na ordem das instruções ou às vezes um erro em como uma instrução se relaciona com as demais. Um bom exemplo de erro de lógica pode ser, "tome um gole de sua garrafa de água, coloque-a na mochila, caminhe para a biblioteca, e depois coloque a tampa de volta na garrafa."

Erros de Semântica: Um erro de semântica é quando a descricao dos passos estão sintaticamente corretos, na ordem certa, mas há existe um erro no programa. O programa está perfeitamente correto, mas ele não faz o que você *deseja* que ele faça. Um exemplo simples poderia ser quando você instrui uma pessoa a chegar até um restaurante e diz, "quando você cruzar a estação de gás, vire à esquerda e ande por um quilômetro e o restaurante estará no prédio vermelho à sua esquerda." Seu amigo está muito atrasado e liga para você para dizer que está em uma fazenda, passando atrás de um celeiro, sem o sinal da existência de um restaurante. Então você diz "você virou à esquerda ou à direita na estação de gás ?" e ele diz: "Eu segui suas instruções perfeitamente, as escrevi em um papel, e dizia para virar à esquerda e andar por um quilômetro até a estação de gás." Então você diz: "Eu sinto muito, embora minhas instruções estivessem sintaticamente corretas, elas infelizmente tinham um pequeno erro semântico não detectado."

Novamente em todos os três tipos de erros, o Python está se esforçando para fazer tudo aquilo que você pediu.

1.11 A jornada do aprendizado

Enquanto você progride para o restante do livro, não tenha medo se os conceitos não parecem se encaixar tão bem em um primeiro momento. Quando você aprendeu a falar, não era um problema que em seus primeiros anos você fizesse sons fofos e desajeitados. Foi tudo certo se levou seis meses para se mover de um vocabulário simples até sentenças simples e levou mais 5-6 anos para se mover de sentenças a parágrafos, e uns anos mais para estar habilitado a escrever uma estória curta e interessante com suas próprias mãos.

Nós queremos que você aprenda Python muito mais rápido, então nós ensinamos tudo ao mesmo tempo nos próximos capítulos. Mas aprender uma nova linguagem leva tempo para ser absorver e entender antes de se tornar natural. Este processo

pode gerar alguma confusão conforme nós visitamos e revisitamos os tópicos para tentar dar a você uma visão completa, nós definimos pequenos fragmentos que aos poucos irão formando a visão completa. Este livro é dividido em capítulos sequenciais e à medida que você avança vai aprendendo diversos assuntos, não se sinta preso na sequência do livro, avance capítulos e depois recue se for preciso, o que importa é o seu aprendizado e em como você sente que deve ser. Ao estudar superficialmente materiais mais avançados sem entender completamente os detalhes, você pode obter um melhor entendimento do "porque?" programar. Revisando materiais mais básicos e até mesmo refazendo exercícios anteriores, você irá perceber que aprendeu muito, até mesmo com aqueles materiais que pareciam impenetráveis de tão difíceis.

Normalmente, quando você aprende sua primeira linguagem de programação, ocorrem vários momentos "Ah Hah!". Aqueles em que você está trabalhando arduamente e quando para para prestar atenção e dar um descanso percebe que está construindo algo maravilhoso.

Se algo estiver particularmente difícil, saiba que não vale a pena ficar acordado a noite inteira encarando o problema. Faça uma pausa, tire um cochilo, faça um lanche, compartilhe o seu problema com alguém (com seu cão talvez) e então retorne ao problema com a mente descansada. Eu asseguro a você que uma vez que você aprenda os conceitos de programação neste livro, irá olhar para trás e perceber que tudo foi muito fácil, elegante e tão simples que tomou de você apenas um tempo para absorver o aprendizado.

1.12 Glossário

bug: Um erro em um programa.

unidade central de processamento: O coração de qualquer computador. É ela que executa o software que nós escrevemos; também chamada de "CPU" ou de "processador".

compilar: Traduzir um programa escrito em uma linguagem de alto nível em uma linguagem de baixo nível tudo de uma vez, em preparação para uma posterior execução.

linguagem de alto nível: Uma linguagem de programação como o Python que é desenhada para ser fácil para humanos ler e escrever.

modo interativo: Um modo de usar o interpretador Python digitando comandos e expressões no prompt.

interpretar: Executar um programa em uma linguagem de alto nível traduzindo uma linha por vez.

linguagem de baixo nível: Uma linguagem de programação que é desenhada para que seja fácil um computador executar; também chamada "código de máquina" ou "linguagem de montagem".

código de máquina: A linguagem mais baixo nível que pode existir em software, é a linguagem que é diretamente executada pela unidade central de processamento (CPU).

memória principal: Armazena programas e dados. A memória principal perde informação quando a energia é desligada.

parse: Examinar um programa e analisar a estrutura sintática.

portabilidade: Uma propriedade de um programa que roda em mais de um tipo de computador.

instrução print: Uma instrução que faz com que o interpretador Python exiba um valor na tela.

resolução de problema: O processo de formular um problema, encontrar a solução e a expressar.

programa: Um conjunto de instruções que especifica uma computação.

prompt: Quando um programa exibe uma mensagem e aguarda o usuário digitar algo para o programa.

memória secundária: Armazena programas e dados, retendo a informação mesmo quando a energia é desligada. Geralmente mais devagar em relação à memória principal. Exemplos de memória secundária são discos rígidos e memória flash nos pendrives USB.

semântica: O significado de um programa.

erro semântico: Um erro em um programa que faz algo diferente daquilo que o programador desejava.

código fonte: Um programa em uma linguagem de alto nível.

1.13 Exercícios

Exercício 1.1 Qual é a função da memória secundária em um computador?

a) Executar todas as computações e lógica de um programa
b) Obter páginas web da internet
c) Armazenar informação por um longo período – mesmo se faltar energia
d) Receber o input de um usuário

Exercício 1.2 O que é um programa?

Exercício 1.3 Qual é a diferença entre um compilador e um interpretador?

Exercício 1.4 Qual das opções a seguir contêm "código de máquina"?

a) O interpretador Python
b) O teclado
c) Arquivo de código fonte Python
d) Um documento do processador de texto

Exercício 1.5 O que está errado no código a seguir:

```
>>> primt 'Ola mundo!'
  File "<stdin>", line 1
    primt 'Ola mundo!'
                      ^
SyntaxError: invalid syntax
>>>
```

Exercício 1.6 Em qual lugar do computador existe uma variável "X" armazenada depois que a seguinte linha de Python finaliza?

```
x = 123
```

a) Unidade central de processamento
b) Memória Principal
c) Memória Secundária
d) Dispositivos de Entrada
e) Dispositivos de Saída

Exercício 1.7 O que o seguinte programa irá imprimir:

```
x = 43
x = x + 1
print x
```

a) 43
b) 44
c) x + 1
d) Um erro porque x = x + 1 não é matematicamente possível

Exercício 1.8 Explique cada item a seguir usando como exemplo uma capacidade humana: (1) Unidade central de processamento, (2) Memória principal, (3) Memória secundária, (4) Dispositivo de entrada, e (5) Dispositivo de saída. Por exemplo, "Qual é a capacidade humana equivalente a Unidade central de processamento"?

Exercício 1.9 Como se conserta um "Erro de Sintaxe"?

Capítulo 2

Variáveis, expressões e instruções

2.1 Valores e tipos

Um **valor** é uma das coisas básicas com a qual um programa trabalha, como uma letra ou um número. Os valores que vimos até agora são são `1`, `2`, and `'Ola, Mundo!'`

Estes valores pertencem a diferentes **tipos**: `2` é um inteiro, e `'Ola, Mundo!'` é uma **string**, assim chamada por conter uma "cadeia" de letras. Você (e o interpretador) podem identificar strings porque elas aparecem entre aspas.

A instrução `print` também funciona com inteiros. Nós usamos o comando `python` para iniciar o interpretador.

```
python
>>> print 4
4
```

Se você não tem certeza que tipo tem um valor, o interpretador pode te dizer.

```
>>> type('Ola, Mundo!')
<type 'str'>
>>> type(17)
<type 'int'>
```

Não surpreendentemente, strings pertencem ao tipo `str` e inteiros pertencem ao tipo `int`. Menos, obviamente, números com ponto decimal pertencem a um tipo chamado `float`, uma vez que estes números são representados em um formato chamado **ponto flutuante**.

```
>>> type(3.2)
<type 'float'>
```

E quanto a valores como `'17'` e `'3.2'`? Eles se parecem com números, mas eles são, quando entre aspas, strings.

```
>>> type('17')
<type 'str'>
>>> type('3.2')
<type 'str'>
```

Eles são strings.

Quando você digita um número inteiro grande, você pode ficar tentado a utilizar vírgulas entre os grupos de três dígitos, como em `1,000,000`. Este não é um número válido em Python, no entanto ele é válido:

```
>>> print 1,000,000
1 0 0
```

Bem, de toda forma, isto não é o que nós esperávamos! Python interpreta `1,000,000` como uma sequência de integers separados por vírgulas, o qual imprimi com espaços entre eles.

Este é o primeiro exemplo que vemos de um erro semântico: o código executa sem produzir uma mensagem de erro, mas ele não faz a coisa "certa".

2.2 Variáveis

Uma das mais poderosas características de uma linguagem de programação é a capacidade de manipular **variáveis**. Uma variável é um nome que se refere a um valor.

Um **comando de atribuição** cria novas variáveis e dá valores a elas:

```
>>> message = 'E agora algo completamente diferente'
>>> n = 17
>>> pi = 3.1415926535897931
```

Este exemplo faz três atribuições. O primeiro atribui uma string a uma nova variável chamada `message`; o segundo atribui o integer `17` à variável `n`; o terceiro atribui valor (aproximado) de π à variável `pi`.

Para mostrar o valor de uma variável, você pode usar o comando print.

```
>>> print n
17
>>> print pi
3.14159265359
```

O tipo de uma variável é o tipo do valor ao qual ela se refere.

```
>>> type(message)
<type 'str'>
>>> type(n)
<type 'int'>
>>> type(pi)
<type 'float'>
```

2.3 Nomes de variáveis e palavras reservadas

Programadores geralmente escolhem nomes, que tenham algum significado, para suas variáveis e documentam para qual finalidade a variável será utilizada.

Nomes de variáveis podem ser arbitrariamente longos. Eles podem conter tanto letras quanto números, porém eles não podem começar com um número. É válido usar letras maiúsculas, porém é uma boa prática começar o nome de uma variável com uma letra minúscula (você verá o porquê, mais tarde).

O caractere sublinhado (_) pode aparecer no nome. Ele é frequentemente usado em nomes com múltiplas palavras, como `my_name` ou `airvelocidade_of_unladen_swallow`.

Nomes de variáveis podem começar como caracter sublinhado, mas nós, geralmente, evitamos isto, a menos que estejamos escrevendo uma biblioteca de código para outros usarem.

Se você der a uma variável um nome inválido, você receberá um erro de sintaxe.

```
>>> 76trombones = 'grande desfile'
SyntaxError: invalid syntax
>>> more@ = 1000000
SyntaxError: invalid syntax
>>> class = 'Avancada Teoria Zymurgy'
SyntaxError: invalid syntax
```

`76trombones` é inválida porquê ela começa com um número. `more@` é inválida porquê ela contém um caractere inválido, `@`. Mas o quê há de errado com `class`? Acontece que a palavra `class` é uma Palavra Reservada do Python **keywords**. O interpretador usa as Palavras Reservadas para reconhecer a estrutura do programa, e elas não podem ser usadas como nomes de variáveis.

Python reserva 31 Palavras Reservadas [1] para seu uso:

```
and       del       from      not       while
as        elif      global    or        with
assert    else      if        pass      yield
break     except    import    print
class     exec      in        raise
continue  finally   is        return
def       for       lambda    try
```

Você pode querer manter esta lista ao alcance das mãos. Se o interpretador reclamar sobre um de seus nomes de variável e você não souber o porquê, verifique se ela se encontra nesta lista.

[1] Em Python 3.0, `exec` não é mais uma palavra reservada, mas `nonlocal` é.

2.4 Instruções

Uma **instrução** é uma unidade de código que o interpretador Python pode executar. Nós vimos dois tipos de instruções: impressão (print) e atribuição (=).

Quando você digita uma instrução no modo interativo, o interpretador a executa e mostra o resultado, se houver um.

Um script geralmente contém uma sequência de instruções. Se houver mais de uma instrução, os resultados aparecem um de cada vez conforme as instruções são executadas.

Por exemplo, o script

```
print 1
x = 2
print x
```

Produz a saída:

```
1
2
```

A instrução de atribuição não produz saída.

2.5 Operadores e operandos

Operadores são símbolos especiais que representam cálculos como adição e multiplicação. Os valores aos quais os operadores são aplicados são chamados de **operandos**.

Os operadores `+`, `-`, `*`, `/`, e `**` realizam, adição, subtração, mumltiplicação, divisão e exponenciação, como no exemplo a seguir:

```
20+32   hora-1   hora*60+minuto   minuto/60   5**2   (5+9)*(15-7)
```

O operador de divisão pode não fazer o que você espera:

```
>>> minuto = 59
>>> minuto/60
0
```

O valor de `minuto` é 59, e na aritmética convencional 59 dividido por 60 é 0.98333, não 0. A razão para esta discrepância é o fato de que o Python realiza um **floor division**[2]

Quando ambos os operandos são integers, o resultado é, também, um integer; floor division corta a parte fracionária, portanto, neste exemplo o resultado foi arredondado para zero.

[2]Em Python 3.0, o resultado desta divisão é do tipo `float`. Em Python 3.0, o novo operador `//` realiza uma divisão to tipo integer.

Se um dos operandos é um número do tipo ponto flutuante, Python realiza uma divisão de ponto flutuante, e o resultado é um `float`:

```
>>> minuto/60.0
0.98333333333333328
```

2.6 Expressões

Uma **expressão** é uma combinação de valores, variáveis e operadores. Um valor, por si só, é considerado uma expressão, e portanto, uma variável, então o que segue são todas expressões válidas (assumindo que a variável `x` tenha recebido um valor):

```
17
x
x + 17
```

Se você digita uma expressão no modo interativo, o interpretador a **avalia** e mostra o resultado:

```
>>> 1 + 1
2
```

Mas em um script, uma expressão por si só não faz nada! Isto é uma fonte comum de confusão para iniciantes.

Exercício 2.1 Digite a seguinte declaraçao no interpretador do Python para ver o que ele faz:

```
5
x = 5
x + 1
```

2.7 Ordem das operações

Quando mais de um operador aparece em uma expressão, a ordem de avaliação depende das **regras de precedência**. Para operadores matemáticos. Python segue a convenção matemática. O Acrônimo **PEMDAS** é uma modo útil para lembrar as regras:

- **P**arenteses têm a mais alta precedência e pode ser usado para forçar que uma expressão seja calculada na ordem que você deseja. Como as expressões entre parênteses são avalidas primeiro, `2 * (3-1)` é 4, e `(1+1)**(5-2)` é 8. Você também pode usar parênteses para tornar uma expressão mais fácil de ser lida, como em `(minute * 100) / 60`, mesmo que isto não mude o resultado.

- **E**xponenciação é a próxima precedência mais alta, `então 2**1+1` é 3, não 4, e `3*1**3` é 3, não 27.
- **M**ultiplicação e **D**ivisão têm a mesma precedência, a qual é mais alta que **A**dição e **S**ubtração, que também têm a mesma precedência entre si. Então `2*3-1` é 5, não 4, e `6+4/2` é 8, não 5.
- Operadores com a mesma precedência são avaliados da esquerda para direita. Portanto na expressão `5-3-1` é 1, não 3 pois o `5-3` acontence primeiro e então o `1` é subtraído de `2`.

Na dúvida, sempre utilize parênteses em suas expressões para ter certeza de que os cálculos serão realizados na ordem que você deseja.

2.8 O operador Módulo

O **operador módulo** funciona em integers e fornece o resto da divisão, quando o primeiro operando é dividido pelo segundo. No Python, o operador módulo é um sinal de percentual (`%`). A sintaxe é a mesma dos outros operadores:

```
>>> quociente = 7 / 3
>>> print quociente
2
>>> resto = 7 % 3
>>> print resto
1
```

Portanto, 7 dividido por 3 é igual a 2, com resto 1.

O operador módulo apresenta-se surpreendentemente útil. Por exemplo, você pode checar se um núnero é divisível por outro—se `x % y` é zero, então `x` é divivisível por `y`.

Você pode, também, testar se um número é dvisível por outro. Por exemplo, `x % 10` nos mostra se o número x é divísível por 10. Similarmente, `x % 100` nos mostra se x é divisível por 100.

2.9 Operações com Strings

O operador + funciona com strings, mas ele não é uma adição no sentido matemático. Ao invés disto, ele realiza **concatenação**, que significa juntar as strings, vinculando-as de ponta-a-ponta. Por exemplo:

```
>>> primeiro = 10
>>> segundo = 15
>>> print primeiro + segundo
25
>>> primeiro = '100'
```

```
>>> segundo = '150'
>>> print primeiro + segundo
100150
```

A saída deste programa é `100150`.

2.10 Solicitando dados de entrada para o usuário

Algumas vezes gostaríamos de solicitar, do usuário, o valor para uma variável por meio do teclado. Python fornece uma função interna chamada `raw_input` que recebe dados de entrada a partir do teclado[3]. Quando esta função é chamada, o programa para e espera para que o usuário digite algo. Quando o usuário pressiona o Return ou Enter, o programa continua e a função `raw_input` retorna o que o usuário digitou, como uma string.

```
>>> entrada = raw_input()
Alguma coisa boba
>>> print entrada
Alguma coisa boba
```

Antes de receber os dados de entrada do usuário, é uma boa idéia imprimir uma mensagem, dizendo ao usuário que o dado deve ser informado. Você pode passar uma string para a função `raw_input` para ser mostrada para o usuário antes da parada para a entrada de dados:

```
>>> nome = raw_input('Qual é o seu nome?\n')
Qual é o seu nome?
Chuck
>>> print nome
Chuck
```

A sequência `\n` no final da mensagem representa uma **nova linha**, que é um caractere especial que causa a quebra de linha. É por este motivo que os dados de entrada informados pelo usuário aparecem abaixo da mensagem.

Se você espera que o usuário digite um integer, você pode tentar converter o valor retornado para `int` usando a função `int()`:

```
>>> pergunta = 'Qual é ... a velocidade de uma andorinha sem carga?\n'
>>> velocidade = raw_input(pergunta)
Qual é ... a velocidade de uma andorinha sem carga?
17
>>> int(velocidade)
17
>>> int(velocidade) + 5
22
```

Porém, se o usuário digita algo diferente de um conjunto de números, você recebe um erro:

[3] Em Python 3.0, esta função é chamada de `input`.

```
>>> velocidade = raw_input(pergunta)
Qual é ... a velocidade de uma andorinha sem carga?
Que tipo de andorinha, uma Africana ou uma Européia?
>>> int(velocidade)
ValueError: invalid literal for int()
```

Nós veremos como tratar este tipo de erro mais tarde.

2.11 Comentários

Como os programas ficam maiores e mais complicados, eles ficam mais difíceis de serem lidos. Linguagens formais são densas, e muitas vezes é difícil olhar para um pedaço de código e descobrir o que ele está fazendo, ou porquê.

Por esta razão, é uma boa ideia adicionar notas em seus programas para explicar, em linguagem natural, o que o programa está fazendo. Estas notas são chamadas de **comentários**, e, em Python, elas começam com o símbolo #:

```
# computa a porcentagem de hora que se passou
porcentagem = (minuto * 100) / 60
```

Neste caso, o comentário aparece sozinho em uma linha. Você pode, também, colocar o comentário no final da linha:

```
porcentagem = (minuto * 100) / 60     # porcentagem de uma hora
```

Todos os caracteres depois do #, até o fim da linha são ignorados—eles não têm efeito sobre o programa. Comentários são mais úteis quando documentam características não obvias do código. É razoável assumir que o leitor pode descobrir *o que* o código faz; é muito mais útil explicar o *porquê*.

Este comentário é redundante e inútil dentro do código:

```
v = 5     # atribui o valor 5 para a variável v
```

Este comentário contem informações úteis que não estão no código.

```
v = 5     # velocidade em metros por segundo
```

Bons nomes de variáveis podem reduzir a necessidade de comentários, porém, nomes longos podem tornar expressões complexas difíceis de serem lidas, então devemos ponderar.

2.12 Escolhendo nomes de variáveis mnemônicos

Contanto que você siga as regras simples de nomenclatura de variáveis, e evite Palavras Reservadas, você tem muitas escolhas quando você nomeia suas variáveis. No início, esta escolha pode ser confusa, tanto quando você lê um programa,

quanto quando você escreve seus próprios programas. Por exemplo, os três programas a seguir são idênticos em termos do que realizam, mas muito diferente quando você os lê e tenta compreendê-los.

```
a = 35.0
b = 12.50
c = a * b
print c

horas = 35.0
taxa = 12.50
pagamento = horas * taxa
print pagamento

x1q3z9ahd = 35.0
x1q3z9afd = 12.50
x1q3p9afd = x1q3z9ahd * x1q3z9afd
print x1q3p9afd
```

O interpretador Python vê todos os três programas *exatamente como o mesmo*, mas os seres humanos veem e entendem esses programas de forma bastante diferente, entenderão mais rapidamente a **intenção** do segundo programa, porque o programador escolheu nomes de variáveis que refletem a sua intenção sobre os dados que serão armazenados em cada variável.

Nós chamamos esses nomes de variáveis sabiamente escolhidos de "nomes de variáveis mnemônicos". A palavra *mnemônico*[4] significa "auxiliar de memória". Nós escolhemos os nomes de variáveis mnemônicos para nos ajudar a lembrar o motivo pelo qual criamos a variável, em primeiro lugar.

Isso tudo soa muito bem, e é uma boa ideia usar nomes de variável mnemônicos, eles podem atrapalhar a capacidade de análise e entendimento do código de um programador iniciante. Isto acontece porque os programadores iniciantes ainda não memorizaram as palavras reservadas (existem apenas 31 delas) e, por vezes, variáveis que têm nomes muito descritivos podem parecer parte da linguagem e não apenas nomes de variáveis bem escolhidas.

Dê uma olhada rápida no seguinte exemplo de código Python que percorre alguns dados. Nós vamos falar sobre loops em breve, mas por agora apenas tente imaginar como isto funciona:

```
for palavra in palavras:
    print palavra
```

O que esta acontecendo aqui? Qual das palavras (for, palavra, in, etc.) são palavras reservadas e quais são apenas nomes de variáveis? O Python entende em um nível fundamental a noção de palavras? Programadores iniciantes têm dificuldade para separar quais partes do código *devem* ser o mesmo que este exemplo e que partes

[4] veja `http://en.wikipedia.org/wiki/Mnemonic` para uma descrição completa da palavra "mnemônico".

do código são simplesmente as escolhas feitas pelo programador. O código a seguir é equivalente ao código acima:

```
for pedaco in pizza:
    print pedaco
```

É mais fácil para o programador iniciante olhar para este código e saber quais partes são palavras reservadas definidas pelo Python e quais partes são, simplesmente, nomes de variáveis escolhidos pelo programador. É bastante claro que o Python não tem nenhuma compreensão fundamental de pizza e pedaços e o fato de que uma pizza é constituída por um conjunto de um ou mais pedaços.

Mas se o nosso programa é verdadeiramente sobre a leitura de dados e a procura de palavras nos dados, `pizza` e `pedaco` são nomes de variáveis não muito mnemônicos. Escolhê-los como nomes de variável, distorce o significado do programa. Depois de um período muito curto de tempo, você vai conhecer as palavras reservadas mais comuns, então vai começar a ver as palavras reservadas saltando para você: **for** `palavra` **in** `palavras`**:**
    print `palavra`

As partes do código que são definidas pelo Python (`for`, `in`, `print`, and `:`) estão em negrito e as variáveis escolhidas pelo programador (`word` and `words`) não estão em negrito. Muitos editores de textos compreendem a sintaxe do Python e vão colorir palavras reservadas de forma diferente para dar a você pistas e manter suas variáveis e palavras reservadas separadas. Depois de um tempo você começará a ler o Python e rapidamente determinar o que é uma variável e o que é uma palavra reservada.

2.13 Debugando

Neste ponto, o erro de sintaxe que você está mais propenso a cometer é um nome de variável ilegal, como `class` e `yield`, que são palavras reservadas ou `emprego~estranho` e `RS$`, que contêm caracteres não permitidos.

Se você colocar um espaço em um nome de variável, o Python interpreta que são dois operandos sem um operador:

```
>>> nome ruim = 5
SyntaxError: invalid syntax
```

Para erros de sintaxe, as mensagens de erro não ajudam muito. As mensagens mais comuns são `SyntaxError: invalid syntax` and `SyntaxError: invalid token`, nenhuma das quais é muito informativa.

O erro de execução que você está mais propenso a a cometer é “use before def;”, isto é, tentando usar uma variável antes de atribuir um valor. Isso pode acontecer se você digitar um nome de variável errado:

```
>>> principal = 327.68
>>> interesse = principal * taxa
NameError: name 'taxa' is not defined
```

Nomes de variáveis são sensíveis a maiúsculo e minúsculo, desta forma, `LaTeX` não é o mesmo que `latex`.

Neste ponto, a causa mais provável de um erro de semântica é a ordem das operações. Por exemplo, para calcular $\frac{1}{2\pi}$, você pode ser tentado a escrever

```
>>> 1.0 / 2.0 * pi
```

Mas a divisão acontece primeiro, então você iria ficar com $\pi/2$, que não é a mesma coisa! Não há nenhuma maneira de o Python saber o que você quis escrever, então, neste caso você não receberia uma mensagem de erro; você apenas receberia uma resposta errada.

2.14 Glossário

atribuição: Uma instrução que atribui um valor a uma variável.

concatenar: Para juntar dois operandos ponta-a-ponta.

Comentário : Informação em um programa que é destinado a outros programadores (ou qualquer pessoa lendo o código fonte) e não tem qualquer efeito sobre a execução do programa.

Avaliar: Para simplificar uma expressão realizando as operações, a fim de se obter um único valor.

Expressão: Uma combinação de variáveis, operadores e valores que representa um valor de resultado único.

Ponto Flutuante: Um tipo que representa números com partes fracionárias.

Floor Division: A operação que divide dois números e corta a parte fracionária.

Integer: Um tipo que representa números inteiros.

Palavra Reservada: Uma palavra reservada usada pelo compilador para analisar um programa; você não pode usar palavras reservadas como `if`, `def`, e `while` como nomes de variáveis.

Mnemônico: Um auxiliar de memória. Nós, muitas vezes, damos nomes mnemônicos a variáveis para nos ajudar lembrar o que está armazenado na mesma.

Operador módulo: Um operador, denotado pelo sinal de porcentagem (`%`), que funciona em inteiros e produz o restante quando um número é dividido por outro.

Operando: Um dos valores sobre os quais um operador opera.

Operador: Um símbolo especial que representa uma cálculo simples, como adição, multiplicação ou concatenação de strings.

Regras de precedência: O conjunto de regras que regem a ordem na qual as expressões, envolvendo múltiplos operadores e operandos, são avaliadas.

Instrução: Uma seção de código que representa um comando ou ação. Até o momento, as instruções que temos visto são instruções de atribuição e impressão.

String: Um tipo que representa sequências de caracteres.

Tipo: Uma categoria de valores. Os tipos que vimos até o momento são inteiros (tipo `int`), números de ponto flutuante (tipo `float`) e strings (tipo `str`).

valor: Uma das unidades básicas de dados, como um número ou string, que um programa manipula.

variável: Um nome que se refere a um valor.

2.15 Exercícios

Exercício 2.2 Escreva um programa que utiliza `raw_input` para solicitar a um usuário o nome dele, em seguida, saudá-lo.

```
Digite o seu nome: Chuck
Ola Chuck
```

Exercício 2.3 Escreva um programa para solicitar ao usuário por, horas e taxa por hora, e então, calcular salário bruto.

```
Digite as horas: 35
Digite a taxa: 2.75
Pagamento: 96.25
```

Não estamos preocupados em fazer com que o nosso pagamento tenha exatamente dois dígitos depois da vírgula, por enquanto. Se você quiser, pode brincar com a função `round` do Python para adequadamente arredondar o pagamento resultante com duas casas decimais.

Exercício 2.4 Suponha que nós executamos as seguintes instruções de atribuição:

```
comprimento = 17
altura = 12.0
```

Para cada uma das seguintes expressões, escrever o valor da expressão e o tipo (do valor da expressão).

1. `comprimento/2`
2. `comprimento/2.0`
3. `altura/3`
4. `1 + 2 * 5`

Utilize o interpretador do Python para conferir suas respostas.

Exercício 2.5 Escreva um programa que pede ao usuário por uma temperatura Celsius, converter a temperatura para Fahrenheit e imprimir a temperatura convertida.

Capítulo 3

Execução Condicional

3.1 Expressões booleanas

Uma **expressão booleana** é uma expressão que é true ou false. Os seguintes exemplos usam o operador ==, que compara dois operadores e produz `True` se eles forem iguais e `False` caso contrário:

```
>>> 5 == 5
True
>>> 5 == 6
False
```

`True` e `False` são valores especiais que pertencem ao tipo `bool`; eles não são strings:

```
>>> type(True)
<type 'bool'>
>>> type(False)
<type 'bool'>
```

O operador == é um dos **operadores de comparação**; os outros são:

```
x != y               # x não é igual a y
x > y                # x é maior que y
x < y                # x é menor que y
x >= y               # x é maior ou igual a y
x <= y               # x é menor ou igual a y
x is y               # x é o mesmo que y
x is not y           # x não é o mesmo que y
```

Embora estas operações sejam provavelmente familiar para você, os simbolos Python são diferentes dos símbolos matemáticos para a mesma operação. Um erro comum é usar um único sinal de igual (=) em vez de um sinal de igual duplo (==). Lembre-se que o = é um operador de atribuição e == é um operador de comparação. Não existe tal coisa como =< ou =>.

3.2 Operador Lógico

Existem três **operadores lógicos**: `and`, `or`, and `not`. A semântica (significado) destes operadores é semelhante ao seu significado em inglês. Por exemplo,

```
x > 0 and x < 10
```

só é verdade se `x` for maior que 0 *e* menor que 10.

`n%2 == 0 or n%3 == 0` é verdadeiro se *qualquer* uma das condições é verdadeira, isto é, se o número é divisível por 2 *ou* 3.

Finalmente, o operador `not` nega uma expressão booleana, então `not (x > y)` é verdadeiro se `x > y` é falso; isto é, se `x` é menor do que ou igual a `y`.

Rigorosamente falando, os operandos dos operadores logicos devem ser expressões booleanas, mas Python não é muito rigoroso. Qualquer numero diferente de zero é interpretado como "verdadeiro."

```
>>> 17 and True
True
```

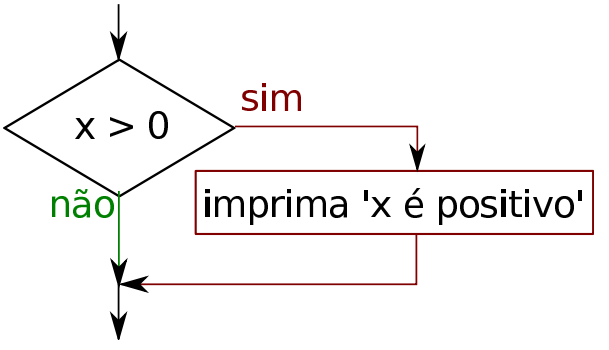
Esta flexibilidade pode ser útil, mas existem algumas sutilezas que podem confundir o Python. Você pode querer evitá-los até você ter certeza que sabe o que está fazendo.

3.3 Execução condicional

Para escrever programas úteis, quase sempre precisamos da capacidade para verificar as condições e mudar o comportamento do programa em conformidade. **Instruções condicionais** nos dão essa capacidade. A forma mais simples é a instrução `if`:

```
if x > 0 :
    imprima 'x é positivo'
```

A expressão booleana depois da declaração `if` é chamado de **condição**. Terminamos a instrução `if` com um caractere dois pontos (:) e a(s) linha(s) após a instrução if são identadas.



Se a condição lógica é verdadeira, então a instrução identada é executada. Se a condição lógica é falsa, a instrução identada é ignorada.

Instruções `if` têm a mesma estrutura que as definições de funções ou loops `for` [1]. A instrução é composta por uma linha de cabeçalho que termina com o caractere dois pontos (:) seguido por um bloco identado. Instruções como esta são chamadas **declarações compostas** porque elas são compostas por mais de uma linha.

Não há limite para o número de intruções que podem aparecer no corpo, mas deve haver pelo menos uma. Às vezes, é util ter um corpo sem instruções (usualmente como um corpo pacificador para o código que você não tenha escrito até o momento). Nesse caso, você pode usar a instrução `pass`, que não faz nada.

```
if x < 0 :
    pass          # precisa lidar com valores negativos!
```

Se você digitar um `if` no interpretador Python, o prompt vai se modificar de três sinais ¿¿¿ para três pontos ... para indicar que você está no meio de um bloco de declarações, como mostrado abaixo:

```
>>> x = 3
>>> if x < 10:
...     print 'pequeno'
...
Small
>>>
```

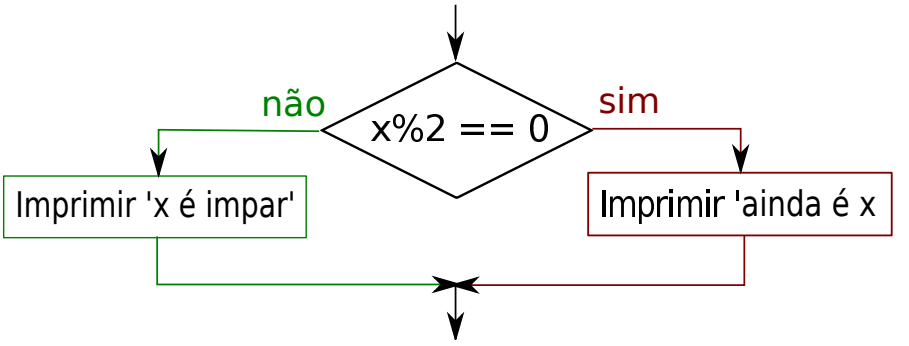
3.4 Execução alternativa

A segunda forma da instrução `if` é a **execução alternativa**, na qual há duas possibilidades e a condição determina qual delas será executada. A sintaxe se parece com esta:

```
if x%2 == 0 :
    print 'x ainda é'
else :
    print 'x é estranho'
```

Se o resto da divisão de `x` por 2 for 0, nós sabemos que `x` é divisível, e o programa exibe uma mensagem para esse efeito. Se a condição for falsa, o segundo conjunto de instruções é executado.

[1] Vamos aprender sobre as funções no Capítulo 4 e loops no Capítulo 5.



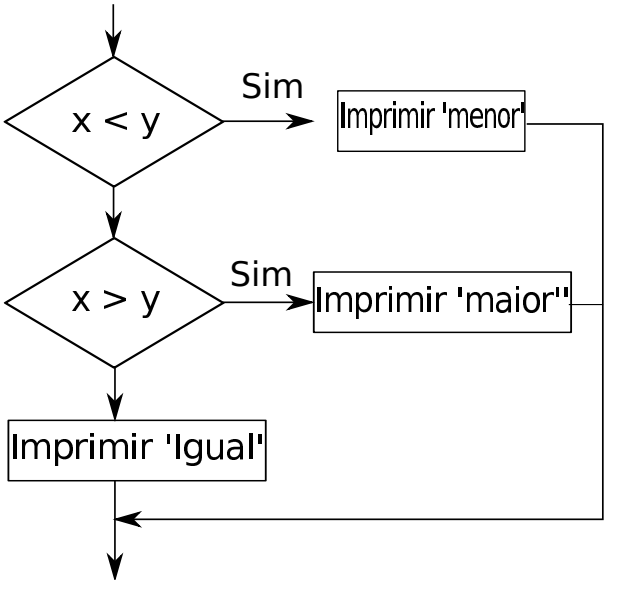
Uma vez que a condição deve ser verdadeira ou falsa, exatamente uma das alternativas será executada. As alternativas são chamadas de **branches**, porque elas dividem o fluxo de execução.

3.5 Condicionais encadeadas

Às vezes, há mais de duas possibilidades e precisamos de mais do que duas condições. Uma maneira de expressar uma computação como essa é uma **condição encadeada**:

```
if x < y:
    print 'x é menor que y'
elif x > y:
    print 'x é maior que y'
else:
    print 'x e y são iguais'
```

`elif` é uma abreviação de " else if. " Mais uma vez, exatamente uma condição será executada.



Não há limite para o número de instruções `elif`. Se houver uma cláusula `else`, ela deve estar no final, mas só pode existir uma única instrução deste tipo.

```
if choice == 'a':
    print 'Escolha ruim'
elif choice == 'b':
    print 'Boa escolha'
elif choice == 'c':
    print 'Perto, mas não correto'
```

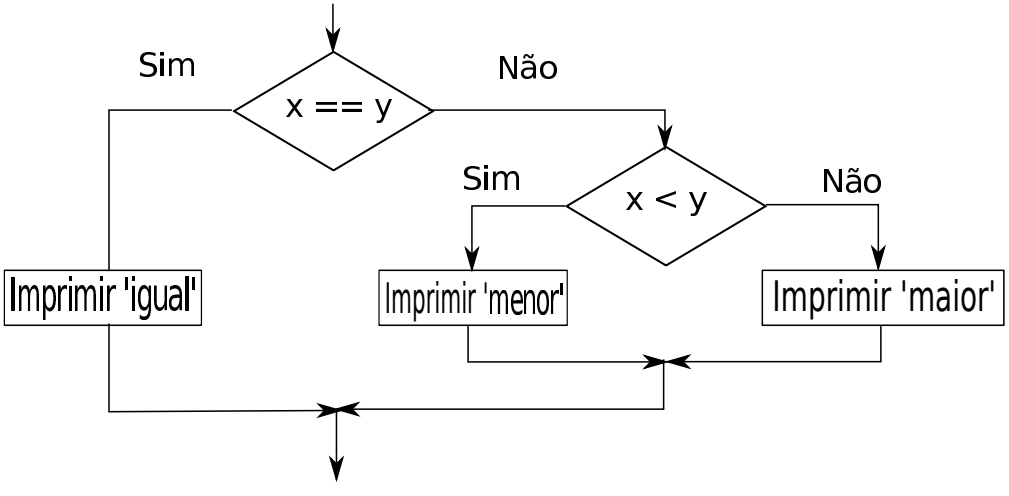
Cada condição é verificada em ordem. Se a primeira é falsa, a próxima será avaliada, e assim por diante. Se um deles é verdadeiro, o fluxo correspondente será executado, e a instrução termina. Mesmo se mais do que uma condição for verdadeira, apenas o primeiro fluxo verdadeiro é executado.

3.6 Condicionais aninhados

Uma instrução condicional também pode ser aninhada dentro de outra. Nós poderíamos ter escrito o exemplo de três ramificações como a seguir:

```
if x == y:
    print 'x e y são iguais'
else:
    if x < y:
        print 'x é menor que y'
    else:
        print 'x é maior que y'
```

A condicional externa contém duas ramificações. A primeira ramificação contém uma instrução simples. A segunda ramificação contém outra instrução `if`, que contém duas ramificações próprias. Aquelas duas ramificações são ambas instruções simples, embora pudessem ter sido instruções condicionais também.



Embora a identação das instruções torna a estrutura visível, **condicionais aninhadas** fica difícil de ler muito rapidamente. Em geral, é uma boa idéia evitá-las

sempre que possível.

Os operadores lógicos muitas vezes fornecem uma maneira de simplificar as instruções condicionais aninhadas. Por exemplo, podemos reescrever o código a seguir usando um condicional simples:

```
if 0 < x:
    if x < 10:
        print 'x é um número positivo de um dígito.'
```

A instrução `print` é executada somente se ambas as condições forem verdadeiras, para que possamos obter o mesmo efeito com o operador `and`:

```
if 0 < x and x < 10:
    print 'x é um número positivo de um dígito.'
```

3.7 Capturando exceções usando try e except

Anteriormente, vimos um segmento de código onde foram utilizadas as funções `raw_input` e `int` para ler e validar um número inteiro informado pelo usuário. Também vimos como pode ser traiçoeiro utilizar isso:

```
>>> speed = raw_input(prompt)
Qual é ... a velocidade aerodinâmica de uma andorinha sem carga?
Você quer saber, uma andorinha Africana ou Européia?
>>> int(speed)
ValueError: invalid literal for int()
>>>
```

Quando estamos executando estas instruções no interpretador Python, temos um novo prompt do interpretador, acho que "oops", e move-se para a próxima instrução.

No entanto, se você colocar esse código em um script Python e este erro ocorrer, seu script para imediatamente e nos retorna sua pilha de execução. Não foi executada a seguinte instrução.

Aqui está um programa de exemplo para converter uma temperatura Fahrenheit para uma temperatura em graus Celsius:

```
inp = raw_input('Digite a Temperatura Fahrenheit:')
fahr = float(inp)
cel = (fahr - 32.0) * 5.0 / 9.0
print cel
```

Se nós executarmos este código e informarmos uma entrada inválida, ele simplesmente falha com uma mensagem de erro não amigável:

```
python fahren.py
Digite a Temperatura Fahrenheit:72
22.2222222222
```

```
python fahren.py
Digite a Temperatura Fahrenheit:fred
Traceback (most recent call last):
  File "fahren.py", line 2, in <module>
    fahr = float(inp)
ValueError: invalid literal for float(): fred
```

Existe uma estrutura de execução condicional do Python para lidar com esses tipos esperados e inesperados de erros chamados "try / except". A ideia de `try` e `except` é a de que você saiba que alguma sequência de instrução pode ter algum problema e você queira adicionar algumas instruções para serem executadas, caso um erro ocorra. Estas instruções adicionais (dentro do bloco except) são ignoradas se não ocorrer um erro.

Você pode associar os recursos `try` e `except` do Python como sendo uma "política segura" em uma seqüência de instruções.

Podemos reescrever nosso conversor de temperaturas da seguinte forma:

```
inp = raw_input('Digite a Temperatura Fahrenheit:')
try:
    fahr = float(inp)
    cel = (fahr - 32.0) * 5.0 / 9.0
    print cel
except:
    print 'Por favor, digite um numero'
```

Python começa executando a sequência de instruções dentro do bloco `try`. Se tudo correr bem, ele ignora o bloco `except` e prossegue. Se uma exceção ocorre no bloco `try`, o Python pula para fora do bloco `try` e executa a sequência de instruções do bloco `except`.

```
python fahren2.py
Digite a Temperatura Fahrenheit:72
22.2222222222

python fahren2.py
Digite a Temperatura Fahrenheit:fred
Por favor, digite um numero
```

Tratar uma exceção com uma instrução `try` é chamado de **capturar** uma exceção. Neste exemplo, a cláusula `except` imprime uma mensagem de erro. Em geral, capturar uma exceção oferece a oportunidade de corrigir o problema, ou tentar novamente, ou pelo menos terminar o programa elegantemente.

3.8 Short-circuit avaliação de expressões lógicas

Quando o Python está processando uma expressão lógica, tal como `x >= 2 e (x / y) > 2`, ele avalia a expressão da esquerda para a direita. Devido à definição do `and`, se `x` é inferior a 2, a expressão `x >= 2` é `False` e assim toda a expressão

é `False` independentemente de saber se `(x / y) > 2` é avaliada como `True` ou `False`.

Quando o Python detecta que não existe nenhum ganho em se avaliar o resto de uma expressão lógica, ele para a sua avaliação e não faz os cálculos para o restante da expressão lógica. Quando a avaliação de uma expressão lógica para porque o valor global já é conhecido, a avaliação é chamada de **short-circuiting**.

Embora esta técnica pareça ter pouca importância, o comportamento de short-circuit leva a uma técnica inteligente chamada **guardian pattern**. Considere a seguinte sequência de código no interpretador Python:

```
>>> x = 6
>>> y = 2
>>> x >= 2 and (x/y) > 2
True
>>> x = 1
>>> y = 0
>>> x >= 2 and (x/y) > 2
False
>>> x = 6
>>> y = 0
>>> x >= 2 and (x/y) > 2
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
ZeroDivisionError: integer division or modulo by zero
>>>
```

O terceiro cálculo falhou porque o Python estava avaliando `(x/y)` e `y` foi zero, o que causou um erro de execução. Mas o segundo exemplo *não* falhou porque a primeira parte da expressão `x >= 2` foi avaliada como `False` então a expressão `(x/y)` não foi executada devido à regra **short-circuit** e não houve erro.

Podemos construir a expressão lógica para colocar estrategicamente uma avaliação do tipo **guardian pattern** antes da avaliação que pode causar um erro, como segue:

```
>>> x = 1
>>> y = 0
>>> x >= 2 and y != 0 and (x/y) > 2
False
>>> x = 6
>>> y = 0
>>> x >= 2 and y != 0 and (x/y) > 2
False
>>> x >= 2 and (x/y) > 2 and y != 0
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
ZeroDivisionError: integer division or modulo by zero
>>>
```

Na primeira expressão lógica, `x >= 2` é `False`, então a avaliação para no `and`. Na segunda expressão lógica, `x >= 2` é `True` mas `y != 0` é `False` então nunca chegamos a avaliar a expressão `(x/y)`.

Na terceira expressão lógica, o `y != 0` encontra-se *depois* do cálculo `(x/y)` de modo que a expressão termina com um erro.

Na segunda expressão, dizemos que `y != 0` atua como um **guard** para garantir que só executaremos `(x/y)` se `y` for diferente de zero.

3.9 Depuração

O Python traceback é exibido quando ocorre um erro, ele contém diversas informações, mas pode ser um pouco confuso com tantos dados. A maioria das informações úteis geralmente são:

- Que tipo de erro ocorreu, e
- Onde ocorreu.

Erros de sintaxe geralmente são fáceis de encontrar, mas há algumas pegadinhas. Erros por espaço em branco podem ser difíceis, porque os espaços e tabs são invisíveis e geralmente os ignoramos.

```
>>> x = 5
>>>  y = 6
  File "<stdin>", line 1
    y = 6
    ^
SyntaxError: invalid syntax
```

Neste exemplo, o problema é que a segunda linha é indentada por um espaço. Mas a mensagem de erro aponta para `y`, que é enganosa. Em geral, as mensagens de erro indicam onde o problema foi descoberto, mas o erro real pode estar no início do código, às vezes em uma linha anterior.

O mesmo ocorre para erros de execução. Suponha que você está tentando calcular uma relação sinal-ruído em decibéis. A fórmula é $SNR_{db} = 10\log_{10}(P_{signal}/P_{noise})$. Em Python, você pode escrever algo como isto:

```
import math
signal_power = 9
noise_power = 10
ratio = signal_power / noise_power
decibels = 10 * math.log10(ratio)
print decibels
```

Mas quando você executá-lo, você recebe uma mensagem de erro [2]:

```
Traceback (most recent call last):
  File "snr.py", line 5, in ?
    decibels = 10 * math.log10(ratio)
OverflowError: math range error
```

[2] Em Python 3.0, você não recebe uma mensagem de erro; o operador de divisão executa a divisão de ponto flutuante, mesmo com operandos do tipo inteiro.

A mensagem de erro indica a linha 5, mas não há nada errado com essa linha. Para encontrar o verdadeiro erro, pode ser útil imprimir o valor da variável `ratio`, que daria 0. O problema está na linha 4, porque dividir dois inteiros causa "floor division"". A solução é representar a potência do sinal e potência de ruído com valores de ponto flutuante.

Em geral, mensagens de erro dizem onde o problema foi descoberto, mas frequentemente não dizem onde ele foi causado.

3.10 Glossário

body: Uma sequência de instruções dentro de uma instrução composta

boolean expression: Uma expressão cujo valor é `True` ou `False`.

branch: Uma das sequências alternativas de instruções em uma instrução condicional.

condicional encadeada: Uma instrução condicional com uma série de ramificações alternativas.

operador de comparação: Um dos operadores que compara seus operandos: ==, !=, >, <, >=, and <=.

instrução condicional: Uma declaração que controla o fluxo de execução dependendo de alguma condição

condição: A expressão booleana em uma declaração condicional que determina qual a condição é executado.

instrução composta: Uma declaração que consiste de um cabeçalho e um corpo. O cabeçalho termina com dois pontos (:). O corpo é identado em relação ao cabeçalho.

guardian pattern: Onde nós construimos uma expressão lógica com comparações adicionais para aproveitar o comportamento de short-circuit.

operador lógico: Um dos operadores que combina expressões booleanas: `and`, `or`, e `not`.

condicional aninhada: Uma instrução condicional que aparece em um dos ramos de uma outra instrução condicional.

traceback: Uma lista das funções que estão em execução, impressa quando ocorre uma exceção.

short circuit: Quando o Python deixa de avaliar uma expressão lógica atéo final e para porque já sabe o valor final para a expressão sem a necessidade de avaliar o resto da expressão.

3.11 Exercícios

Exercício 3.1 Reescrever o seu cálculo de pagamento para dar ao trabalhador 1.5 vezes o valor da hora para horas trabalhadas acima de 40 horas.

```
Digite as Horas: 45
Digite a Taxa: 10
Pagamento: 475.0
```

Exercício 3.2 Reescrever seu programa de pagamento usando `try` e `except` para que o programa trate entradas não numérica amigavelmente imprimindo uma mensagem e saindo do programa. A seguir mostra duas execuções do programa:

```
Digite as Horas: 20
Digite a Taxa: nove
Erro, por favor, informe entrada numérica

Digite as Horas: quarenta
Erro, por favor, informe entrada numérica
```

Exercício 3.3 Escreva um programa para solicitar uma pontuação entre 0.0 e 1.0. Se o resultado estiver fora da faixa, imprimir uma mensagem de erro. Se a pontuação estiver entre 0.0 e 1.0, imprimir uma nota utilizando a seguinte tabela:

```
Ponto   Nota
>= 0.9   A
>= 0.8   B
>= 0.7   C
>= 0.6   D
< 0.6    F

Digite a Pontuação: 0.95
A

Digite a Pontuação: perfeito
Pontuação incorreta

Digite a Pontuação: 10.0
Pontuação incorreta

Digite a Pontuação: 0.75
C

Digite a Pontuação: 0.5
F
```

Executar o programa repetidamente, como mostrado acima, para testar os diversos resultados para as diferentes entradas.

Capítulo 4

Funções

4.1 Chamadas de funções

Em programação, uma **função** é uma sequência de condições que executa uma tarefa. Quando você define uma função, você especifica o nome e a sequência de condições. Posteriormente, você pode "chamar" a função pelo nome. Nós já vimos um exemplo de **chamada de função**:

```
>>> type(32)
<type 'int'>
```

O nome da função é `type`. A expressão em parênteses é chamada de **argumento** da função. O argumento é um valor ou variável que passamos como entrada para a função. O resultado, para a função `type`, é o tipo do argumento.

É comum dizer que uma função "recebe" um arqumento e "retorna" um resultado. O resultado é chamado de **valor de retorno**.

4.2 Funções embutidas ("baterias inclusas")

O Python provê um grande número de funções embutidas importantes que podemos utilizar sem a necessidade de definir como novas funções. Os criadores de Python escreveram um conjunto de funções para a resoluções de problemas comuns e incluiram-nas no Python para que as utilizássemos.

As funções `max` e `min` nos dão o maior e o menor valor em uma lista, respectivamente:

```
>>> max('Hello world')
'w'
>>> min('Hello world')
' '
>>>
```

A função `max` retorna o "maior caractere" quando usado com *strings* (que acaba sendo a letra "w") e função `min` retorna o menor caractere (que é o espaço).

Outra função muito comum é a função `len` que nos diz quantos ítens tem no seu argumento. Se o argumento do `len` é uma *string*, ela vai retornar o número de caracteres na *string*.

```
>>> len('Hello world')
11
>>>
```

Estas funções não estão limitadas ao uso com *strings*. Elas podem ser utilizadas em qualquer conjunto de valores, como veremos nos próximos capítulos.

Você deve tratar os nomes das funções embutidas como palavras reservadas (i.e., evitando utilizar "max" como nome de variável).

4.3 Funções de conversões de tipos

Python também provê funções para converter valores de um tipo para outro. A função `int` pega um valor e converte para um inteiro, se ela conseguir, caso contrário ela vai "reclamar":

```
>>> int('32')
32
>>> int('Hello')
ValueError: invalid literal for int(): Hello
```

A função `int` pode converter um ponto-flutuante para um inteiro, mas ela não arredonda; ela somente corta a parte fracionária:

```
>>> int(3.99999)
3
>>> int(-2.3)
-2
```

A função `float` converte inteiros e *strings* para pontos-flutuantes:

```
>>> float(32)
32.0
>>> float('3.14159')
3.14159
```

E por fim, `str` converte seu arqumento para uma *string*:

```
>>> str(32)
'32'
>>> str(3.14159)
'3.14159'
```

4.4 Números aleatórios

Dada a mesma entrada, a maioria dos programas de computadores geram sempre a mesma saída, por isso são chamados de **determinísticos**. Determinismo é normalmente uma coisa boa, uma vez que esperamos que o mesmo cálculo retorne o mesmo resultado. Para algumas aplicações, no entanto, nós desejamos que o computador seja imprevisível. Os jogos são exemplos óbvios, mas existem outros.

Fazer um programa realmente não-determinístico não é uma coisa tão fácil, mas existem formas de fazê-lo ao menos parecer não-determinístico. Umas das formas é utilizar **algoritmos** que geram números pseudoaleatórios. Números pseudoaleatórios não são realmente aleatórios porque são gerados por uma computação determinística, mas somente olhando para os números é quase impossível distinguilos de números aleatórios.

O módulo `random` provê funções que geram números pseudoaleatórios (que eu vou chamar simplesmente de "aleatórios" a partir de agora).

A função `random` retorna um *float* aleatório entre 0.0 e 1.0 (incluindo o 0.0, mas não o 1.0). Cada vez que a função `random` é chamada você obtém um número com uma grande série. Para ver um exemplo disto, execute este laço:

```
import random

for i in range(10):
    x = random.random()
    print x
```

Este programa produziu a seguinte lista de 10 números aleatórios entre 0.0 e até, mas não incluindo, o 1.0.

```
0.301927091705
0.513787075867
0.319470430881
0.285145917252
0.839069045123
0.322027080731
0.550722110248
0.366591677812
0.396981483964
0.838116437404
```

Exercício 4.1 Execute o programa em seu computador e veja quais números você obtém. Execute o programa mais de uma vez no seu computador e veja quais números você obtém.

A função `random` é uma de muitas funções que tratam números aleatórios. A função `randint` usa como parâmetros `baixo` e `alto`, e retorna um inteiro entre estes números (incluindo ambos os números passados).

```
>>> random.randint(5, 10)
5
>>> random.randint(5, 10)
9
```

Para escolher um elemento de uma sequência aleatória, você pode utilizar a função `choice`:

```
>>> t = [1, 2, 3]
>>> random.choice(t)
2
>>> random.choice(t)
3
```

O módulo `random` também provê funções para geração de valores, distribuições contínuas incluindo Gaussianas, exponenciais, gama e algumas outras.

4.5 Funções matemáticas

Python tem o módulo `math` que provê as funções matemáticas mais conhecidas. Antes de utilizar o módulo, temos que importá-lo:

```
>>> import math
```

Esta declaração cria um módulo objeto chamado math. Se você exibir o objeto módulo, obterá algumas informações sobre ele:

```
>>> print math
<module 'math' from '/usr/lib/python2.5/lib-dynload/math.so'>
```

O módulo contém funções e variáveis definidas. Para acessar umas destas funções, tem que especificar o nome do módulo e o nome da função, separados por um ponto (também conhecido como período). Este formato é conhecido como **notação de ponto**.

```
>>> ratio = signal_power / noise_power
>>> decibels = 10 * math.log10(ratio)

>>> radians = 0.7
>>> height = math.sin(radians)
```

O primeiro exemplo calcula o logaritmo de base 10 da relação sinal-ruído. O módulo `math` também provê uma função chamada `log` que calcula o logaritmo de base `e`.

O segundo exemplo descobre o seno de `radianos`. O nome da variável é uma dica para informar que o `sin` e as outras funções trigonométricas (`cos`, `tan`, etc.) recebem como argumento valores em radianos. Para converter de graus para radianos, divide-se o valor por 360 e multiplica-se por 2π:

```
>>> degrees = 45
>>> radians = degrees / 360.0 * 2 * math.pi
>>> math.sin(radians)
0.707106781187
```

A expressão `math.pi` pega a variável `pi` do módulo `math`. O valor desta variável é uma aproximação do π, em exatos 15 dígitos.

Se você conhece trigometria, você pode verificar o resultado anterior comparando-o a raiz de 2 dividido por 2:

```
>>> math.sqrt(2) / 2.0
0.707106781187
```

4.6 Adicionando novas funções

Até agora, utilizamos somente funções que já estão no Python, mas também é possível adicionar novas funções. Uma definição de uma função especifica o nome de uma nova função e a sequência das condições que serão executadas quando a função é chamada. Uma vez definida a função, podemos reutilizá-la diversas vezes em nosso programa.

Aqui temos um exemplo:

```
def print_lyrics():
    print "I'm a lumberjack, and I'm okay."
    print 'I sleep all night and I work all day.'
```

A palavra-chave `def` indica o início de uma função. O nome da função é `print_lyrics`. As regras para nomes de função são os mesmos das variavéis: letras, números e alguns caracteres especiais, mas o primeiro caractere não pode ser um número. Você não pode usar uma palavra-chave para o nome de uma função, e deve evitar ter uma variável e uma função com o mesmo nome.

A primeira linha em uma função é chamada de **header** (cabeçalho); o resto é chamado de **body** (corpo). O cabeçalho tem que terminar com o sinal de dois pontos **:** e o corpo deve ser indentado. Por convenção, a indentação são sempre 4 espaços. O corpo pode ter um número indefinido de declarações.

A cadeia de caracteres na declaração *print* são delimitadas entre aspas. Aspas simples e aspas duplas tem o mesmo resultado; a maioria das pessoas utiliza aspas simples exceto nos casos onde uma aspas simples (que também é um apóstrofe) aparece na cadeia de caracteres.

Se você for escrever uma função no modo interativo (*Python shell*), o interpretador irá exibir pontos (...) para que você perceba que a definição da função está incompleta:

```
>>> def print_lyrics():
...     print "I'm a lumberjack, and I'm okay."
...     print 'I sleep all night and I work all day.'
...
```

Para terminar uma função, você precisa inserir uma linha vazia (isto não é necessário em um script).

Ao definir uma função, se cria uma variável com o mesmo nome.

```
>>> print print_lyrics
<function print_lyrics at 0xb7e99e9c>
>>> print type(print_lyrics)
<type 'function'>
```

O valor de `print_lyrics` é uma **função objeto**, que tem o tipo `'function'`.

A sintaxe para chamar a nova função é a mesma para as funções embutidas:

```
>>> print_lyrics()
I'm a lumberjack, and I'm okay.
I sleep all night and I work all day.
```

Uma vez definida uma função, você pode utilizá-la dentro de outra função. Por exemplo, para repetir o refrão anterior, podemos escrever uma função chamada `repeat_lyrics`:

```
def repeat_lyrics():
    print_lyrics()
    print_lyrics()
```

E então chamá-la `repeat_lyrics`:

```
>>> repeat_lyrics()
I'm a lumberjack, and I'm okay.
I sleep all night and I work all day.
I'm a lumberjack, and I'm okay.
I sleep all night and I work all day.
```

Mas isto não é realmente como a música toca.

4.7 Definitions and uses

4.8 Definições e usos

Colocando juntos as partes do código da seção anterior, o programa inteiro se parece com isto:

```
def print_lyrics():
    print "I'm a lumberjack, and I'm okay."
    print 'I sleep all night and I work all day.'
```

```
def repeat_lyrics():
    print_lyrics()
    print_lyrics()

repeat_lyrics()
```

Este programa contém duas funções definidas: `print_lyrics` e `repeat_lyrics`. Funções definidas são executadas da mesma forma como outras declarações, mas o efeito é a criação de funções objetos. As declarações dentro de uma função não são executadas até que a função seja chamada, e a definição de uma função não gera um resultado de saída.

Como você deve imaginar, primeiro é necessário criar uma função antes de executá-la. Em outras palavras, a definição de uma função tem que ser realizada antes da primeira vez que esta função é chamada.

Exercício 4.2 Mova a última linha deste programa para o início, de forma que a chamada da função esteja antes da definição da mesma. Execute o programa e veja a mensagem de erro que aparecerá.

Exercício 4.3 Mova a chamada da função para a última linha e mova a definição da função `print_lyrics` para depois da definição da função `repeat_lyrics`. O que acontece quando você executa o programa?

4.9 Fluxo de execução

A fim de garantir que uma função seja definida antes do primeiro uso, você tem que saber a ordem em que as declarações serão executadas, o que chamamos de **fluxo de execução**.

A execução sempre começará na primeira declaração do programa. Declarações são executadas uma por vez, em ordem do início ao fim.

Definições de funções não alteram o fluxo de executação de um programa, mas lembre-se que as declarações dentro de uma função não são executadas até que a função seja chamada.

Uma chamada de função é como um desvio no fluxo de execução. Ao invés de ir para a próxima declaração, o fluxo salta para o corpo da função, executa todas as declarações que a função possuir, e então volta para o lugar onde tinha parado.

Isto pode parecer simples o suficiente, até que você se lembra que uma função pode chamar outra função. Enquanto estiver no meio de uma função, o programa pode ter que executar declarações em outra função. Mas enquanto executa esta nova função, o programa pode ter que executar ainda outra função!

Felizmente, Python é bom o suficiente para manter o rastro de onde está, então cada vez que uma função termina, o programa volta para onde estava na função que a chamou. Quando alcançar o final do programa, ele termina.

Qual a moral deste conto sórdido? Quando você lê um programa, você nem sempre quer fazê-lo do início até o final. Algumas vezes faz mais sentido se você seguir o fluxo de executação.

4.10 Parâmetros e argumentos

Algumas das funções embutidas que vimos, requerem argumentos. Por exemplo, quando chamamos a função `math.sin` você passa um número como argumento. Algumas funções tem mais de um argumento: `math.pow`, precisa de dois, a base e o expoente.

Dentro da função, os argumentos são atribuídos a variáveis chamadas de **parâmetros**. Aqui está um exemplo de uma função que tem um argumento:

```
def print_twice(bruce):
    print bruce
    print bruce
```

Esta função atribui o argumento ao parâmetro chamado `bruce` Quando a função é chamada, ela exibe o valor do parâmetro (independente de qual seja) duas vezes.

Esta função funciona com qualquer valor que possa ser impresso.

```
>>> print_twice('Spam')
Spam
Spam
>>> print_twice(17)
17
17
>>> print_twice(math.pi)
3.14159265359
3.14159265359
```

As mesmas regras de composição que se aplicam a funções embutidas são aplicadas a funções definidas pelo usuário, então podemos utilizar qualquer tipo de expressão como argumento para `print_twice`:

```
>>> print_twice('Spam '*4)
Spam Spam Spam Spam
Spam Spam Spam Spam
>>> print_twice(math.cos(math.pi))
-1.0
-1.0
```

O argumento é avaliado antes da função ser chamada, então no exemplo a expressão `'Spam '*4` e `math.cos(math.pi)` são avaliadas somente uma vez.

Você também pode usar variáveis como argumento:

```
>>> michael = 'Eric, the half a bee.'
>>> print_twice(michael)
Eric, the half a bee.
Eric, the half a bee.
```

O nome da variável que passamos como argumento (`michael`) não tem relação com o nome do parâmetro (`bruce`). Não importa qual valor foi chamado (na chamada); aqui em `print_twice`, chamamos todos de `bruce`

4.11 Funções férteis e funções vazias

Algumas funções que utilizamos, como as funções `math`, apresentam resultados; pela falta de um nome melhor, vou chamá-las de **funções férteis**. Outras funções como `print_twice`, realizam uma ação mas não retornam um valor. Elas são chamadas **funções vazias**.

Quando você chama uma função fértil, você quase sempre quer fazer alguma coisa com o resultado; por exemplo, você pode querer atribuir a uma variável ou utilizar como parte de uma expressão:

```
x = math.cos(radians)
golden = (math.sqrt(5) + 1) / 2
```

Quando você chama uma função no modo interativo, o Python apresenta o resultado:

```
>>> math.sqrt(5)
2.2360679774997898
```

Mas em um script, se você chamar uma função fértil e não armazenar o resultado da função em uma variável, o valor retornado desaparecerá em meio a névoa!

```
math.sqrt(5)
```

Este script calcula a raiz quadrada de 5, mas desde que não se armazene o resultado em uma variável ou mostre o resultado, isto não é muito útil.

Funções vazias podem mostrar alguma coisa na tela ou ter algum outro efeito, mas elas não tem valor de retorno. Se você tentar atribuir o retorno a uma variável, você vai receber um valor especial chamado `None`.

```
>>> result = print_twice('Bing')
Bing
Bing
>>> print result
None
```

O valor `None` não é o mesmo de uma string `'None'`. É um valor especial que tem seu próprio tipo:

```
>>> print type(None)
<type 'NoneType'>
```

Para retornar um resultado de uma função, utilizamos a declaração `return` na nossa função. Por exemplo, podemos criar uma função bem simples chamada, `addtwo` que soma dois números e retorna o resultado.

```
def addtwo(a, b):
    added = a + b
    return added

x = addtwo(3, 5)
print x
```

Quando este script executa, a declaração `print` irá exibir o valor "8" porque a função `addtwo` foi chamada com 3 e 5 como argumento. Dentro da função, o parâmetro `a` e `b` são 3 e 5, respectivamente. A função calcula a soma dos dois números e coloca isto em uma variável local da função chamada `added`. Depois é utilizada pela declaração `return` para enviar o resultado calculado de volta para o código como resultado da função, ao qual foi atribuido a variável `x` e exibido na tela.

4.12 Por que funções?

Pode não ficar claro por que vale a pena o trabalho de dividir um programa em funções. Existem diversas razões:

- Criar uma nova função dá a você a oportunidade de nomear um grupo de declarações, o que tornará seu programa mais fácil de ler, entender e depurar
- Funções podem tornar seu programa menor, eliminando código repetido. Posteriormente, se você fizer uma alteração, você só precisa fazer em um lugar.
- Dividir um programa grande em funções, permite que você depure uma parte por vez e depois junte esta parte ao programa inteiro.
- Funções bem definidas serão úteis para muitos programas. Uma vez que você tenha escrito e depurado uma destas funções, você pode reutilizá-la.

Ao longo do resto do livro, normalmente utilizaremos uma função para explicar um conceito. Parte das habilidades para criar e utilizar uma função é de ter uma função que captura de forma apropriada uma ideia como "encontrar o menor valor em uma lista". Depois mostraremos códigos que encontram os menores valores em uma lista e apresentaremos a uma função chamada `min` que pega uma lista como argumento e retorna o menor valor desta lista.

4.13 Depuração

Se você estiver utilizando um editor de texto para escrever seus scripts, você pode ter problemas com espaços e tabs. A melhor forma de evitar estes problemas é utilizar sempre espaços (não tabs). A maioria dos editores de texto que conhecem Python, fazem isto por padrão, mas alguns não.

Tabs e espaços são usualmente invisíveis, o que torna difícil de depurar, então tente encontrar um editor de texto que gerencie a indentação pra você.

E também, não se esqueça de salvar seus programas antes de executá-los. Alguns ambientes de desenvolvimento (IDE) fazem isto automaticamente, mas alguns não. Nestes casos, o programa que você estará vendo no editor de texto não é o mesmo que você estará executando.

Depuração pode consumir uma grande quantidade de tempo, se você estiver executando o mesmo programa errado diversas vezes!

Tenha certeza que o código que você está olhando é o mesmo que você está executando. Se você não tiver certeza, coloque algo como um `print 'hello'` no começo do programa e execute novamente. Se você não ver o `hello`, você não está executando o mesmo programa!

4.14 Glossário

algoritmo: Um processo genérico para solução de problemas

argumento: Um valor dado a uma função quando a função é chamada. Este valor é atribuido a um parâmetro correspondente na função.

body (corpo): Sequência de declarações dentro de uma função.

composição: Utilizando uma expressão como parte de uma expressão maior ou uma declaração como parte de uma declaração maior.

deterministico: Refere-se a um programa que faz as mesmas coisas cada vez que é executado, dado os mesmos valores de entrada.

notação de ponto: Sintaxe para chamada de uma função em outro módulo especificado pelo nome do módulo seguido de um ponto (.) e o nome da função.

fluxo de execução: A ordem em que cada declaração será executada durante a execução do programa.

função fértil: Uma função que retorna um valor.

função: Nome para uma sequência de declarações que executam alguma operação. Funções podem ou não receber argumentos e podem ou não produzir um resultado.

chamada de função: Uma declaração que executa uma função. Consiste do nome de uma função seguida por uma lista de argumento.

definição de função: Uma declaração que cria uma nova função, especificando seu nome, parâmetros e as declarações que serão executadas.

função objeto: Um valor criado pela definição de uma função. O nome da função é uma variável que se refere a função objeto.

header (cabeça): A primeira linha de uma função.

declaração *import*: Uma declaração que lê um arquivo de módulo e cria um módulo objeto.

objeto módulo: Um valor criado pela chamada de uma declaração `import` que provê acesso aos dados e códigos definidos em um módulo.

parâmetro: Um nome utilizado dentro de uma função para referenciar ao valor passado por um argumento.

pseudoaleatório: Refere-se a uma sequência de número que parecem ser aleatórios, mas são gerados por um programa determinístico.

valor de retorno: O resultado de uma função. Se uma função for utilizada como uma expressão, o valor de retorno é o valor da expressão.

função vazia: Uma função que não possui um valor de retorno.

4.15 Exercícios

Exercício 4.4 Qual o propósito da palavra-chave ”def”em Python?

a) É uma gíria que significa ”Este código é muito maneiro”
b) Indica o início de uma função
c) Indica que a seção de código a seguir indentada deve ser guardada para depois
d) b e c são ambas verdadeiras
e) Nenhuma das questões acima.

Exercício 4.5 Qual será o resultado do programa abaixo?

```
def fred():
   print "Zap"

def jane():
   print "ABC"

jane()
fred()
jane()
```

a) Zap ABC jane fred jane
b) Zap ABC Zap
c) ABC Zap jane
d) ABC Zap ABC
e) Zap Zap Zap

Exercício 4.6 Reescreva o cálculo de pagamento com a metade do tempo por hora extra e crie uma função chamada `computepay` que recebe dois parâmetros (`hours` e `rate`).

```
Enter Hours: 45
Enter Rate: 10
Pay: 475.0
```

Exercício 4.7 Reescreva o programa de escala dos capítulos anteriores utilizando uma função chamada `computegrade` que recebe os pontos como parâmetros e retorna a escala como uma cadeia de caracteres (*string*).

```
Score   Grade
> 0.9     A
> 0.8     B
> 0.7     C
> 0.6     D
<= 0.6    F

Execução do Programa:

Digite o score: 0.95
A

Digite o score: perfect
Score invalido

Digite o score: 10.0
Score invalido

Digite o score: 0.75
C

Digite o score: 0.5
F
```

Execute o programa repetidas vezes para testar os diferentes valores de entrada.

Capítulo 5

Iteração

5.1 Atualizando variáveis

Um padrão comum nas instruções de atribuição é uma instrução de atribuição que atualiza uma variável – onde o novo valor da variável depende da antiga.

```
x = x+1
```

Isto significa "pega o valor atual de `x`, adicione 1, e depois atualize `x` com o novo valor."

Se você tentar atualizar uma variável que não existe, você receberá um erro, porque Python avalia o lado direito antes de atribuir um valor a `x`:

```
>>> x = x+1
NameError: name 'x' is not defined
```

Antes de você atualizar uma variável, é necessário **inicializá-la**, usualmente com uma simples atribuição:

```
>>> x = 0
>>> x = x+1
```

Atualizando uma variável, adicionando 1, é o que chamamos **incremento**; subtraindo 1 é o que chamamos de **decremento**.

5.2 A instrução `while`

Computadores são normalmente utilizados para automatizar tarefas repetitivas. A repetição de tarefas idênticas ou similares sem produzir erros é algo que computadores fazem bem e as pessoas não muito. Pelo fato de iterações serem tão comuns, Python disponibiliza muitas funcionalidades para tornar isto fácil.

Uma das formas de iterações em Python é a instrução `while`. Aqui está um programa simples que realiza uma contagem regressiva a partir de cinco e depois diz "Blastoff!".

```
n = 5
while n > 0:
    print n
    n = n-1
print 'Blastoff!'
```

Você quase pode ler a instrução `while` como se ela fosse escrita em Português. Ou seja, "Enquanto `n` for maior que 0, mostre o valor de `n` e então subtraia o valor de `n` em 1. Quando chegar ao 0, saia da declaração do `while` e mostre a palavra `Blastoff!`".

Formalmente, este é o fluxo de execução de uma declaração `while`:

1. Avalia a condição, produzindo `True` ou `False`.

2. Se a condição for falsa, sai da instrução `while` e continua a execução para a próxima declaração.

3. Se a condição for verdadeira, executa o corpo do `while` e depois volta para o passo 1.

Este tipo de fluxo é chamado de **laço** ou (*loop*) devido ao terceiro passo que retorna para o início da instrução. Chamamos cada vez que executamos o corpo do laço da **iteração**. Para o laço anterior, podemos dizer que, "tem cinco iterações", que significa que o corpo do laço será executado cinco vezes.

O corpo do laço deve mudar o valor de uma ou mais variáveis para que a condição eventualmente se torne false e o laço termine. Podemos chamar a variável que muda a cada vez que o laço executa e controla quando ele irá terminar de **variável de iteração**. Se não houver variável de iteração, o laço irá se repetir para sempre, resultando em um **laço infinito**.

5.3 Laços infinitos

Um recurso interminável de diversão para programadores é a observação do ato de se ensaboar, "ensaboe, enxague e repita", é um laço infinito porque não há variável de iteração dizendo quantas vezes o laço deve ser executado.

No caso de `contagem regressiva`, nós provamos que o laço terminou porque sabemos que o valor de `n` é finito, e podemos ver que o valor de `n` diminui cada vez que passa pelo laço, então eventualmente nós teremos 0. Em outros casos, o laço é obviamente infinito porque não tem variável de iteração.

5.4 “Laços infinitos” e break

Algumas vezes você não sabe se é hora de acabar o laço até que você percorra metade do corpo. Neste caso você pode escrever um laço infinito de propósito e então usar a declaração break para sair do laço.

Este laço é obviamente um **laço infinito** porque a expressão lógica do while é a constante lógica True

```
n = 10
while True:
    print n,
    n = n - 1
print 'Done!'
```

Se você cometer o erro e executar este código, você aprenderá rapidamente como parar um processo Python no seu sistema ou onde está o botão de desligar do seu computador. Este programa executará eternamente ou até que sua bateria acabe por que a expressão lógica no início do laço será sempre verdadeiro em virtude do fato que a expressão é o valor constante True.

Enquanto este laço é um laço infinito disfuncional, nós continuamos utilizando este padrão para construir laços úteis desde que adicionemos código de forma cuidadosa no corpo do laço para explicitamente sair do laço utilizando break quando alcançarmos a condição de saída.

Por exemplo, suponha que você queira obter a entrar do usuário, até que ele digite done. Podemos escrever:

```
while True:
    line = raw_input('> ')
    if line == 'done':
        break
    print line
print 'Done!'
```

A condição do laço é True, ou seja, é sempre verdade, então o laço executará de forma repetida até que chegue a declaração do break.

A cada vez, pergunta-se ao usuário com um sinal de menor. Se o usuário digitar done, a declaração break sai do laço. Caso contrário, o programa irá imprimir qualquer coisa que o usuário digitar e retornar para o início do laço. Veja um exemplo:

```
> hello there
hello there
> finished
finished
> done
Done!
```

Esta forma de escrever um laço while é muito comum, porque você pode verificar a condição em qualquer lugar do laço (não somente no início) e pode definir

explicitamente a condição de parar (“pare quando isto acontecer”) contrário de negativamente (“continue até que isto aconteça.”).

5.5 Terminando as iterações com `continue`

Algumas vezes você está em uma iteração de um laço e quer acabar a iteração atual e pular para a próxima iteração. Neste caso você pode utilizar a declaração `continue` para passar para a próxima iteração sem terminar o corpo do laço da iteração atual.

Aqui temos um exemplo de um laço que copia sua entrada até que o usuário digite “done”, mas trata a linha que inicia com um caractere cerquilha como linha para não ser impressa (como um comentário em Python).

```
while True:
    line = raw_input('> ')
    if line[0] == '#' :
        continue
    if line == 'done':
        break
    print line
print 'Done!'
```

Aqui temos um exemplo deste novo programa com o uso do `continue`.

```
> hello there
hello there
> # don't print this
> print this!
print this!
> done
Done!
```

Todas as linhas serão impressas, exceto aquela que inicia com o sinal de cerquilha porque quando o `continue` é executado, ele termina a iteração atual e pula de volta para a declaração `while` para começar uma nova iteração, mas passando a declaração `print`.

5.6 Usando `for` para laços

Algumas vezes queremos que um laço passe por um **conjunto** de coisas como uma lista de palavras, as linhas de um arquivo, ou uma lista de números. Quando temos uma lista de coisas para percorrer, construímos um laço *limitado* utilizando a declaração `for`. Nós chamamos uma declaração `while` como um laço `ilimitado` por que o laço executa até que alguma condição se torne `False`, enquanto o laço `for` é executado em um conjunto de itens conhecidos, então ele executa quantas iterações forem a quantidade de itens do conjunto.

A sintaxe do laço `for` é similar ao do `while` em que há uma declaração `for` e um corpo para o laço percorrer:

```
friends = ['Joseph', 'Glenn', 'Sally']
for friend in friends:
    print 'Happy New Year:', friend
print 'Done!'
```

Em Python, a variável `friends` é uma lista[1] de três strings e o laço `for` passa através da lista e executa o corpo uma vez para cada uma das três strings na lista, resultando na saída:

```
Happy New Year: Joseph
Happy New Year: Glenn
Happy New Year: Sally
Done!
```

Traduzindo este laço `for` para o Português, não é tão direto como o laço `while`, mas se você pensar em amigos como um **conjunto**, fica parecido com isto: "Execute a declaração no corpo do laço for uma vez para cada amigo *nos* nomes dos amigos".

Olhando ao laço `for`, **for** e **in** são palavras reservadas do Python, e `friend` e `friends` são variáveis.

for `friend` **in** `friends`**:**
    print `'Happy New Year', friend`

Em particular, `friend` é a **variável de iteração** do laço `for`. A variável `friend` muda para cada iteração do laço e controla quando o laço `for` completa. A **variável de iteração** passa sucessivamente através das três strings armazenadas na variável `friends`.

5.7 Padrões de Laços

Normalmente, utilizamos os laços `for` ou `while` para percorrer uma lista de itens ou o conteúdo de um arquivo procurando por alguma coisa como o maior ou o menor valor do dado que estamos percorrendo.

Estes laços são normalmente construídos da seguinte forma:

- Inicializando uma ou mais variáveis antes de iniciar o laço
- Realizando alguma verificação em cada item no corpo do laço, possivelmente mudando as variáveis no corpo do laço
- Olhando o resultado das variáveis quando o laço finaliza

Utilizamos uma lista de números para demonstrar os conceitos e os padrões para construção de laços.

[1] Nós analisaremos as listas em mais detalhes em um capítulo mais adiante.

5.7.1 Contando e somando laços

Por exemplo, para contar o números de itens em uma lista, podemos escrever o seguinte laço `for`:

```
count = 0
for itervar in [3, 41, 12, 9, 74, 15]:
    count = count + 1
print 'Count: ', count
```

Nós definimos a variável `count` em zero antes do laço iniciar, então escrevemos um laço `for` para percorrer uma lista de números. Nossa variável de iteração é chamada de `itervar` e enquanto não usamos a variável `itervar` no laço, ele controla o laço que o será executado somente uma vez para cada valor na lista.

No corpo do laço, adicionamos 1 ao valor atual de `count` para cada valor da lista. Enquanto o laço é executado, o valor da variável `count` é o número de valores que nós vimos "até agora".

Uma vez que o laço termina, o valor de `count` é o total de itens. O total de itens "cai no seu colo" no final do laço. Construímos o laço para que tenhamos o que queremos quando o laço terminar.

Outro laço similar que calcula o total de um conjunto de números pode ser visto a seguir:

```
total = 0
for itervar in [3, 41, 12, 9, 74, 15]:
    total = total + itervar
print 'Total: ', total
```

Neste laço, nós *fazemos* uso da **variável de iteração**. Ao invés de simplesmente adicionar um a variável `count`, como vimos no laço anterior, nós adicionamos o número atual (3, 41, 12, etc.) ao total atual na iteração de cada vez que o laço é executado. Se você pensar sobre a variável `total`, ela contém o "o total dos valores até então". Então, antes do laço iniciar o `total` é zero porque nós não vimos nenhum valor, e durante o laço, o valor de `total` é o total atual, e no final do laço, `total` é a soma total de todos os valores na lista.

Enquanto o laço é executado, `total` acumula a soma dos elementos; uma variável utilizada desta maneira é chamada de **acumulador**.

Nem o laço contador, nem o laço somador são particularmente úteis na prática porque Python tem funções nativas `len()` e `sum()` que calcula o número e o total de itens em uma lista, respectivamente.

5.7.2 Laços de máximos e mínimos

Para encontrar o maior valor em uma lista ou sequência, construímos o seguinte laço:

```
largest = None
print 'Before:', largest
for itervar in [3, 41, 12, 9, 74, 15]:
    if largest is None or itervar > largest :
        largest = itervar
    print 'Loop:', itervar, largest
print 'Largest:', largest
```

Ao executar o programa, a saída é a seguinte:

```
Before: None
Loop: 3 3
Loop: 41 41
Loop: 12 41
Loop: 9 41
Loop: 74 74
Loop: 15 74
Largest: 74
```

A variável `largest` é vista como o "maior valor que temos". Antes do laço nós definimos `largest` com a constante `None`. `None` é um valor especial que podemos utilizar em uma variável para definir esta variável como "vazia".

Antes que o laço inicia, o maior valor que temos até então é `None`, uma vez que nós ainda não temos valor nenhum. Enquanto o laço está executando, se `largest` é `None` então, pegamos o primeiro valor que temos como o maior. Você pode ver na primeira iteração quando o valor de `itervar` é 3, uma vez que `largest` é `None`, nós imediatamente definimos a variável `largest` para 3.

Depois da primeira iteração, `largest` não é mais `None`, então a segunda parte composta da expressão lógica que verifica o gatilho `itervar > largest` somente quando o valor é maior que o "maior até agora". Quando temos um novo valor "ainda maior" nós pegamos este novo valor e definimos como `largest`. Você pode ver na saída do programa o progresso do `largest` de 3 para 41 para 74.

No final do laço, analisamos todos os valores e a variável `largest` agora contém o maior valor na lista.

Para calcular o menor número, o código é muito similar com pequenas diferenças:

```
smallest = None
print 'Before:', smallest
for itervar in [3, 41, 12, 9, 74, 15]:
    if smallest is None or itervar < smallest:
        smallest = itervar
    print 'Loop:', itervar, smallest
print 'Smallest:', smallest
```

Novamente, `smallest` é o "menor até agora" antes, durante e depois do laço ser executado. Quando o laço se completa, `smallest` contém o mínimo valor na lista.

De novo, assim como contagem e soma, as funções nativas `max()` e `min()` tornam estes laços desnecessários.

A seguir uma versão simples da função nativa `min()` do Python:

```
def min(values):
    smallest = None
    for value in values:
        if smallest is None or value < smallest:
            smallest = value
    return smallest
```

Nesta pequena versão da função, retiramos todos as declarações de `print` para que fosse equivalente a função `min` que é nativa no Python.

5.8 Depurando

Assim que você começar a escrever programas maiores, você se encontrará gastando mais tempo depurando. Mais códigos significam mais chances de se fazer mais erros e mais bugs para se esconder.

Uma forma de diminuir o tempo de depuração é "depuração por bisseção". Por exemplo, se você tiver 100 linhas em seu programa e você verificá-la uma por vez, isto levaria 100 passos.

Ao invés disto, tente quebrar o programa pela metade. Olhe para a metade do programa, ou próximo dele, por um valor intermediário que você possa verificar. Adicione a declaração de `print` (ou alguma coisa que tenha um efeito verificável) e execute o programa.

Se a verificação do ponto intermediário estiver incorreta, o problema pode estar na primeira metade do programa. Se estiver correto, o problema está na segunda parte.

Toda vez que você executar uma verificação como esta, você reduzirá o número de linha que você tem que procurar. Depois de seis passos (o que é muita menos que 100), você poderia diminuir para uma ou duas linha de código, pelo menos em teoria.

Na prática, nem sempre está claro qual é a "metade do programa" e nem sempre é possível verificar. Não faz sentido contar as linhas e achar exatamente o meio do programa. Ao contrário, pense sobre lugares no programa onde podem haver erros e lugares onde é fácil colocar uma verificação, (um `print`) Então escolha um lugar onde você acha que pode ocorrer erros e faça uma verificação antes e depois da análise.

5.9 Glossário

acumulador: Uma variável utilizada em um laço para adicionar e acumular o resultado.

contador: Uma variável utilizada em um laço para contar um número de vezes que uma coisa aconteça. Nós inicializamos o contador em zero e depois incrementamos o contador cada vez que quisermos "contar" alguma coisa.

decremento: Uma atualização que diminui o valor de uma variável.

inicializador: Uma atribuição que dá um valor inicial para a variável que será atualizada.

incremento: Uma atualização que aumenta o valor de uma variável (muitas vezes em um).

laço infinito: Um laço onde a condição terminal nunca é satisfeita ou para o qual não exista condição terminal.

iteração: Execução repetida de um conjunto de declarações utilizando uma função ou um laço que se executa.

5.10 Exercícios

Exercício 5.1 Escreva um programa que repetidamente leia um número até que o usuário digite "done". Uma vez que "done" é digitada, imprima o total, soma e a média dos números. Se o usuário digitar qualquer coisa diferente de um número, detecte o engano utilizando `try` e `except` e imprima uma mensagem de erro e passe para o próximo número.

```
Enter a number: 4
Enter a number: 5
Enter a number: bad data
Invalid input
Enter a number: 7
Enter a number: done
16 3 5.33333333333
```

Exercício 5.2 Escreva outro programa que solicita uma lista de números, como acima, e no final imprima o máximo e o mínimo dos números ao invés da média.

Capítulo 6

Strings

6.1 Uma string é uma sequência

Uma string é uma **sequência** de caracteres. Você pode acessar cada caractere, um por vez com o operador colchete, como pode ser visto no código a seguir:

```
>>> fruit = 'banana'
>>> letter = fruit[1]
```

A segunda declaração extrai o caractere no índice de posição 1 da variável `fruit` e atribui a variável `letter`.

A expressão entre o colchetes é chamada de **índice**. O índice indica qual caractere na sequência você quer (daí o nome).

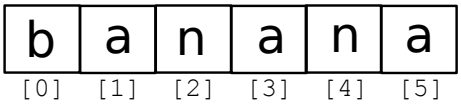
Mas você pode não ter o que espera:

```
>>> print letter
a
```

Para a maioria das pessoas, a primeira palavra de `'banana'` é `b`, não `a`. Mas em Python, o índice é alinhado com o começo da *string* e o alinhamento da primeira letra é a partir do zero.

```
>>> letter = fruit[0]
>>> print letter
b
```

Então `b` é a letra 0 (“posição zero”) de `'banana'`, `a` é letra 1 (“posição um”), e `n` é letra 2 (“posição 2”) e assim por diante até o fim da palavra.



Você pode utilizar qualquer expressão, variável e operador, como um índice, mas o valor do índice tem que ser um inteiro. Caso contrário você terá:

```
>>> letter = fruit[1.5]
TypeError: string indices must be integers
```

6.2 Obtendo o tamanho de uma *string* usando `len`

A função `len` nativa do Python, que retorna o número de caracteres de uma *string*:

```
>>> fruit = 'banana'
>>> len(fruit)
6
```

Para obter a última letra da *string*, você pode tentar fazer algo como isto:

```
>>> length = len(fruit)
>>> last = fruit[length]
IndexError: string index out of range
```

A razão para o `IndexError` é que não existe letra em banana no índice 6. Uma vez que começamos a contar a partir do zero, a seis letras são numeradas de 0 até 5. Para mostrar o último caractere, você tem que subtrair 1 do tamanho (`length`):

```
>>> last = fruit[length-1]
>>> print last
a
```

Como alternativa, é possível utilizar índices negativos, que contam a string de trás pra frente. A expressão `fruit[-1]` mostra a última letra, `fruit[-2]` mostra a segunda a partir do final, e assim por diante.

6.3 Percorrendo uma *string* com um *loop*

Processar uma string, um caractere por vez, envolve uma série de computação. Normalmente, eles começam no começo da palavra, selecionam um caractere por vez, fazem alguma coisa com ele, e continuam até o final. Este padrão de processamento é chamado de **percorrer**. Uma forma de percorrer uma *string*, por exemplo, é através de um loop com `while`:

```
index = 0
while index < len(fruit):
    letter = fruit[index]
    print letter
    index = index + 1
```

Este loop percorre a string e apresenta cada letra em uma linha própria. A condição do loop é `índice < len(fruit)`, assim quando o `índice` for igual ao tamanho da *string*, a condição se torna falsa, e o loop não é executado. O último caractere acessado é o caractere com o índice `len(fruit)-1`, que é o último caractere na *string*.

Exercício 6.1 Escreva um loop `while` que comece no último caractere da string e volte de trás pra frente até o primeiro caractere da string, imprimindo cada letra em uma linha separada.

Outra forma de percorrer uma string é com um loop `for`:

```
for char in fruit:
    print char
```

Cada vez que percorrer o loop, o caractere na string é atribuido a variável `char`. O loop continua até que não haja mais caractere na string.

6.4 Fatiando *strings*

Um segmento de uma string é chamado de **fatia**. Selecionar uma fatia é similar a selecionar um caractere:

```
>>> s = 'Monty Python'
>>> print s[0:5]
Monty
>>> print s[6:12]
Python
```

O operador `[n:m]` retorna a parte da string da posição "n" até a posição "m", incluindo o primeiro, mas excluindo o último.

Se você omitir o primeiro índice (antes dos dois pontos), a fatia inicia no começo da string. Se você omitir o segundo índice, a fatia irá até o fim da string:

```
>>> fruit = 'banana'
>>> fruit[:3]
'ban'
>>> fruit[3:]
'ana'
```

Se o primeiro índice for maior ou igual ao segundo, o resultado é uma **string vazia**, representado entre duas aspas.

```
>>> fruit = 'banana'
>>> fruit[3:3]
''
```

Uma string vazia não contém caracteres e tem tamanho 0 (zero), mas diferente disto, isto é igual a qualquer outra string.

Exercício 6.2 Dada uma string, `fruit`, o que significa a declaração `fruit[:]`?

6.5 Strings são imutáveis

É tentador utilizar o operador `[]` no lado esquerdo de uma atribuição, com a intenção de mudar um caractere em uma string. Por exemplo:

```
>>> greeting = 'Hello, world!'
>>> greeting[0] = 'J'
TypeError: object does not support item assignment
```

O "objeto" nesse caso é a string e o "item" é o caractere que você tentou atribuir. Agora, um **objeto** é a mesma coisa que um valor, mas vamos refinar esta definição posteriormente. Um **item** é um dos valores em uma sequência.

A razão para o erro é que strings são **imutáveis**, que significa que você não pode mudar uma string já existente. O melhor que você pode fazer é criar uma nova string que é uma variação da original:

```
>>> greeting = 'Hello, world!'
>>> new_greeting = 'J' + greeting[1:]
>>> print new_greeting
Jello, world!
```

Neste exemplo, concatenamos uma nova letra em uma fatia de `greeting`. Isto não tem efeito na string original.

6.6 Looping e contabilização

O programa a seguir conta o número de vezes que a letra `a` aparece em uma string:

```
word = 'banana'
count = 0
for letter in word:
    if letter == 'a':
        count = count + 1
print count
```

Este programa demonstra outro padrão de computação chamado **contador**. A variável `count` é iniciada em 0 e depois incrementada cada vez que uma letra `a` é encontrada. Quando o laço existe, `count` contém o resultado—o número total de `a`'s.

Exercício 6.3 Encapsule este código em uma função chamada `count`, e generalize para que aceite a string e a letra como argumento.

6.7 O operador `in`

A palavra `in` é um operador booleano que pega duas strings e retorna `True` se a primeira aparecer como substring na segunda:

```
>>> 'a' in 'banana'
True
>>> 'seed' in 'banana'
False
```

6.8 Comparação de string

O operador de comparação funciona com strings. Para verificar se duas strings são iguais:

```
if word == 'banana':
    print  'All right, bananas.'
```

Outras operações de comparações são úteis para colocar as palavras em ordem alfabética:

```
if word < 'banana':
    print 'Your word,' + word + ', comes before banana.'
elif word > 'banana':
    print 'Your word,' + word + ', comes after banana.'
else:
    print 'All right, bananas.'
```

Python não manipula letras em maiúscula ou minúscula da mesma forma que as pessoas fazem. Todas as palavras em maiúsculas vem antes das minúsculas, então:

```
%Your word, Pineapple, comes before banana.
Sua palavra, Abacaxi, vem antes de banana.
```

Uma maneira de tratar este problema é converter strings para um formato padrão, todas como minúsculas, antes de realizar a comparação. Mantenha isto em mente em caso de ter que se defender contra alguém armado com um abacaxi.

6.9 Método `string`

Strings são um exemplo de um **objeto** em Python. Um objeto contém ambos dado (a string atual) e os **métodos**, que são efetivamente funções construídas dentro do objeto e disponível para quaisquer instâncias do objeto.

Python tem uma função chamada `dir` que lista os métodos disponíveis de um objeto. A função `type` mostra o tipo de um objeto e a função `dir` os métodos disponíveis.

```
>>> stuff = 'Hello world'
>>> type(stuff)
<type 'str'>
>>> dir(stuff)
['capitalize', 'center', 'count', 'decode', 'encode',
'endswith', 'expandtabs', 'find', 'format', 'index',
'isalnum', 'isalpha', 'isdigit', 'islower', 'isspace',
'istitle', 'isupper', 'join', 'ljust', 'lower', 'lstrip',
'partition', 'replace', 'rfind', 'rindex', 'rjust',
'rpartition', 'rsplit', 'rstrip', 'split', 'splitlines',
'startswith', 'strip', 'swapcase', 'title', 'translate',
'upper', 'zfill']
>>> help(str.capitalize)
```

```
Help on method_descriptor:

capitalize(...)
    S.capitalize() -> string

    Return a copy of the string S with only its first character
    capitalized.
>>>
```

A função `dir` lista os métodos, e você pode utilizar `help` para obter ajuda na documentação de um método, uma melhor fonte de documentação para métodos de string pode ser vista através do endereço `https://docs.python.org/2/library/stdtypes.html#string-methods`.

Chamar um **método** é similar a chamar uma função—recebe argumentos e retorna um valor—mas a sintaxe é diferente. Nós chamamos um método anexando o nome do método a variável utilizando um ponto como delimitador.

Por exemplo, o método `upper` transforma uma string, retornando uma nova string com todas as letras em maiúsculo:

Ao invés de usar a sintaxe de uma função `upper(word)`, usa-se a sintaxe de método `word.upper()`.

```
>>> word = 'banana'
>>> new_word = word.upper()
>>> print new_word
BANANA
```

Esta forma de notação de ponto especifica o nome do método, `upper`, e o nome da string para aplicar o método, `word`. O parêntese vazio indica que este método não recebe argumento.

Uma chamado de método é dito como uma **invocação**; neste caso, podemos dizer que estamos invocando `upper` na palavra `word`.

Por exemplo, existe um método de string chamado `find` que procura pela posição de uma string em outra:

```
>>> word = 'banana'
>>> index = word.find('a')
>>> print index
1
```

Neste exemplo, nós invocamos `find` na palavra `word` e passamos a letra que estamos procurando como um parâmetro.

O método `find` consegue encontrar substrings, assim como caracteres:

```
>>> word.find('na')
2
```

Pode receber como segundo argumento, o índice que indica onde deve começar:

```
>>> word.find('na', 3)
4
```

Uma tarefa comum é remover espaços em branco (espaços, tabs ou novas linhas) do início e final de uma string é usado o método `strip`:

```
>>> line = '  Here we go  '
>>> line.strip()
'Here we go'
```

Alguns métodos como o **startswith** retorna valores booleanos.

```
>>> line = 'Please have a nice day'
>>> line.startswith('Please')
True
>>> line.startswith('p')
False
```

Você perceberá que `startswith` precisa ser case sensitive para funcionar, então algumas vezes nós pegamos uma linha e convertemos para minúscula antes de fazer qualquer verificação, utilizando o método `lower`.

```
>>> line = 'Please have a nice day'
>>> line.startswith('p')
False
>>> line.lower()
'please have a nice day'
>>> line.lower().startswith('p')
True
```

No último exemplo, o método `lower` é chamado e depois utilizamos `startswith` para verificar se a string resultante em minúsculo começa com a letra "p". Contanto que nos preocupemos com a ordem, podemos realizar múltiplas chamadas de métodos em uma única expressão.

Exercício 6.4 Existe um método de strings chamado `count` que é similar a função do exercício anterior. Leia a documentação deste método no endereço: `https://docs.python.org/2/library/stdtypes.html#string-methods` e escreva uma invocação que conte o número de vezes que a letra "a" ocorre em `'banana'`.

6.10 Analisando strings

Normalmente queremos olhar uma string e procurar uma substring. Por exemplo se forem apresentadas uma série de linhas formatadas como a seguir:

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
```

e quisermos tirar somente a segunda metade do endereço (i.e., `uct.ac.za`) de cada linha, nós podemos fazer isto utilizando o método `find`, fatiando a string.

Primeiro, nós encontraremos a posição do arroba ("@") na string. Depois acharemos a posição do primeiro espaço, *depois* do arroba. E então usaremos o fatiamento da string para extrair a porção da string que estamos procurando.

```
>>> data = 'From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008'
>>> atpos = data.find('@')
>>> print atpos
21
>>> sppos = data.find(' ',atpos)
>>> print sppos
31
>>> host = data[atpos+1:sppos]
>>> print host
uct.ac.za
>>>
```

Utilizamos uma versão do método `find`, que nos permite especificar uma posição na string, onde queremos começar a procura. Quando fatiamos, extraímos os caracteres de "um além do arroba até *mas não incluindo* o caractere espaço".

A documentação para o método `find` está disponível no endereço `https://docs.python.org/2/library/stdtypes.html#string-methods`.

6.11 Operador format

O operador **format**, `%` nos permite construir strings, substituindo parte da string com dados armazenados em variáveis. Quando aplicados a inteiros, o `%` é o operador módulo. Mas quando o primeiro operando é uma string, o `%` é o operador `format`.

O primeiro operando é o **format** de string, que contém uma ou mais **sequências** que especifica como o segundo operador é formatado. O resultado é uma string.

Por exemplo, a sequência de formatação `'%d'` significa que o segundo operando deve ser formatado como um inteiro (`d` significa "decimal"):

```
>>> camels = 42
>>> '%d' % camels
'42'
```

O resultado é a string `'42'`, que não pode ser confundido com o valor inteiro `42`.

Um sequência de `format` pode aparecer em qualquer lugar em uma string, então você pode embutir um valor em uma sequência:

```
>>> camels = 42
>>> 'I have spotted %d camels.' % camels
'I have spotted 42 camels.'
```

Se houver mais de uma sequência de `format` na string, o segundo argumento tem que ser uma tupla[1]. Cada sequência de `format` é combinada com um elemento da tupla, em ordem.

O seguinte exemplo utiliza `'%d'` para formatar um inteiro, o `'%g'` para formatar um ponto-flutuante (não pergunte o por quê), e `'%s'` para formatar string:

```
>>> 'In %d years I have spotted %g %s.' % (3, 0.1, 'camels')
'In 3 years I have spotted 0.1 camels.'
```

O número de elementos na tupla deve combinar com o número de sequência para formatar em uma string. Os tipos de elementos também devem combinar com a sequência a ser formatada:

```
>>> '%d %d %d' % (1, 2)
TypeError: not enough arguments for format string
>>> '%d' % 'dollars'
TypeError: illegal argument type for built-in operation
```

No primeiro exemplo, não existem elementos suficientes; no segundo o elemento possui o tipo errado.

O operador `format` é muito poderoso, mas pode ser difícil de ser utilizado. Você pode ler mais sobre ele no endereço `https://docs.python.org/2/library/stdtypes.html#string-formatting`.

Você pode especificar o número de dígitos como parte do formato de uma sequência. Por exemplo, a sequência `'%8.2f'` formata um número em ponto flutuante para ter 8 caracteres de comprimento, com 2 dígitos depois do ponto decimal:

```
>>> '%8.2f' % 3.14159
'    3.14'
```

O resultado ocupa oito casas com dois dígitos depois do ponto decimal;

6.12 Depurando

Uma habilidade que você deve cultivar como programador é se perguntar sempre "O que poderia dar errado?" ou alternativamente, "Que coisa louca nosso usuário pode fazer para quebrar nosso programa (aparentemente) perfeito?".

Por exemplo, olhe para o programa que utilizamos para demonstrar o laço `while` no capítulo de iterações:

```
while True:
    line = raw_input('> ')
    if line[0] == '#' :
```

[1] Um tupla é uma sequência de valores, separados por vírgula dentro de um par de colchetes. Vamos abordar tuplas no Capítulo 10

```
        continue
    if line == 'done':
        break
    print line

print 'Done!'
```

Olhe o que acontece quando o usuário entrar com uma linha em branco no `input`:

```
> hello there
hello there
> # don't print this
> print this!
print this!
>
Traceback (most recent call last):
  File "copytildone.py", line 3, in <module>
    if line[0] == '#' :
```

O código funciona, até que se use uma linha vazia. Então existe um caractere não zero, e assim recebemos um `traceback`. Existem duas soluções para isto para tornar a linha três "segura" mesmo se a linha estiver vazia.

Uma possibilidade é utilizando o método `startswith` que retorna `False` se a string estiver vazia.

```
    if line.startswith('#') :
```

Outra forma é escrever de forma segura uma condição de `if` utilizando o padrão de **guarda** e garantir que a segunda expressão lógica seja avaliada somente onde existe pelo menos um caractere na string:

```
    if len(line) > 0 and line[0] == '#' :
```

6.13 Glossário

contador: Uma variável utilizada para contar alguma coisa, normalmente inicializada em zero e depois incrementada.

string vazia: Uma string sem caracteres e tamanho 0, representado por duas aspas.

operador format: Um operador, `%`, que pega uma string formatada e uma tupla gerando um string que inclui elementos da tupla formatada especificada pela string formatada.

sequência formatada: Uma sequência de caracteres em uma string formatada, como `%d`. que especifica como um valor deve ser formatado.

string formatada: Uma string, utilizada com o operador `format`, que contém uma sequência formatada.

flag: Uma variável booleana utilizada para indicar se uma condição é verdadeira.

invocação: Uma condição que chama um método.

imutável: Propriedades de uma sequência dos itens que não podem ser atribuídos.

índice: Um valor inteiro usado para selecionar um item em uma sequência, como um caractere em uma string.

item: Um dos valores em uma sequência.

método: Uma função que é associada com um objeto e acessado utilizando a notação de ponto.

objeto: Algum valor ao qual uma variável se refere. Desta forma você pode utilizar "objeto" e "valor" de forma intercambiável.

procura: Um padrão que percorre transversalmente e para quando encontra o que está procurando.

sequência: Um conjunto ordenado; que é, um conjunto de valores onde cada valor é identificado por um índice inteiro.

fatia: Uma parte da string especificada por uma

percorrer: Percorrer através de itens em uma sequência, executando uma operação similar em cada um dos itens.

6.14 Exercícios

Exercício 6.5 Use o código Python a seguir para armazenar a string:'

```
str = 'X-DSPAM-Confidence: 0.8475'
```

Use o `find` e o fatiamento de strings para extrair a parte da string depois da vírgula e então use a função `float` para converter a string extraída em um número de ponto flutuante.

Exercício 6.6 Leia a documentação dos métodos de `string` no endereço `https://docs.python.org/2/library/stdtypes.html#string-methods`. Você pode querer experimentar alguns destes métodos para ter certeza que entendeu como funcionam. Por exemplo, `strip` e `replace` são particularmente úteis.

A documentação utiliza uma sintaxe que pode parecer confusa. Por exemplo, no `find(sub[, start[, end]])`, os colchetes indicam que o argumento é opcional. Desta forma, `sub` é obrigatório, mas `start` é opcional, e se você incluir o `start`, então o `end` é opcional.

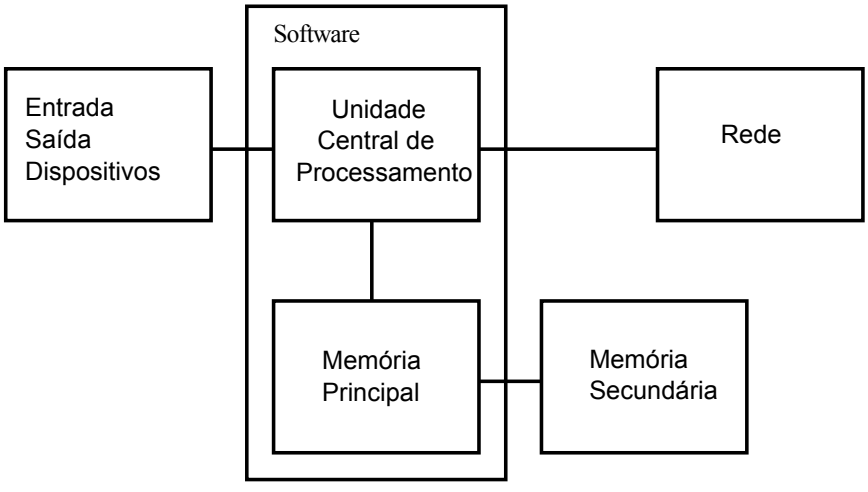
Capítulo 7

Arquivos

7.1 Persistência

Até agora, aprendemos como escrever programas e comunicar nossas intenções para a **Unidade de Processamento Central** usando execução condicional, funções e iterações. Aprendemos também como criar e usar estruturas de dados na **Memória Principal**. A CPU e a memória é onde nosso software executa. É o lugar onde todo "o pensamento" acontece.

Mas se você se recordar das nossas discussões sobre arquitetura de hardware, uma vez que a energia for desligada, tudo que estiver armazenado na CPU ou na memória principal será apagado. Até agora, nossos programas tem sido breves exercícios para aprender Python.



Neste capítulo, começaremos a trabalhar com **Memória Secundária** (ou arquivos). A memória secundária não é apagada quando a energia é desligada. Ou no caso de um pen drive USB, o dado que nós escrevemos a partir de nossos programas, pode ser removido e transportado para outro sistema.

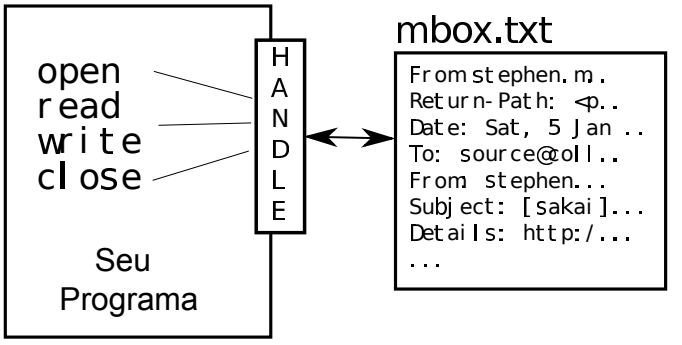
Nós focaremos primeiramente na leitura e escrita de arquivos texto tais como aqueles que criamos em um editor de texto. Depois iremos trabalhar com arquivos de banco de dados que são arquivos binários, especificamente desenhados para serem lidos e escritos através do nosso software de banco de dados.

7.2 Lendo arquivos

Quando queremos ler ou gravar um arquivo (nosso disco rígido, por exemplo), devemos sempre abrir o arquivo primeiro através do comando **open**. Abrir um arquivo é uma comunicação com o seu sistema operacional, que sabe onde o dado para cada arquivo é armazenado. Quando você abre um arquivo, você está pedindo ao sistema operacional para encontrar o arquivo pelo nome e certificar-se de que ele existe. Neste exemplo, abrimos o arquivo `mbox.txt`, o qual deve ser armazenado no mesmo diretório onde o seu programa Python está executando. Você pode fazer o download deste arquivo a partir de: `www.py4inf.com/code/mbox.txt`

```
>>> fhand = open('mbox.txt')
>>> print fhand
<open file 'mbox.txt', mode 'r' at 0x1005088b0>
```

Se o comando `open` rodar com sucesso, o sistema operacional nos retorna um **manipulador de arquivo**. Este manipulador não contém os dados do arquivo, mas apenas o "ponteiro" que nós podemos usar para ler um dado. Você recebe um ponteiro se o arquivo requisitado existir e se você tiver permissão para lê-lo.



Se o arquivo não existir, `open` ocorrerá um erro com a pilha de execução (traceback) e você não conseguirá obter um ponteiro (handle) para acessar o conteúdo do arquivo:

```
>>> fhand = open('stuff.txt')
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
IOError: [Errno 2] No such file or directory: 'stuff.txt'
```

Mais tarde, vamos aprender a utilizar `try` e `except` para lidar com a situação onde tentamos abrir um arquivo que não existe.

7.3 Arquivos texto e linhas

Podemos imaginar um arquivo texto como um sequência de linhas, assim como uma string em Python é uma sequência de caracteres. Por exemplo, esta é um exemplo de um arquivo texto com registros de atividade de e-mail de várias pessoas em um time de desenvolvimento em um projeto open source:

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
Return-Path: <postmaster@collab.sakaiproject.org>
Date: Sat, 5 Jan 2008 09:12:18 -0500
To: source@collab.sakaiproject.org
From: stephen.marquard@uct.ac.za
Subject: [sakai] svn commit: r39772 - content/branches/
Details: http://source.sakaiproject.org/viewsvn/?view=rev&rev=39772
...
```

O arquivo completo de iterações por e-mail está disponível em: `www.py4inf.com/code/mbox.txt` e uma versão reduzida do arquivo está disponível em: `www.py4inf.com/code/mbox-short.txt`. Estes arquivos estão em um formato padrão de um arquivo contendo múltiplas mensagens de e-mail. A expressão "From " separa as mensagens e as linhas que começam com "From:" são parte da mensagem. Para maiores informações sobre o formato mbox, veja: `en.wikipedia.org/wiki/Mbox`.

Para separar o arquivo em linhas, existe um caractere especial que representa o "fim da linha" chamado de **newline** caractere.

Em Python, representamos o caractere **newline** como a string `\n`, uma constante string. Mesmo que essa expressão pareça ser dois caracteres, ela é na verdade apenas um caractere simples. Quando imprimimos o valor da variável "stuff" no interpretador, ele nos mostra o `\n` na string, mas quando usamos `print` para exibir, nós vemos uma string quebrada em duas linhas pelo caractere newline.

```
>>> stuff = 'Hello\nWorld!'
>>> stuff
'Hello\nWorld!'
>>> print stuff
Hello
World!
>>> stuff = 'X\nY'
>>> print stuff
X
Y
>>> len(stuff)
3
```

Você também pode ver que o tamanho da string `'X\nY'` é *três [three]* caracteres porque o caractere newline é um único caractere simples.

Então, quando olhamos as linhas em um arquivo, nós precisamos *imaginar* que ele é uma espécie de caractere invisível que faz com que o fim de cada linha seja de fato, o fim da linha.

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008\n

Return-Path: <postmaster@collab.sakaiproject.org>\n

Date: Sat, 5 Jan 2008 09:12:18 -0500\n

To: source@collab.sakaiproject.org\n

From: stephen.marquard@uct.ac.za\n

Subject: [sakai] svn commit: r39772 - content/branches/\n

Details: http://source.sakaiproject.org/viewsvn/?view=rev&rev=39772\n

...
```

Observe que o caractere newline separa os caracteres no arquivo em linhas.

7.4 Lendo arquivos

O **ponteiro para o arquivo** não contém o dado do arquivo, é muito fácil construir um laço `for` para ler o arquivo inteiro e contar quantas linhas existem.

```
fhand = open('mbox.txt')
count = 0
for line in fhand:
    count = count + 1
print 'Line Count:', count

python open.py
Line Count: 132045
```

Nós podemos utilizar o ponteiro do arquivo como uma sequência no nosso loop `for`. Nosso loop `for` conta o número de linhas no arquivo e então imprime. Uma tradução grotesca do loop `for` para o português seria, "para cada linha do arquivo representada pelo ponteiro do arquivo, adicione um à variável `count`."

A razão pela qual a função `open` não lê o arquivo inteiro é que o arquivo pode ser muito grande com vários gigabytes de dados. A instrução `open` recebe a mesma quantidade de tempo sem levar em consideração o tamanho do arquivo.

Quando um arquivo é lido usando um laço `for` desta maneira, o Python divide o dado do arquivo em linhas separadas pelo caractere newline. O Python lê cada linha até encontrar o newline e então inclui o newline como o último caractere da variável `line` para cada iteração do laço `for`.

Pelo fato de o laço `for` ler o dado uma linha de cada vez, ele consegue eficientemente ler e contar as linhas em um arquivos grandes sem estourar a memória do computador para armazenar os dados. O programa acima pode contar as linhas

em qualquer tamanho de arquivo usando pouca quantidade de memória uma vez que cada linha é lida, contada e então descartada.

Se você souber que o arquivo é relativamente pequeno comparado ao tamanho total da memória principal, você pode ler o arquivo inteiro para uma única string usando o método `read` no ponteiro do arquivo `handle`.

```
>>> fhand = open('mbox-short.txt')
>>> inp = fhand.read()
>>> print len(inp)
94626
>>> print inp[:20]
From stephen.marquar
```

Neste exemplo, o conteúdo total (todos os 94.626 caracteres) do arquivo `mbox-short.txt` são lidos diretamente para a variável `inp`. Nós usamos o método de fatiar a string `slice` para imprimir os primeiros 20 caracteres dos dados armazenados na string `inp`.

Quando o arquivo é lido deste modo, todos os caracteres incluindo todas as linhas e caracteres newline são uma única e grande string dentro da variável **inp**. Lembre que este modo de utilizar a função `open` deve somente ser usado se o tamanho do arquivo lido couber perfeitamente na memória principal do seu computador.

Se o arquivo for muito grande para a memória principal, você deve escrever seu programa para ler o arquivo em blocos, usando um laço `for` ou `while`.

7.5 Fazendo buscas em um arquivo

Quando você estiver procurando algo dentro de um arquivo, esta é uma forma comum de se percorrer todo o arquivo, ignorando a maioria das linhas e somente processando aquelas que atendam a uma condição particular. Nós podemos combinar padrões para leitura em um arquivo com os métodos da classe string para construir mecanismos simples de busca.

Por exemplo, se quisermos ler o arquivo, imprimindo apenas as linhas que iniciarem com o prefixo “From:”, podemos usar o método da classe string **startswith** para selecionar apenas as linhas com o prefixo desejado:

```
fhand = open('mbox-short.txt')
for line in fhand:
    if line.startswith('From:') :
        print line
```

Quando este programa executa, obtemos a seguinte saída:

```
From: stephen.marquard@uct.ac.za

From: louis@media.berkeley.edu
```

```
From: zqian@umich.edu

From: rjlowe@iupui.edu
...
```

O programa funcionou, uma vez que a saída imprimiu apenas aquelas linhas que iniciam com o prefixo "From:". Mas porque iríamos querer as linhas em branco? Isto se deve ao caractere invisível **newline**. Cada uma das linhas terminam com um newline, então a instrução `print` imprime a string contida na variável **line** o que inclui um newline e então a instrução print adiciona *outro* newline, resultando no efeito de duplo espaço que pudemos visualizar.

Nós podemos utilizar o método slicing para imprimir todos os caracteres menos o último, mas um método mais interessante é utilizar o método **strip** para remover o espaço em branco do lado direito da string, como segue:

```
fhand = open('mbox-short.txt')
for line in fhand:
    line = line.rstrip()
    if line.startswith('From:') :
        print line
```

Quando este programa executa, obtemos a seguinte saída:

```
From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
From: cwen@iupui.edu
...
```

Conforme seus programas de processamento de arquivo se tornam mais complicados, você pode estruturar seus laços com a instrução `continue`. A ideia básica do laço de busca é que você procura por linhas "interessantes" e efetivamente pula aquelas "não interessantes". E então quando encontrarmos uma linha interessante, podemos fazer algo com ela.

Podemos estruturar o laço para seguir o padrão de pular linhas que não interessam, como segue:

```
fhand = open('mbox-short.txt')
for line in fhand:
    line = line.rstrip()
    # Skip 'uninteresting lines'
    if not line.startswith('From:') :
        continue
    # Process our 'interesting' line
    print line
```

A saída do programa é a mesma. As linhas que não são interessantes são aquelas que não começam com "From:", as quais nós pulamos através da instrução

`continue`. As linhas interessantes (i.e., aquelas que começam com "From:") são processadas pelo nosso programa.

Podemos usar o método da classe string `find`, para simular uma busca de um editor de texto que procura por uma string em todas as linhas de um arquivo onde ela aparecer, não importa a posição da linha. A instrução `find` procura pela ocorrência de uma string em outra, retornando o índice da posição encontrada ou -1 caso não encontre. Podemos escrever o seguinte laço para mostrar as linhas que contém a string "@uct.ac.za" (i.e. originadas na Universidade de Cape Town na África do Sul):

```
fhand = open('mbox-short.txt')
for line in fhand:
    line = line.rstrip()
    if line.find('@uct.ac.za') == -1 :
        continue
    print line
```

Que produz a seguinte saída:

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
X-Authentication-Warning: set sender to stephen.marquard@uct.ac.za using -f
From: stephen.marquard@uct.ac.za
Author: stephen.marquard@uct.ac.za
From david.horwitz@uct.ac.za Fri Jan  4 07:02:32 2008
X-Authentication-Warning: set sender to david.horwitz@uct.ac.za using -f
From: david.horwitz@uct.ac.za
Author: david.horwitz@uct.ac.za
...
```

7.6 Deixando o usuário escolher o nome do arquivo

Nós não queremos ter que editar nosso código Python toda vez que tivermos que processar um arquivo diferente. É melhor pedir que o usuário digite o nome do arquivo cada vez que o programa executar, assim nosso programa pode ser utilizado para executar diferentes arquivos sem ter que ficar alterando o script Python.

Isto é muito fácil de se fazer, basta utilizarmos a instrução `raw_input` como mostrado a seguir:

```
fname = raw_input('Enter the file name: ')
fhand = open(fname)
count = 0
for line in fhand:
    if line.startswith('Subject:') :
        count = count + 1
print 'There were', count, 'subject lines in', fname
```

O nome do arquivo é lido através da entrada do usuário e armazenado em uma variável chamada `fname` e então o arquivo é aberto. Desta forma podemos executar o programa diversas vezes na leitura de diferentes arquivos.

```
python search6.py
Enter the file name: mbox.txt
There were 1797 subject lines in mbox.txt

python search6.py
Enter the file name: mbox-short.txt
There were 27 subject lines in mbox-short.txt
```

Antes de espiar a próxima seção, dê uma olhada no programa acima e pergunte a você mesmo, "O que pode dar errado aqui?" ou "O que será que o nosso amigo usuário pode querer fazer que vá fazer com que o nosso pequeno programa terminar com um erro inesperado e um traceback, fazendo com que olhemos de uma forma não tão bacana para os olhos dos nossos queridos usuários?"

7.7 Usando `try, except,` e `open`

Eu disse para você não espiar. Esta é a sua última chance

O que aconteceria se o usuário digitasse qualquer outra coisa que não fosse o nome de um arquivo?

```
python search6.py
Enter the file name: missing.txt
Traceback (most recent call last):
  File "search6.py", line 2, in <module>
    fhand = open(fname)
IOError: [Errno 2] No such file or directory: 'missing.txt'

python search6.py
Enter the file name: na na boo boo
Traceback (most recent call last):
  File "search6.py", line 2, in <module>
    fhand = open(fname)
IOError: [Errno 2] No such file or directory: 'na na boo boo'
```

Não dê risada, os usuários tentarão de todas as formas fazer com que o nosso programa dê erros—seja com um propósito ou com intenção maliciosa. Na verdade, uma importante atividade de qualquer time de desenvolvimento de software é uma pessoa ou grupo chamado Quality Assurance (ou QA), cuja principal responsabilidade é fazer as coisas mais loucas possíveis na tentativa de quebrar o software que o programador criou.

O time de QA é responsável por encontrar falhas em programas antes que ele seja entregue aos usuários finais que estão pagando o software ou o salário dos programadores. Então, o time QA são os melhores amigos dos desenvolvedores.

Então, agora que encontramos uma falha no programa, podemos consertá-lo usando a estrutura `try`/`except`. Podemos assumir que a chamada `open` pode falhar e adicionar um código de tratamento para quando o `open` falhar, como segue:

```
fname = raw_input('Enter the file name: ')
try:
    fhand = open(fname)
except:
    print 'File cannot be opened:', fname
    exit()

count = 0
for line in fhand:
    if line.startswith('Subject:') :
        count = count + 1
print 'There were', count, 'subject lines in', fname
```

A função `exit` faz que com o programa termine. Esta é uma função que chamamos e que nunca retorna. Agora quando nosso usuário (ou o time QA) digitar nomes bobos ou ruins para o nome do arquivo, nós "capturamos" os erros e tratamos de uma forma adequada.

```
python search7.py
Enter the file name: mbox.txt
There were 1797 subject lines in mbox.txt

python search7.py
Enter the file name: na na boo boo
File cannot be opened: na na boo boo
```

Proteger a chamada da função `open` é um bom exemplo do uso correto da instrução `try` e `catch` em um programa Python. Utilizamos o termo "Pythônico" quando estamos fazendo do "jeito Python". Podemos dizer que o exemplo acima é o jeito Pythônico de se abrir um arquivo.

Quando você se tornar mais qualificado em Python, pode ajudar outros programadores Python a decidir qual de duas soluções equivalentes para um determinado problema é "mais Pythônica". O objetivo de ser "mais Pythônico" remete à noção de que programação é parte da engenharia e da arte. Não estamos interessados em apenas fazer algo funcionar, queremos que a nossa solução seja elegante e apreciada por nossos colegas.

7.8 Escrevendo arquivos

Para escrever um arquivo, você deve abri-lo no modo `'w'` como segundo parâmetro.

```
>>> fout = open('output.txt', 'w')
>>> print fout
<open file 'output.txt', mode 'w' at 0xb7eb2410>
```

Se o arquivo já existir, abri-lo no modo escrita irá limpar o conteúdo do arquivo e iniciar uma escrita limpa, então tenha cuidado! Se o arquivo não existir, um novo será criado.

O método `write` de um objeto tratador "(handle)" de arquivo colocará dados dentro dele.

```
>>> line1 = "This here's the wattle,\n"
>>> fout.write(line1)
```

Novamente, o objeto file mantém o endereço de onde o arquivo está, assim, se você chamar `write` novamente, irá adicionar dados ao final do arquivo.

Devemos nos certificar de gerenciar o fim das linhas conforme escrevemos em um arquivo, explicitamente inserindo o caractere newline quando quisermos finalizar a linha. A instrução `print` adiciona automaticamente uma nova linha. A instrução `print` automaticamente adiciona uma nova linha, mas o método `write` não adiciona automaticamente uma nova linha.

```
>>> line2 = 'the emblem of our land.\n'
>>> fout.write(line2)
```

Quando você terminar de escrever, terá que fechar o arquivo para se certificar de que o último bit de dados será escrito fisicamente para o disco, assim a informação não será perdida quando a energia desligar.

```
>>> fout.close()
```

Podemos fechar os arquivos que abrirmos para leitura também, mas podemos ser um pouco negligentes somente se estivermos abrindo alguns poucos arquivos desde que o Python se certifique de fechar todos os arquivos que foram abertos quando o programa finalizar. Quando escrevermos arquivos, temos que fechar explicitamente usando a instrução `close` para não corromper o arquivo.

7.9 Depurando ou "Debugando"

Quando você estiver lendo e escrevendo arquivos, você pode ter problemas com espaços em branco. Estes erros podem ser difíceis de se depurar porque espaços, tabs e novas linhas são normalmente invisíveis:

```
>>> s = '1 2\t 3\n 4'
>>> print s
1 2  3
 4
```

A função padrão `repr` pode ajudar. Recebe qualquer objeto como um argumento e retorna a representação da string de um objeto. Para strings, os espaços em branco são representados como caracteres com sequências de \n:

```
>>> print repr(s)
'1 2\t 3\n 4'
```

Isto pode ser muito interessante para depuração.

Um outro problema que você pode ter é que diferentes sistemas usam diferentes caracteres para indicar o fim da linha. Alguns sistemas usam o newline, representado por `\n`. Outros usam um caractere de retorno, representado por `\r`. Alguns usam os dois. Se você mover-se entre estes diferentes sistemas, algumas inconsistências podem causar problemas.

Para a maioria dos sistemas, existem aplicativos para converter de um formato para o outro. Você pode achá-los (e ler mais sobre este assunto) em `wikipedia.org/wiki/Newline`. Ou, naturalmente, você pode escrever o seu próprio aplicativo.

7.10 Glossário

catch: Para prevenir uma exceção de terminar um programa usando as instruções `try` e `except`.

newline: Um caractere especial utilizado em arquivos e strings para indicar o fim de uma linha.

Pythonic: Uma técnica que funciona elegantemente no Python. "Usar try e except é um jeito *Pythônico* de se recuperar de arquivos não existentes, por exemplo".

Controle da Qualidade - QA: Uma pessoa ou time focado em garantir todo o fluxo de qualidade de um produto de software. QA é frequentemente envolvido nos testes de um produto afim de identificar problemas antes que ele seja lançado.

arquivo texto: Uma sequência de caracteres armazenada em um storage, como em um hard drive por exemplo. storage like a hard drive.

7.11 Exercícios

Exercício 7.1 Escreva um programa para ler um arquivo linha a linha e imprimir o seu conteúdo inteiro em letra maiúscula. O resultado da execução deve se parecer com o exemplo abaixo:

```
python shout.py
Enter a file name: mbox-short.txt
FROM STEPHEN.MARQUARD@UCT.AC.ZA SAT JAN  5 09:14:16 2008
RETURN-PATH: <POSTMASTER@COLLAB.SAKAIPROJECT.ORG>
RECEIVED: FROM MURDER (MAIL.UMICH.EDU [141.211.14.90])
 BY FRANKENSTEIN.MAIL.UMICH.EDU (CYRUS V2.3.8) WITH LMTPA;
 SAT, 05 JAN 2008 09:14:16 -0500
```

You can download the file from `www.py4inf.com/code/mbox-short.txt`

Exercício 7.2 Escreva um programa para perguntar o nome de um arquivo e então ler suas linhas procurando por aquelas que se enquadram no seguinte formato:

```
X-DSPAM-Confidence: 0.8475
```

Quando você encontrar uma linha que inicia com "X-DSPAM-Confidence:" destrinche a linha para extrair o ponto flutuante dela. Conte as linhas e compute o total de valores "spam confidence" que forem encontrados. Quando você atingir o final do arquivo, imprima a porcentagem de "spam confidence" encontrados.

```
Digite o nome do arquivo: mbox.txt
Porcentagem de spam confidence: 0.894128046745

Digite o nome de um arquivo: mbox-short.txt
Porcentagem de spam confidence: 0.750718518519
```

Teste seu programa utilizando os arquivos `mbox.txt` e `mbox-short.txt`.

Exercício 7.3 Algumas vezes quando programadores se entediam ou querem ter um pouco de diversão, eles adicionam um recurso escondido, que não faz mal, **Easter Egg** aos seus programas (`en.wikipedia.org/wiki/Easter_egg_(media)`). Modifique o programa que pergunta ao usuário pelo nome do arquivo e imprima uma mensagem engraçada quando o usuário digitar exatamente a expressão: "na na boo boo". O programa deve se comportar normalmente para todos os arquivos que existem ou não existem. Aqui está um exemplo da execução do programa:

```
python egg.py

Digite o nome do arquivo: mbox.txt
Existem 1797 linhas ``subject'' em mbox.txt

python egg.py
Digite o nome do arquivo: missing.tyxt
File cannot be opened: missing.tyxt

python egg.py
Digite o nome do arquivo: na na boo boo
NA NA BOO BOO PARA VOCE TAMBEM - Voceê caiu na pegadinha!
```

Nós não estamos encorajando você a colocar Easter Eggs nos seus programas—isto é apenas um exercício.

Capítulo 8

Listas

8.1 Uma lista é uma sequência

Assim como uma string, uma **lista** é uma sequência de valores. Em uma string, os valores são caracteres, já em uma lista, eles podem ser de qualquer tipo. Os valores em uma lista são chamados de **elementos** e por vezes também chamados de **itens**.

Existem diversas maneiras de se criar uma nova lista; a mais simples é colocar os elementos dentro de colchetes (`[` e `]`):

```
[10, 20, 30, 40]
['crunchy frog', 'ram bladder', 'lark vomit']
```

O primeiro exemplo é uma lista de quatro inteiros. A segunda é uma lista de três strings. Os elementos de uma lista não precisam ter o mesmo tipo. A lista a seguir contém uma string, um número flutuante, um inteiro e (lo!) outra lista:

```
['spam', 2.0, 5, [10, 20]]
```

Uma lista dentro de outra lista é chamada de lista **aninhada**.

Uma lista que não contenha elementos é chamada de uma lista vazia; você pode criar uma com colchetes vazios, `[]`.

Como você deve imaginar, você pode atribuir valores de uma lista para variáveis:

```
>>> cheeses = ['Cheddar', 'Edam', 'Gouda']
>>> numbers = [17, 123]
>>> empty = []
>>> print cheeses, numbers, empty
['Cheddar', 'Edam', 'Gouda'] [17, 123] []
```

8.2 Listas são mutáveis

A sintaxe para acessar os elementos de uma lista é a mesma utilizada para acessar os caracteres de de uma string—o operador colchetes. A expressão dentro dos colchetes especifica o índice. Lembre que os índices iniciam no 0:

```
>>> print cheeses[0]
Cheddar
```

Diferente das strings, listas são mutáveis pois é possível modificar a ordem dos itens em uma lista ou reatribuir um item da lista. Quando um operador colchete aparece ao lado esquerdo da atribuição, ele identifica o elemento da lista que será atribuído.

```
>>> numbers = [17, 123]
>>> numbers[1] = 5
>>> print numbers
[17, 5]
```

O element 1 de `numbers`, que era 123, agora é 5.

Você pode pensar em uma lista como um relacionamento entre índices e elementos. Este relacionamento é chamado de **mapeamento**; cada índice "mapeia para" um dos elementos.

índices de lista funcionam da mesma maneira que os índices de strings:

- Qualquer expressão de um inteiro pode ser usada como um índice.
- Se você tentar ler ou escrever um elemento que não existe, você terá um `IndexError`.
- Caso um índice tenha um valor negativo, ele contará ao contrário, do fim para o início da lista.

O operador `in` também funciona em listas.

```
>>> cheeses = ['Cheddar', 'Edam', 'Gouda']
>>> 'Edam' in cheeses
True
>>> 'Brie' in cheeses
False
```

8.3 Percorrendo uma lista

A maneira mais comum de se percorrer os elementos de uma lista é com um laço `for`. A sintaxe é a mesma da utilizada para strings:

```
for cheese in cheeses:
    print cheese
```

Isto funciona se você precisa ler apenas os elementos da lista. Porém, caso você precise escrever ou atualizar elementos, você precisa de índices. Um forma comum de fazer isto é combinar as funções `range` e `len`:

```
for i in range(len(numbers)):
    numbers[i] = numbers[i] * 2
```

Este laço percorre a lista e atualiza cada elemento. `len` retorna o número de elementos na lista. `range` retorna uma lista de índices de 0 a $n-1$, onde n é o tamanho da lista. Cada vez que passa pelo laço, `i` recebe o índice do próximo elemento. A instrução de atribuição no corpo, utiliza `i` para ler o valor antigo do elemento e atribuir ao novo valor.

Um laço `for` em uma lista vazia nunca executa as instruções dentro do laço:

```
for x in empty:
    print 'Esta linha nunca será executada.'
```

Embora uma lista possa conter outra lista, a lista aninhada ainda conta como um único elemento. O tamanho dessa lista é quatro:

```
['spam', 1, ['Brie', 'Roquefort', 'Pol le Veq'], [1, 2, 3]]
```

8.4 Operações de Lista

O operador + concatena listas:

```
>>> a = [1, 2, 3]
>>> b = [4, 5, 6]
>>> c = a + b
>>> print c
[1, 2, 3, 4, 5, 6]
```

De modo parecido, o operador `*` repete uma lista pelo número de vezes informado:

```
>>> [0] * 4
[0, 0, 0, 0]
>>> [1, 2, 3] * 3
[1, 2, 3, 1, 2, 3, 1, 2, 3]
```

O primeiro exemplo repete `[0]` quatro vezes. O segundo exemplo repete a lista `[1, 2, 3]` três vezes.

8.5 Fatiamento de Lista

O operador de fatiamento também funciona em listas:

```
>>> t = ['a', 'b', 'c', 'd', 'e', 'f']
>>> t[1:3]
['b', 'c']
>>> t[:4]
['a', 'b', 'c', 'd']
>>> t[3:]
['d', 'e', 'f']
```

Se você omite o primeiro índice, o fatiamento é iniciado no começo da lista. Se omitir o segundo, o fatiamento vai até fim. Então se ambos são omitidos, a fatia é uma cópia da lista inteira.

```
>>> t[:]
['a', 'b', 'c', 'd', 'e', 'f']
```

Uma vez que lista são mutáveis, com frequência é útil fazer uma cópia antes de realizar operações que dobram, reviram ou mutilam listas.

Um operador de fatiamento do lado esquerdo de uma atribuição pode atualizar múltiplos elementos.

```
>>> t = ['a', 'b', 'c', 'd', 'e', 'f']
>>> t[1:3] = ['x', 'y']
>>> print t
['a', 'x', 'y', 'd', 'e', 'f']
```

8.6 Métodos de lista

O Python provê métodos que operam nas listas. Por exemplo, `append` adiciona um novo elemento ao fim da lista:

```
>>> t = ['a', 'b', 'c']
>>> t.append('d')
>>> print t
['a', 'b', 'c', 'd']
```

`extend` recebe uma lista como argumento e adiciona todos seus elementos.

```
>>> t1 = ['a', 'b', 'c']
>>> t2 = ['d', 'e']
>>> t1.extend(t2)
>>> print t1
['a', 'b', 'c', 'd', 'e']
```

Este exemplo deixa `t2` sem modificação.

`sort` organiza os elementos da lista do menor para o maior:

```
>>> t = ['d', 'c', 'e', 'b', 'a']
>>> t.sort()
>>> print t
['a', 'b', 'c', 'd', 'e']
```

A maior parte dos métodos de lista são vazios; eles modificam a lista e retornam None. Caso você acidentalmente escreva t = t.sort(), ficará desapontado com o resultado.

8.7 Deletando elementos

Existem diversas maneiras de se deletar elementos de uma lista. Se você souber o índice do elemento que você quer, pode usar o pop:

```
>>> t = ['a', 'b', 'c']
>>> x = t.pop(1)
>>> print t
['a', 'c']
>>> print x
b
```

pop modifica a lista e retorna o elemento que foi removido. Se você não informa um índice, ele deletará e retornará o último elemento da lista.

Se você não precisa do valor removido, poderá usar o operador del:

```
>>> t = ['a', 'b', 'c']
>>> del t[1]
>>> print t
['a', 'c']
```

Se você sabe qual elemento você quer remover (mas não sabe o índice), você pode usar o remove:

```
>>> t = ['a', 'b', 'c']
>>> t.remove('b')
>>> print t
['a', 'c']
```

O valor retornado de remove é None.

Para remover mais de um elemento, você pode usar del com um índice de fatiamento:

```
>>> t = ['a', 'b', 'c', 'd', 'e', 'f']
>>> del t[1:5]
>>> print t
['a', 'f']
```

Como de costume, uma fatia seleciona todos os elementos até o segundo índice, porém sem incluí-lo.

8.8 Listas e funções

Existem várias funções built-in que podem ser usadas em listas, permitindo que você tenha uma visão rápida da lista sem a necessidade de escrever o seu próprio laço:

```
>>> nums = [3, 41, 12, 9, 74, 15]
>>> print len(nums)
6
>>> print max(nums)
74
>>> print min(nums)
3
>>> print sum(nums)
154
>>> print sum(nums)/len(nums)
25
```

A função `sum()` funciona apenas quando os elementos da lista são números. As outras funções (`max()`, `len()`, etc.) funcionam com listas de strings e outros tipos que são comparáveis.

Nós podemos reescrever um programa anterior que computou a média de uma lista de números adicionados pelo usuário utilizando uma lista.

Primeiramente, o programa para calcular uma média sem uma lista:

```
total = 0
count = 0
while ( True ) :
    inp = raw_input('Digite um número: ')
    if inp == 'done' : break
    value = float(inp)
    total = total + value
    count = count + 1

average = total / count
print 'Average:', average
```

Neste programa, temos as variáveis `count` e `total` para armazenar a contagem e o total da soma dos número que o usuário digitou, enquanto pedimos mais números para o usuário.

Nós poderíamos simplesmente guardar cada número a medida que o usuário vai adicionando e usar funções built-in para calcular a soma e a contagem no final.

```
numlist = list()
while ( True ) :
    inp = raw_input('Digite um número: ')
    if inp == 'done' : break
    value = float(inp)
    numlist.append(value)

average = sum(numlist) / len(numlist)
print 'Média:', average
```

Nós criamos uma lista vazia antes do loop iniciar, e então sempre que tivermos um número, este será adicionado na lista. Ao final do programa, calcularemos a soma dos números da lista e dividiremos o total pela contagem de números na lista para chegar a média.

8.9 Listas e strings

Uma string é uma sequência de caracteres e uma lista é uma sequência de valores, porém, uma lista de caracteres não é o mesmo que uma string. Para converter uma string para lista de caracteres você pode usar `list`:

```
>>> s = 'spam'
>>> t = list(s)
>>> print t
['s', 'p', 'a', 'm']
```

Em razão de `list` ser o nome de uma função built-in, você deve evitar usar isto como nome de variável. Eu também evito a letra `l` pois se parece muito com o número `1`. Por essa razão utilizo `t`.

A função `list` quebra uma string em letras individuais. Se você deseja quebrar uma string em palavras, você deve usar o método `split`.

```
>>> s = 'pining for the fjords'
>>> t = s.split()
>>> print t
['pining', 'for', 'the', 'fjords']
>>> print t[2]
the
```

Uma vez que você usou `split` para quebrar uma string em uma lista de palavras, você pode usar o operador de índice (colochete) para ver uma palavra em particular dentro da lista.

Você pode chamar `split` com um argumento opcional chamado **delimitador** que especifica quais caracteres a serem usados como delimitadores de palavra. O exemplo a seguir usa um hífen como delimitador:

```
>>> s = 'spam-spam-spam'
>>> delimiter = '-'
>>> s.split(delimiter)
['spam', 'spam', 'spam']
```

`join` é o inverso de `split`. Ele recebe uma lista de strings e concatena seus elementos. `join` é um método da classe string, então você pode invocá-lo no delimitador e passar a lista como parâmetro.

```
>>> t = ['pining', 'for', 'the', 'fjords']
>>> delimiter = ' '
>>> delimiter.join(t)
'pining for the fjords'
```

Neste caso, o delimitador é um caractere espaço, então `join` coloca um espaço entre as palavras. Para concatenar strings sem espaços você pode usar uma string vazia, `''`, como delimitador.

8.10 Analisando linhas de um texto

Normalmente quando estamos lendo um arquivo, queremos fazer algo com as linhas e não somente imprimir a linha inteira. Frequentemente queremos encontrar as "linhas interessantes" e então **analisar** a linha para encontrar a *parte* interessante da linha. E se quiséssemos imprimir o dia da semana das linhas que começam com "From "?

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
```

O método `split` é muito efetivo quando temos este tipo de problema. Podemos escrever um pequeno programa que procure por linhas onde a linha inicia com "From ", dividir essas linhas, e então imprimir a terceira palavra da linha:

```
fhand = open('mbox-short.txt')
for line in fhand:
    line = line.rstrip()
    if not line.startswith('From ') : continue
    words = line.split()
    print words[2]
```

Aqui também utilizamos o `if` de forma contraída onde colocamos o `continue` na mesma linha do `if`. A forma contraída do `if` funciona da mesma maneira que funcionaria se o `continue` estivesse na próxima linha e indentado.

O programa produz a saída a seguir:

```
Sat
Fri
Fri
Fri
    ...
```

Futuramente, iremos aprender técnicas cada vez mais sofisticadas para pegar as linhas e como separar essas linhas para encontrar a informação exata que estamos procurando.

8.11 Objetos e valores

Se executarmos estas instruções de atribuição:

```
a = 'banana'
b = 'banana'
```

Sabemos que ambos `a` e `b` se referem a uma string, mas não sabemos se eles se referem a *mesma* string. Aqui estão duas possibilidades:

Em um caso, `a` e `b` se referem a dois objetos diferentes que tem o mesmo valor. No segundo caso, eles se referem ao mesmo objeto.

Para checar se duas variáveis referem-se ao mesmo objeto, você pode utilizar o operador `is`.

```
>>> a = 'banana'
>>> b = 'banana'
>>> a is b
True
```

Neste exemplo, o Python apenas criou um objeto string, e ambos `a` e `b` referem-se a ele.

Porém, quando você cria duas listas, você tem dois objetos:

```
>>> a = [1, 2, 3]
>>> b = [1, 2, 3]
>>> a is b
False
```

Neste caso, diríamos que as duas listas são **equivalentes**, pois possuem os mesmos elementos, mas não são **idênticas**, já que não são o mesmo objeto. Se dois objetos são idênticos, eles também são equivalentes, porém se eles são equivalentes, não são necessariamente idênticos.

Até agora estivemos utilizando a nomenclatura "objeto" ou "valor", mas, é mais preciso dizer que um objeto tem um valor. Se você executa `a = [1,2,3]`, `a` refere-se a um objeto lista do qual o valor é uma sequência particular de elementos. Se outra lista tem os mesmos elementos, diríamos que tem o mesmo valor.

8.12 Aliasing - Interferência entre variáveis

Se `a` refere-se a um objeto e você atribui `b = a`, então ambas as variáveis referem-se ao mesmo objeto:

```
>>> a = [1, 2, 3]
>>> b = a
>>> b is a
True
```

A associação de uma variável com um objeto é chamada **referência**. Neste exemplo existem duas referências para o mesmo objeto.

Um objeto com mais de uma referência tem mais de um nome, então dizemos que o objeto é **aliased**.

Se o objeto **aliased** é mutável, modificações feitas com um alias afetará as outras:

```
>>> b[0] = 17
>>> print a
[17, 2, 3]
```

Embora este comportamento possa ser útil, é passível de erro. De maneira geral é mais seguro evitar **aliasing** quando você está trabalhando com objetos mutáveis.

Para objetos imutáveis como strings, **aliasing** não chega a ser um problema. Neste exemplo:

```
a = 'banana'
b = 'banana'
```

Isso quase nunca faz diferença, se `a` e `b` fazem referência à mesma string ou não.

8.13 Argumentos de Lista

Quando você passa uma lista para uma função, a função pega uma referência para a lista. Se a função modifica a lista passada como argumento, o "caller"vê a mudança. Por exemplo, `delete_head` remove o primeiro elemento da lista:

```
def delete_head(t):
    del t[0]
```

Aqui está como isto é utilizado:

```
>>> letters = ['a', 'b', 'c']
>>> delete_head(letters)
>>> print letters
['b', 'c']
```

O parâmetro `t` e a variável `letters` são **aliases** para o mesmo objeto.

Isso é importante para distinguir entre operações que modificam listas e operações que criam novas listas. Por exemplo, o método `append` modifica uma lista, mas o operador + cria uma nova lista:

```
>>> t1 = [1, 2]
>>> t2 = t1.append(3)
>>> print t1
[1, 2, 3]
>>> print t2
None

>>> t3 = t1 + [3]
>>> print t3
[1, 2, 3]
>>> t2 is t3
False
```

Esta diferença é importante quando você escreve funções que supostamente devem modificar listas. Por exemplo, esta função *não* deleta o início de uma lista:

```
def bad_delete_head(t):
    t = t[1:]              # ERRADO!
```

O operador de fatiamento cria uma nova lista e a atribuição faz `t` se referir a isto, porém nada disso tem efeito na lista passada como argumento.

Uma alternativa é escrever uma função que cria e retorna uma nova lista. Por exemplo, `tail` retorna tudo, menos o primeiro elemento de uma lista:

```
def tail(t):
    return t[1:]
```

Esta função deixa a lista original inalterada. Aqui está como isto é utilizado:

```
>>> letters = ['a', 'b', 'c']
>>> rest = tail(letters)
>>> print rest
['b', 'c']
```

Exercício 8.1 Escreva uma função chamada `chop` que recebe uma lista e a modifica, removendo o primeiro e o último elementos e retorna `None`.

Então escreva uma função chamada `middle` que recebe uma lista e retorna uma nova lista que contenha tudo menos o primeiro e o último elementos.

8.14 Depurando

O uso descuidado de listas (e outros objetos mutáveis) pode levar a longas horas de depuração. Aqui estão algumas das armadilhas mais comuns e maneiras de evitá-las.

1. Não esqueça que a maioria dos métodos de lista modificam o argumento e retornam `None`. Isto é o oposto dos métodos de string, os quais retornam uma nova string e deixam o original inalterado.

   Se você está acostumado a escrever código para strings assim:

   ```
   word = word.strip()
   ```

   É tentador escrever código para lista assim:

   ```
   t = t.sort()           # ERRADO!
   ```

   Por `sort` retornar `None`, a próxima operação que você executar com `tt` provavelmente falhará.

   Antes de usar métodos e operadores de lista você deveria ler a documentação com cuidado e então testá-los no modo interativo. Os métodos e operadores que as listas compartilham com outras sequências (como strings) são documentados em `https://docs.python.org/2/library/stdtypes.html#string-methods`. Os métodos e operadores que se aplicam apenas a sequências mutáveis são documentados em: `https://docs.python.org/2/library/stdtypes.html#mutable-sequence-types`.

2. Pegue um idioma e fique como ele.

   Parte do problema com listas é que existem muitas maneiras de fazer as coisas. Por exemplo, para remover um elemento de uma lista, você pode usar `pop`, `remove`, `del`, ou mesmo atribuição de um fatiamento (slice).

   Para adicionar um elemento, você pode utilizar os métodos `append` ou o operador `+`. Mas não esqueça que esses estão corretos:

   ```
   t.append(x)
   t = t + [x]
   ```

   E esses estão errados:

   ```
   t.append([x])          # ERRADO!
   t = t.append(x)        # ERRADO!
   t + [x]                # ERRADO!
   t = t + x              # ERRADO!
   ```

   Experimente cada um desses exemplos no modo interativo para ter certeza que você entende o que eles fazem. Note que apenas o último causa um erro de runtime; os outros três são legais, mas fazem a coisa errada.

3. Faça cópias para evitar aliasing.

   Se você quer usar um método como `sort` que modifica o argumento, mas você também precisa manter a lista original, você pode fazer uma cópia.

   ```
   orig = t[:]
   t.sort()
   ```

   Neste exemplo você também pode usar a função built-in `sorted`, a qual retorna uma nova lista ordenada e deixa a original inalterada. Mas, neste caso você deve evitar `sorted` como um nome de variável!

4. Listas, `split`, e arquivos

   Quando lemos e analisamos arquivos, existem muitas oportunidades para encontrar entradas que podem causar falhas em nosso programa, então é uma boa ideia revisitar o padrão **protetor** quando se trata de escrever programas que leiam de um arquivo e procurem por uma "agulha no palheiro".

   Vamos revisitar nosso programa que procura pelo dia da semana nas linhas do nosso arquivo:

   ```
   From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
   ```

   Já que estamos quebrando esta linha em palavras, poderíamos distribuir isso com o uso do `startswith` e simplesmente olhar a primeira palavra da linha para determinar se estamos interessados na linha. Podemos usar `continue` para pular linhas que não possuem "From" como primeira palavra:

```
fhand = open('mbox-short.txt')
for line in fhand:
    words = line.split()
    if words[0] != 'From' : continue
    print words[2]
```

Isso parece muito mais simples e nós nem mesmo precisamos fazer o `rstrip` para remover o newline ao final do arquivo. Mas, é melhor assim?

```
python search8.py
Sat
Traceback (most recent call last):
  File "search8.py", line 5, in <module>
    if words[0] != 'From' : continue
IndexError: list index out of range
```

Funciona de certa maneira e vemos o dia da primeira (Sat), mas então o programa falha com um erro traceback. O que deu errado? Que dados bagunçados causaram a falha do nosso elegante, inteligente e Pythonico programa?

Você pode ficar olhando por um longo tempo e tentar decifrá-lo ou pedir ajuda para alguém, porém a abordagem mais rápida e inteligente é adicionar um `print`. O melhor lugar para colocar um print é logo antes da linha onde o programa falhou e imprimir os dados que parecem estar causando a falha.

Essa abordagem deve gerar muitas linhas na saída do programa, mas, ao menos você imediatamente terá alguma pista sobre o problema. Então adicione um print da variável `words` logo antes da linha cinco. Nós até mesmo colocamos um prefixo: "Debug:" na linha, assim podemos manter nossa saída normal separada da saída de debug.

```
for line in fhand:
    words = line.split()
    print 'Debug:', words
    if words[0] != 'From' : continue
    print words[2]
```

Quando executamos o programa, há muita saída passando pela tela, mas ao fim vemos nossa saída de debug e um traceback, dessa forma sabemos o que aconteceu antes do traceback.

```
Debug: ['X-DSPAM-Confidence:', '0.8475']
Debug: ['X-DSPAM-Probability:', '0.0000']
Debug: []
Traceback (most recent call last):
  File "search9.py", line 6, in <module>
    if words[0] != 'From' : continue
IndexError: list index out of range
```

Cada linha de debug imprime uma lista de palavras que temos quando quando dividimos a linha em palavras `split`. Quando o programa falha,

a lista de palavras está vazia `[]`. Se abrirmos um arquivo em um editor de texto e olharmos neste ponto, ele parecerá conforme a seguir:

```
X-DSPAM-Result: Innocent
X-DSPAM-Processed: Sat Jan  5 09:14:16 2008
X-DSPAM-Confidence: 0.8475
X-DSPAM-Probability: 0.0000

Detalhes: http://source.sakaiproject.org/viewsvn/?view=rev&rev=39772
```

O erro ocorre quando nosso programa encontra uma linha em branco! Claro, uma linha em branco tem "zero palavras". Porque não pensamos nisso quando estávamos escrevendo o código? Quando o código procura pela primeira palavra (`word[0]`) para ver se encontra "From", nós então temos um erro "index out of range".

Este é claro, o lugar perfeito para adicionar algum código **protetor** para evitar a checagem da primeira palavra caso ela não exista. Existem muitas maneiras de proteger este código; escolheremos checar o número de palavras que temos antes de olharmos a primeira palavra:

```
fhand = open('mbox-short.txt')
count = 0
for line in fhand:
    words = line.split()
    # print 'Debug:', words
    if len(words) == 0 : continue
    if words[0] != 'From' : continue
    print words[2]
```

Primeiramente, comentamos o print de debug ao invés de removê-lo, para caso nossa modificação falhe, precisaremos investigar novamente. Então adicionamos uma instrução protetora que verifica se temos zero palavras, caso positivo, usamos `continue` para pular para a próxima linha no arquivo.

Podemos pensar nas duas instruções `continue` nos ajudando a refinar o conjunto de linhas que são "interessantes" para nós e quais queremos processar mais um pouco. Uma linha que não tenha palavras "não é interessante" para nós então, pulamos para a próxima linha. Uma linha que não tenha "From" como a sua primeira palavra não é interessante para nós, então nós a pulamos.

O programa, da forma como foi modificado, executa com sucesso, então talvez esteja correto. Nossa instrução protetora assegurará que `words[0]` nunca falhará, mas talvez isso não seja o suficiente. Quando estamos programando, devemos sempre estar pensando, "O que pode dar errado?"

Exercício 8.2 Descubra qual linha do programa acima, ainda não está corretamente protegida. Veja se você pode construir um arquivo de texto que causará falha no programa e então modifique o programa para que então a linha esteja corretamente protegida e teste para ter certeza de que o programa processará o novo arquivo de texto.

Exercício 8.3 Reescreva o código protetor, no exemplo acima, sem as duas instruções `if`. Ao invés disso, use uma expressão lógica combinada com o operador lógico `and` com apenas uma instrução `if`.

8.15 Glossário

aliasing: Uma circunstância onde duas ou mais variáveis, referem-se ao mesmo objeto.

delimitador: Um caractere (ou string) usada para indicar onde uma string deve ser dividida.

elemento: Um dos valores em uma lista (ou outra sequência); também chamado de itens.

equivalente: Ter os mesmos valores.

index: Um valor inteiro que indica um elemento em uma lista.

idêntico: É o mesmo objeto (o que indica equivalência).

lista: Uma sequência de valores.

percorrer lista: Acesso sequencial a cada elemento de uma lista.

lista aninhada: Uma lista que é um elemento de outra lista.

objeto: Algo a que uma variável pode se referir. Uma objeto tem um tipo e valor.

referência: Uma associação entre uma variável e seu valor.

8.16 Exercícios

Exercício 8.4 Faça o download de uma cópia do arquivo em `www.py4inf.com/code/romeo.txt`

Escreva um programa para abrir o arquivo `romeo.txt` e ler linha por linha. Para cada linha, divida a linha em uma lista de palavras usando a função `split`.

Para cada palavra, verifique se a palavra já está em uma lista. Se a palavra não está na lista, adicione à lista.

Quando o programa completar, ordene e imprima as palavras resultantes em ordem alfabética.

```
Enter file: romeo.txt
['Arise', 'But', 'It', 'Juliet', 'Who', 'already',
'and', 'breaks', 'east', 'envious', 'fair', 'grief',
'is', 'kill', 'light', 'moon', 'pale', 'sick', 'soft',
'sun', 'the', 'through', 'what', 'window',
'with', 'yonder']
```

Exercício 8.5 Escreva um programa para ler os dados do mail box e quando você achar uma linha que inicie com "From", você dividirá a linha em palavras usando a função `split`. Estamos interessados em quem enviou a mensagem, que é a segunda palavra na linha do From.

```
From stephen.marquard@uct.ac.za Sat Jan 5 09:14:16 2008
```

Você irá analisar a linha do From, imprimir a segunda palavra para cada linha com From, então você também contará o número de linhas com From (e não From:) e imprimirá e calculará ao final.

Este é um bom exemplo de saída com algumas linhas removidas:

```
python fromcount.py
Enter a file name: mbox-short.txt
stephen.marquard@uct.ac.za
louis@media.berkeley.edu
zqian@umich.edu

[...Parte da saída removida...]

ray@media.berkeley.edu
cwen@iupui.edu
cwen@iupui.edu
cwen@iupui.edu
Existiam 27 linhas no arquivos onde From era a primeira palavra
```

Exercício 8.6 Reescreva o programa que leva o usuário para uma lista de números e imprime o máximo e o mínimo para os números no fim quando o usuário digita "done". Escreva um programa para armazenar em uma lista, os números que o usuário digitar e use as funções `max()` e `min()` para calcular o máximo e o mínimo ao fim do loop.

```
Digite um número: 6
Digite um número: 2
Digite um número: 9
Digite um número: 3
Digite um número: 5
Digite um número: done
Maximum: 9.0
Minimum: 2.0
```

Capítulo 9

Dicionários

Um **dicionário** é como uma lista, porém mais abrangente. Em uma lista, os índices devem ser valores inteiros; em um dicionário, os índices podem ser de qualquer tipo (praticamente).

Pode-se considerar um dicionário como um mapeamento entre um conjunto de índices (chamados de **chaves**) e um conjunto de valores. Cada chave é mapeada a um valor. A associação entre uma chave e um valor é chamada de **par chave-valor** ou também como um **item**.

Como exemplo, construiremos um dicionário que mapeia palavras inglesas para palavras em espanhol, portanto chaves e valores são strings.

A função `dict` cria um novo dicionário sem itens. Pelo fato de `dict` ser o nome de uma função padrão da linguagem, esse termo não pode ser usado como nome de variável.

```
>>> eng2ptbr = dict()
>>> print eng2ptbr
{}
```

Os caracteres chaves, `{}`, representam um dicionário vazio. Colchetes podem ser utilizados para adicionar itens ao dicionário:

```
>>> eng2ptbr['one'] = 'um'
```

Esta linha cria um item que mapeia da chave `'one'` para o valor `'um'`. Se exibirmos o dicionário novamente, veremos um par chave-valor com o caractere dois-pontos entre a chave e o valor:

```
>>> print eng2ptbr
{'one': 'um'}
```

Esse formato de saída também é um formato de entrada. Por exemplo, pode-se criar um novo dicionário com três itens:

```
>>> eng2ptbr = {'one': 'um', 'two': 'dois', 'three': 'tres'}
```

Mas se exibirmos `eng2ptbr`, podemos nos surpreender:

```
>>> print eng2ptbr
{'one': 'um', 'three': 'tres', 'two': 'dois'}
```

A ordem dos pares chave-valor não é a mesma. De fato, se esse mesmo exemplo for executado em outro computador, um resultado diferente pode ser obtido. Em linhas gerais, a ordem dos elementos em um dicionário é imprevisível.

Entretanto, isso não é um problema, uma vez que os elementos de um dicionário nunca são indexados por índices inteiros. Ao invés disso, usa-se as chaves para se buscar os valores correspondentes:

```
>>> print eng2ptbr['two']
'dois'
```

A ordem dos itens não importa, já que a chave `'two'` sempre é mapeada ao valor `'dois'`.

Se a chave não está no dicionário, uma exceção é levantada:

```
>>> print eng2ptbr['four']
KeyError: 'four'
```

A função `len` também pode ser usada em dicionários; ela devolve o número de pares chave-valor:

```
>>> len(eng2ptbr)
3
```

Pode-se utilizar o operador `in` para se verificar se algo está representado como uma *chave* no dicionário (não serve para verificar diretamente a presença de um valor).

```
>>> 'one' in eng2ptbr
True
>>> 'um' in eng2ptbr
False
```

Para verificar se algo está representado como um valor no dicionário, pode-se usar o método `values`, o qual devolve os valores como uma lista e, desse modo, o operador `in` pode ser usado:

```
>>> vals = eng2ptbr.values()
>>> 'um' in vals
True
```

O operador `in` usa algoritmos diferentes para listas e dicionários. Para listas é usado um algoritmo de busca linear. Conforme o tamanho da lista aumenta, o tempo de busca aumenta de maneira diretamente proporcional ao tamanho da lista. Para dicionários, Python usa um algoritmo chamado **tabela de hash**, a qual possui

uma propriedade notável—o operador `in` consume a mesma quantidade de tempo para se realizar a busca independente do número de itens existente no dicionário. Aqui não será explicado o porquê das funções de hash serem tão mágicas, mas informações adicionais sobre esse assunto podem ser lidas em `pt.wikipedia.org/wiki/Tabela_de_disperso`.

Exercício 9.1 Escreva um programa que leia as palavras do arquivo `words.txt` e armazene-as como chaves em um dicionário. Os valores não importam. Então, use o operador `in` como uma maneira rápida de verificar se uma string está no dicionário.

9.1 Dicionário como um conjunto de contagens

Suponha que dada uma string deseja-se saber quantas vezes aparece cada letra. Há várias maneiras para que isso seja feito:

1. Poderiam ser criadas 26 variáveis, cada uma contendo uma letra do alfabeto. Então, a string poderia ser travessada e, para cada caractere, seria incrementado o contador correspondente, provavelmente utilizando-se operadores condicionais encadeado.

2. Poderia ser criada uma lista com 26 elementos. Assim, cada caractere poderia ser convertido em um número (usando a função embutida `ord`), o qual seria usado como um índice na lista, e se incrementaria o contador apropriado.

3. Poderia ser criado um dicionário, onde os caracteres são as chaves e os valores são as contagens correspondentes. Ao se encontrar um caractere pela primeira vez, um item é adicionado ao dicionário. Em seguida, o valor de um dado item seria incrementado.

Essas opções realizam a mesma computação, porém cada uma a implementa de um modo diferente.

Uma **implementação** é um modo de se executar uma computação; algumas implementações são melhores do que outras. Por exemplo, uma das vantagens de se utilizar a implementação com dicionário é que não há a necessidade de se saber de antemão quais letras aparecem na string, sendo que as letras serão adicionadas ao dicionário conforme for demandado.

Eis como o código ficaria:

```
word = 'brontosaurus'
d = dict()
for c in word:
    if c not in d:
        d[c] = 1
```

```
    else:
        d[c] = d[c] + 1
print d
```

De fato está sendo construído um **histograma**, que é um termo estatístico para um conjunto de contagens (ou frequências).

O laço `for` caminha por toda a string. Em cada iteração, se o caractere `c` não está no dicionário, cria-se um novo item com chave `c` e valor inicial 1 (já que essa letra foi encontrada um vez). Se `c` já está no dicionário, o valor `d[c]` é incrementado.

Eis a saída do programa:

```
{'a': 1, 'b': 1, 'o': 2, 'n': 1, 's': 2, 'r': 2, 'u': 2, 't': 1}
```

O histograma indica que as letras `'a'` e `'b'` aparecem uma vez; `'o'` aparece duas vezes, e assim por diante.

Dicionários têm um método chamado `get`, que recebe como argumento uma chave e um valor padrão. Se a chave se encontra no dicionário, `get` devolve o valor correspondente; caso contrário, devolve o valor padrão. Por exemplo:

```
>>> counts = { 'chuck' : 1 , 'annie' : 42, 'jan': 100}
>>> print counts.get('jan', 0)
100
>>> print counts.get('tim', 0)
0
```

O método `get` pode ser usado para escrever o histograma de maneira mais concisa. Pelo fato de `get` automaticamente lidar com a ausência de uma chave no dicionário, quatro linhas de código podem ser reduzidas para uma e o bloco `if` pode ser removido.

```
word = 'brontosaurus'
d = dict()
for c in word:
    d[c] = d.get(c,0) + 1
print d
```

O uso do método `get` para simplificar esse laço de contagem é um "idiomatismo" comum em Python e será usado diversas vezes no decorrer do livro. Desse modo, vale a pena dedicar um tempo e comparar o laço usando `if` e o operador `in` com o laço usando o método `get`. Eles fazem exatamente a mesma coisa, mas o segundo é mais sucinto.

9.2 Dicionários e arquivos

Um dos usos comuns de dicionários é na contagem da ocorrência de palavras em arquivos de texto. Comecemos com um arquivo muito simples contendo palavras extraídas de *Romeu e Julieta*.

Para os primeiros exemplos, usaremos uma versão mais curta e simplificada do texto, sem pontuações. Em seguida, trabalharemos com o texto da cena com as pontuações incluídas.

```
But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

Escreveremos um programa em Python, que lerá as linhas do arquivo, transformará cada linha em uma lista de palavras e, então, iterará sobre cada palavra na linha contando-a usando um dicionário.

Veremos que temos dois laços `for`. O laço externo lê as linhas do arquivo e o interno percorre cada palavra de uma linha em particular. Este é um exemplo de um padrão chamado **laços aninhados** porque um dos laços é *externo* e o outro é *interno*.

Pelo fato do laço interno executar todas suas iterações para cada uma que o laço externo faz, diz-se que o laço interno itera "mais rapidamente" ao passo que o externo itera mais lentamente.

A combinação dos laços aninhados garante que contaremos todas as palavra de todas as linhas do arquivo de entrada.

```
fname = raw_input('Digite o nome do arquivo: ')
try:
    fhand = open(fname)
except:
    print 'Arquivo nao pode ser aberto:', fname
    exit()

counts = dict()
for line in fhand:
    words = line.split()
    for word in words:
        if word not in counts:
            counts[word] = 1
        else:
            counts[word] += 1

print counts
```

Quando rodamos o programa, vemos o resultado bruto das contagens de modo não sorteado. (o arquivo `romeo.txt` está disponível em `www.py4inf.com/code/romeo.txt`)

```
python count1.py
Digite o nome do arquivo: romeo.txt
{'and': 3, 'envious': 1, 'already': 1, 'fair': 1,
'is': 3, 'through': 1, 'pale': 1, 'yonder': 1,
'what': 1, 'sun': 2, 'Who': 1, 'But': 1, 'moon': 1,
'window': 1, 'sick': 1, 'east': 1, 'breaks': 1,
```

```
'grief': 1, 'with': 1, 'light': 1, 'It': 1, 'Arise': 1,
'kill': 1, 'the': 3, 'soft': 1, 'Juliet': 1}
```

É um tanto quanto inconveniente procurar visualmente em um dicionário por palavras mais comuns e suas contagens. Desse modo, precisamos adicionar mais código Python para obter um resultado que seja mais útil.

9.3 Laços de repetição e dicionário

Se um dicionário for usado como a sequência em um bloco `for`, esse iterará sobre as chaves do dicionário. Este laço exibe cada chave e o valor correspondente:

```
counts = { 'chuck' : 1 , 'annie' : 42, 'jan': 100}
for key in counts:
    print key, counts[key]
```

Que resulta em:

```
jan 100
chuck 1
annie 42
```

Mais uma vez, as chaves não respeitam nenhum tipo de ordenamento.

Podemos usar este padrão para implementar os diferentes estilos de laço que foram descritos anteriormente. Por exemplo, se quiséssemos encontrar todas as entradas em um dicionário com valor acima de dez, poderíamos escrever o seguinte código:

```
counts = { 'chuck' : 1 , 'annie' : 42, 'jan': 100}
for key in counts:
    if counts[key] > 10 :
        print key, counts[key]
```

O laço `for` itera pelas *chaves* do dicionário, então devemos usar o operador de índice para obter o *valor* correspondente para cada chave. Eis o resultado da execução:

```
jan 100
annie 42
```

Vemos apenas as entradas com valor acima de dez.

Para exibir as chaves em ordem alfabética, deve-se gerar uma lista das chaves do dicionário por meio do método `keys`, disponível em objetos dicionário, e então ordenar essa lista. Em seguida, itera-se pela lista ordenada, procurando cada chave e exibindo os pares chave-valor de modo ordenado, como em:

```
counts = { 'chuck' : 1 , 'annie' : 42, 'jan': 100}
lst = counts.keys()
print lst
lst.sort()
for key in lst:
    print key, counts[key]
```

O que gera a seguinte saída:

```
['jan', 'chuck', 'annie']
annie 42
chuck 1
jan 100
```

Primeiramente, pode-se ver a lista não ordenada das chaves, obtida pelo método `keys`. Em seguida, vemos os pares chave-valor gerados no laço `for`.

9.4 Processamento avançado de texto

No exemplo acima, no qual usamos o arquivo `romeo.txt`, todas as pontuações foram removidas para tornar o texto o mais simples possível. O texto original possui muitas pontuações, como mostrado abaixo.

```
But, soft! what light through yonder window breaks?
It is the east, and Juliet is the sun.
Arise, fair sun, and kill the envious moon,
Who is already sick and pale with grief,
```

Uma vez que a função do Python `split` procura por espaços e trata palavras como tokens separados por espaços, as palavras "soft!" e "soft" seriam tratadas como *diferentes* e seriam criadas entradas separadas no dicionário para cada uma delas.

Além disso, como o arquivos possui letras capitalizadas, as palavras "who" e "Who" seriam tratadas como diferentes e teriam contagens diferentes.

Podemos solucionar ambos os problemas usando os métodos de string `lower`, `punctuation` e `translate`. Dentre esses três o método `translate` é o mais complexo. Eis a documentação para `translate`:

```
string.translate(s, table[, deletechars])
```

Deleta todos os caracteres de s que estão em deletechars (se presente) e traduz os caracteres usando table, que deve ser uma string com o comprimento de 256 caracteres fornecendo a tradução para cada valor de caractere, indexado pelo sua posição. Se table é None, então apenas a deleção de caracteres é realizada.

Não iremos especificar o parâmetro `table`, mas iremos usar `deletechars` para deletar todas as pontuações. Iremos utilizar a lista de caracteres que o próprio Python considera como "pontuação":

```
>>> import string
>>> string.punctuation
'!"#$%&\'()*+,-./:;<=>?@[\\]^_`{|}~'
```

Faremos a seguinte modificação em nosso programa:

```
import string                                          # New Code

fname = raw_input('Digite o nome do arquivo: ')
try:
    fhand = open(fname)
except:
    print 'Arquivo nao pode ser aberto:', fname
    exit()

counts = dict()
for line in fhand:
    line = line.translate(None, string.punctuation)    # New Code
    line = line.lower()                                # New Code
    words = line.split()
    for word in words:
        if word not in counts:
            counts[word] = 1
        else:
            counts[word] += 1

print counts
```

O programa se manteve praticamente o mesmo,com a exceção de que usamos `translate` para remover todas as pontuações e `lower` para tornar a linha em caixa baixa. Note que para Python 2.5 e versões anteriores, `translate` não aceita `None` como primeiro parâmetro. Então, use este código para chamar translate:

```
print a.translate(string.maketrans(' ',' '), string.punctuation)
```

Parte de aprender a "Arte do Python" ou "Pensar pythonicamente" está em perceber que Python geralmente tem capacidades embutidas para analisar muitos dados de problemas comuns. No decorrer do tempo, vê-se exemplos de código e documentação suficientes para se saber onde procurar para ver se alguém já escreveu alguma coisa que faça seu trabalho mais fácil.

A seguir esté um versão abreviada da saída:

```
Digite o nome do arquivo: romeo-full.txt
{'swearst': 1, 'all': 6, 'afeard': 1, 'leave': 2, 'these': 2,
'kinsmen': 2, 'what': 11, 'thinkst': 1, 'love': 24, 'cloak': 1,
a': 24, 'orchard': 2, 'light': 5, 'lovers': 2, 'romeo': 40,
'maiden': 1, 'whiteupturned': 1, 'juliet': 32, 'gentleman': 1,
'it': 22, 'leans': 1, 'canst': 1, 'having': 1, ...}
```

Buscar informações nessa saída ainda é difícil e podemos usar Python para nos fornecer exatamente o que estamos procurando; contudo, para tanto, precisamos aprender sobre as **tuplas** do Python. Retornaremos a esse exemplo uma vez que aprendermos sobre tuplas.

9.5 Depuração

Conforme se trabalha com conjuntos de dados maiores, pode ser difícil de depurá-los por exibição e checagem à mão. Eis algumas sugestões para depuração de conjuntos de dados grandes:

Reduza a entrada: Se possível, reduza o tamanho do conjunto de dados. Por exemplo, se o programa lê um arquivo de texto, comece com apenas 10 linhas, ou com o menor exemplo que pode ser construído. Pode-se ainda editar os próprios arquivos, ou (melhor) modificar o programa de tal modo a ler apenas as `n` linhas.

Se houver um erro, pode-se reduzir `n` até o menor valor que manifesta o erro, e, então, aumentá-lo gradualmente conforme se encontra e se corrige os erros.

Verificar sumários e tipos: Ao invés de exibir e verificar o conjunto de dados por completo, considera-se exibir sumarizações dos dados: por exemplo, o número de itens em um dicionário ou o total de uma lista de números.

Valores que não são do tipo correto são uma causa comum de erros de execução. Para depurar esse tipo de erro, geralmente basta exibir o tipo dos valores em questão.

Escreva auto-verificações: Há momentos em que se pode escrever código para verificar erros automaticamente. Por exemplo, se está calculando-se a média de uma lista de números, pode-se verificar se o resultado não é maior que o maior valor na lista nem menor que o menor valor. Isso é chamado de "verificação de sanidade" porque ele detecta resultados que sejam "completamente ilógicos".

Há outro tipo de teste que compara resultados de duas computações diferentes para ver se esses são consistentes. Tal verificação é chamada de "verificação de consistência"

Exiba saídas de maneira aprazível: Formatar a saída da depuração pode fazer com que seja mais fácil de se detectar erros.

Novamente, tempo gasto construindo arcabouços pode reduzir o tempo gasto com depuração.

9.6 Glossário

busca: Uma operação de dicionário que encontra um valor a partir de uma dada chave.

chave: Um objeto que aparece em um dicionário como a primeira parte de um par chave-valor.

dicionário: Um mapeamento entre um conjunto de chaves e seus valores correspondentes.

função de hash: A função usada por uma tabela de hash para calcular a posição de uma chave.

histograma: Um conjunto de contagens.

implementação: Uma maneira de se realizar uma computação.

item: Outro nome para um par chave-valor.

laços aninhados: Quando há um ou mais laços "dentro" de outro laço. O laço interno é executado completamente para cada execução do laço externo.

par chave-valor: A representação de um mapeamento de uma chave a um valor.

tabela de hash: O algoritmo usado para implementar os dicionários de Python.

valor: Um objeto que aparece em um dicionário como a segunda parte em um par chave-valor. Esse é mais específico do que nosso uso anterior da palavra "valor".

9.7 Exercícios

Exercício 9.2 Escreva um programa que categorize cada mensagem de e-mail pelo dia da semana que o commit (`https://pt.wikipedia.org/wiki/Commit`) foi feito. Para tanto, procure por linhas que comecem com "From", então busque pela terceira palavra e mantenha um procedimento de contagem para cada dia da semana. Ao final do programa, exiba o conteúdo do dicionário (ordem não importa).

```
Amostra de linha:
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008

Amostra de execução:
python dow.py
Enter a file name: mbox-short.txt
{'Fri': 20, 'Thu': 6, 'Sat': 1}
```

Exercício 9.3 Escreva um programa que leia um log (`https://pt.wikipedia.org/wiki/Log_de_dados`) de correio eletrônico, escreva um histograma usando um dicionário para contar quantas mensagens vieram de cada endereço de e-mail e, por fim, exiba o dicionário.

```
Enter file name: mbox-short.txt
{'gopal.ramasammycook@gmail.com': 1, 'louis@media.berkeley.edu': 3,
'cwen@iupui.edu': 5, 'antranig@caret.cam.ac.uk': 1,
'rjlowe@iupui.edu': 2, 'gsilver@umich.edu': 3,
```

```
'david.horwitz@uct.ac.za': 4, 'wagnermr@iupui.edu': 1,
'zqian@umich.edu': 4, 'stephen.marquard@uct.ac.za': 2,
'ray@media.berkeley.edu': 1}
```

Exercício 9.4 Insira código no programa acima para descobrir quem tem mais mensagens no arquivo.

Após todos os dados terem sido lidos e o dicionário criado, percorra o dicionário usando um laço de máximo (veja Sessão 5.7.2) para encontrar quem tem mais mensagens e exiba quantas mensagens existem para essa pessoa.

```
Enter a file name: mbox-short.txt
cwen@iupui.edu 5

Enter a file name: mbox.txt
zqian@umich.edu 195
```

Exercício 9.5 Este programa leva em consideração o nome do domínio (ao invés do endereço) de onde a mensagem foi mandada e não de quem essa veio (isto é, o endereço de e-mail inteiro). Ao final do programa, exiba o conteúdo do dicionário.

```
python schoolcount.py
Enter a file name: mbox-short.txt
{'media.berkeley.edu': 4, 'uct.ac.za': 6, 'umich.edu': 7,
'gmail.com': 1, 'caret.cam.ac.uk': 1, 'iupui.edu': 8}
```

Capítulo 10

Expressões regulares

Até agora, temos percorrido os arquivos procurando por padrões e extraindo pedaços de linhas que achamos interessantes. Temos usado métodos string como `split` e `find` e usamos listas e fatiamento de strings para extrair partes das linhas.

Esta tarefa de busca e extração é tão comum que Python tem uma biblioteca muito poderosa chamada **expressões regulares** que lida com muitas destas tarefas de forma muito elegante. O motivo de não introduzirmos expressões regulares antes no livro é que enquanto elas são muito poderosas, também são um pouco complicadas e leva algum tempo para se acostumar com sua sintaxe.

Expressões regulares são quase uma própria pequena linguagem de programação para pesquisa e análise de strings. Na verdade, livros inteiros foram escritos sobre o tema expressões regulares. Neste capítulo, cobriremos apenas noções básicas do assunto. Para mais detalhes sobre regulares expressões, veja:

`http://en.wikipedia.org/wiki/Regular_expression`

`https://docs.python.org/2/library/re.html`

A biblioteca expressão regular `re` deve ser importada para o seu programa antes que você possa usá-la. O uso mais simples da biblioteca de expressão regular é a função `search()`. O programa a seguir demonstra um uso trivial da função search.

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    if re.search('From:', line) :
        print line
```

Abrimos o arquivo, iteramos linha por linha, e usamos a expressão regular `search()` para imprimir apenas as linhas que contém a string “From:”. Este programa não usa o real poder das expressões regulares, uma vez que poderíamos simplesmente usar `line.find()` para obter o mesmo resultado.

O poder das expressões regulares surge quando adicionamos caracteres especiais à string de busca, isso nos permite controlar com mais precisão quais as linhas que casam com a string. Adicionar estes caracteres especiais em nossa expressão regular nos permite um casamento e extração sofisticada com pouco código.

Por exemplo, o acento circunflexo é usado em expressões regulares para identificar "o início" de uma linha. Nós poderíamos mudar nosso programa para casar apenas linhas em que "From:" estivesse no início como a seguir:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    if re.search('^From:', line) :
        print line
```

Agora vamos casar apenas as linhas que *começam com* a string "From:". Esse é um exemplo simples no qual poderíamos ter feito equivalente com o método `startswith()` da biblioteca string. Mas serve para introduzir a noção de que expressões regulares contém caracteres de ação especiais que nos dão mais controle sobre o que irá casar com a expressão regular.

10.1 Casamento de caractere em expressões regulares

Existe um grande número de caracteres especiais que nos permitem escrever expressões regulares ainda mais poderosas. O caractere especial mais utilizado é o ponto, que casa com qualquer caracter.

No exemplo a seguir, a expressão regular "F..m:" casaria com qualquer uma das strings "From:", "Fxxm:", "F12m:", ou "F!@m:" porque o caractere ponto casa com qualquer caractere em um expressão regular.

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    if re.search('^F..m:', line) :
        print line
```

Isso é particularmente poderoso quando combinado à habilidade de indicar que um caractere pode ser repetido algumas vezes utilizando os caracteres "*" ou "+" em suas expressões regulares. Esses caracteres especiais significam que ao invés de casar com um único caractere na string de busca, eles casam com zero-ou-mais caracteres (no caso do asterisco) ou um-ou-mais caracteres (no caso do sinal de adição).

Podemos ainda diminuir as linhas que casam utilizando um caractere repetido **curinga** no exemplo seguinte:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    if re.search('^From:.+@', line) :
        print line
```

A string de busca "^From:.+@" casará com sucesso as linhas que comecem com "From:", seguidas por um ou mais caracteres (".+"), seguidas por uma arroba. Então casará a seguinte linha:

From: stephen.marquard @uct.ac.za

Você pode pensar no curinga ".+" como uma expansão para casar todos os caracteres entre os dois pontos e o arroba.

From: .+ @

É bom pensar no sinal de adição e no asterisco como "insistentes". Por exemplo, a string a seguir casaria o último arroba na string com o ".+", como mostrado abaixo:

From: stephen.marquard@uct.ac.za, csev@umich.edu, and cwen @iupui.edu

É possível dizer ao asterisco ou ao sinal de adição para não serem tão "ganancioso" adicionando outro caracter. Veja a documentação detalhada para mais informações sobre como desligar o comportamento ganancioso.

10.2 Extraindo dados com expressões regulares

Se quisermos extrair dados de uma string em Python podemos usar o método `findall()` para extrair tudo das substrings que casam com a expressão regular. Vamos usar o exemplo de querer extrair qualquer coisa que se pareça com um endereço de email a partir de qualquer linha, independentemente do formato. Por exemplo, queremos pegar os endereços de email de cada uma das seguintes linhas:

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
Return-Path: <postmaster@collab.sakaiproject.org>
          for <source@collab.sakaiproject.org>;
Received: (from apache@localhost)
Author: stephen.marquard@uct.ac.za
```

Não queremos escrever código para cada tipo de linha, dividindo e fatiando diferentemente cada linha. O programa seguinte usa `findall()` para encontrar as linhas com endereço de e-mail e extrair um ou mais endereços de cada uma dessas linhas.

```
import re
s = 'Hello from csev@umich.edu to cwen@iupui.edu about the meeting @2PM'
lst = re.findall('\S+@\S+', s)
print lst
```

O método `findall()` procura a string no segundo argumento e retorna uma lista de todas as strings que se parecem com um endereço de e-mail. Estamos usando uma sequência de dois caracteres que casam com um caractere sem espaço em branco (\S).

A saída do programa seria:

```
['csev@umich.edu', 'cwen@iupui.edu']
```

Traduzindo a expressão regular, estamos procurando por substrings que tenham ao menos um caractere sem espaço em branco, seguido de um arroba, seguido de ao menos mais um caractere sem espaço em branco. A instrução "\S+" casa com o máximo de caracteres sem espaço em branco possíveis.

A expressão regular casaria duas vezes (csev@umich.edu e cwen@iupui.edu), mas não casaria com a string "@2PM" porque não há caracteres sem espaço em branco *antes* do arroba. Podemos usar essa expressão regular em um programa para ler todas as linhas de um arquivo e imprimir qualquer coisa que se pareça com um endereço de email como a seguir:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    x = re.findall('\S+@\S+', line)
    if len(x) > 0 :
        print x
```

Nós lemos cada linha e então extraímos todas as substrings que casam com nossa expressão regular. Uma vez que `findall()` retorna uma lista, nós simplesmente checamos se o número de elementos na lista retornada é maior que zero para imprimir apenas as linhas onde encontramos ao menos uma substring que se pareça com um endereço de email.

Se rodarmos o programa em `findall()` teremos a seguinte saída:

```
['wagnermr@iupui.edu']
['cwen@iupui.edu']
['<postmaster@collab.sakaiproject.org>']
['<200801032122.m03LMFo4005148@nakamura.uits.iupui.edu>']
['<source@collab.sakaiproject.org>;']
['<source@collab.sakaiproject.org>;']
['<source@collab.sakaiproject.org>;']
['apache@localhost)']
['source@collab.sakaiproject.org;']
```

Alguns de nossos endereços de email tem caracteres incorretos como "<" ou ";" no começo ou no fim. Vamos declarar que estamos interessados apenas no pedaço da string que começa e termina com uma letra ou um número.

Para fazer isso, nós usamos outra funcionalidade das expressões regulares. Colchetes são usados para indicar um conjunto de vários caracteres aceitáveis que

estamos dispostos a considerar. Num certo sentido, o "\S" está pedindo para casar o conjunto de caracteres sem espaço em branco. Agora seremos um pouco mais explícitos em termos de caracteres que vamos casar.

Aqui está nossa nova expressão regular:

```
[a-zA-Z0-9]\S*@\S*[a-zA-Z]
```

Isso está ficando um pouco complicado e você pode começar a ver por que expressão regular é uma pequena linguagem em si mesma. Traduzindo esta expressão regular, estamos procurando por substrings que começam com uma *única* letra minúscula, letra maiúscula, ou número "[a-zA-Z0-9]", seguida por zero ou mais caracteres sem espaço em branco ("\S*"), seguida por um arroba, seguida por zero ou mais caracteres sem espaço em branco ("\S*"), seguida por uma letra maiúscula ou minúscula. Note que mudamos de "+" para "*" para indicar zero ou mais caracteres sem espaço em branco uma vez que "[a-zA-Z0-9]" já é um caractere sem espaço em branco. Lembre-se que o "*" ou "+" se aplica ao caractere imediatamente à esquerda do sinal de adição ou do asterisco.

Se usarmos essa expressão em nosso programa, nossos dados serão muito mais limpos:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    x = re.findall('[a-zA-Z0-9]\S*@\S*[a-zA-Z]', line)
    if len(x) > 0 :
        print x

...
['wagnermr@iupui.edu']
['cwen@iupui.edu']
['postmaster@collab.sakaiproject.org']
['200801032122.m03LMFo4005148@nakamura.uits.iupui.edu']
['source@collab.sakaiproject.org']
['source@collab.sakaiproject.org']
['source@collab.sakaiproject.org']
['apache@localhost']
```

Observe que na linha "source@collab.sakaiproject.org", nossa expressão regular eliminou duas letras no fim da string (">;"). Isso ocorre porque quando nós adicionamos "[a-zA-Z]" ao final de nossa expressão regular, nós estamos demandando que qualquer string que o analisador de expressão regular encontre precisa terminar com uma letra. Assim quando se vê ">" depois de "sakaiproject.org>;" ele simplesmente para na última letra que encontrou e casou. (i.e., o "g" foi o último que casou).

Note também que a saída do programa é uma lista do Python que tem uma string como único elemento da lista.

10.3 Combinando busca e extração

Se quisermos encontrar números em linhas que comecem com a string "X-" como:

```
X-DSPAM-Confidence: 0.8475
X-DSPAM-Probability: 0.0000
```

Nós não queremos simplesmente os números de ponto flutuante de quaisquer linhas. Nós queremos apenas extrair números de linhas que tenham a sintaxe acima.

Nós podemos construir a seguinte expressão regular para selecionar as linhas:

```
^X-.*: [0-9.]+
```

Traduzindo isso, estamos dizendo que queremos linhas que comecem com "X-", seguido de zero ou mais caracteres (".*"), seguido de dois pontos (":") e, em seguida, um espaço. Depois do espaço, estamos buscando por um ou mais caracteres que são ou um dígito (0-9) ou um ponto "[0-9.]+". Repare que dentro dos colchetes, o ponto corresponde a um ponto real (i.e., não é um curinga dentro dos colchetes).

Essa é uma expressão regular muito justa que casará apenas as linhas em que estamos interessados como a seguir:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    if re.search('^X\S*: [0-9.]+', line) :
        print line
```

Quando rodamos o programa, vemos os dados muito bem filtrados exibindo apenas as linhas que estamos buscando.

```
X-DSPAM-Confidence: 0.8475
X-DSPAM-Probability: 0.0000
X-DSPAM-Confidence: 0.6178
X-DSPAM-Probability: 0.0000
```

Mas agora temos que resolver o problema da extração de números. Enquanto isso seria simples o suficiente usando `split`, nós podemos usar outra funcionalidade de expressões regulares tanto para buscar quanto para analisar as linhas ao mesmo tempo.

Parênteses são outro caractere especial em expressões regulares. Quando você adiciona parênteses a uma expressão regular, eles são ignorados quando encontram a string. Mas quando você está usando `findall()`, os parênteses indicam que enquanto você quer que a expressão inteira case, você apenas está interessado em extrair o pedaço da substring que case com a expressão regular.

Então, faremos a seguinte alteração em nosso programa:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    x = re.findall('^X\S*: ([0-9.]+)', line)
    if len(x) > 0 :
        print x
```

Em vez de chamar `search()`, podemos adicionar parênteses ao redor da parte da expressão regular que representa o número de ponto flutuante para indicar que só desejamos que `findall()` nos devolva o pedaço de número de ponto flutuante da string correspondente.

A saída desse programa é a seguinte:

```
['0.8475']
['0.0000']
['0.6178']
['0.0000']
['0.6961']
['0.0000']
..
```

Os números ainda estão em uma lista e precisam ser convertidos de strings para ponto flutuante, mas temos usado o poder das expressões regulares tanto para buscar quanto para extrair as informações interessantes que encontramos.

Como outro exemplo dessa técnica, se você olhar para o arquivo há uma série de linhas da seguinte forma:

```
Details: http://source.sakaiproject.org/viewsvn/?view=rev&rev=39772
```

Se quisermos extrair todos os números de revisão (o número inteiro no fim destas linhas) utilizando a mesma técnica vista acima, podemos escrever o programa seguinte:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    x = re.findall('^Details:.*rev=([0-9]+)', line)
    if len(x) > 0:
        print x
```

Traduzindo nossa expressão regular, estamos a procura de linhas que comecem com “Details:”, seguido por algum número de caracteres (“.*”), seguido por “rev=”, e então por um ou mais dígitos. Queremos encontrar linhas que casem com a expressão regular inteira, mas queremos extrair apenas o número inteiro no fim da linha, então colocamos o “[0-9]+” entre parênteses.

Quando rodamos o programa, obtemos a seguinte saída:

```
['39772']
['39771']
['39770']
['39769']
...
```

Lembre-se que o "[0-9]+" é "ganancioso" e tenta criar uma string de dígitos tão grande quanto possível antes de extrair esses dígitos. Esse comportamento "gananciosо" é por que pegamos todos os cinco dígitos de cada número. A biblioteca de expressão regular se expande em ambos os sentidos até encontrar um não-dígito, ou no início ou no fim de uma linha.

Agora podemos utilizar expressões regulares para refazer um exercício do início do livro no qual estávamos interessados na hora do dia em que uma mensagem foi enviada. Olhamos as linhas da seguinte forma:

```
From stephen.marquard@uct.ac.za Sat Jan  5 09:14:16 2008
```

e queremos extrair a hora do dia para cada linha. Anteriormente fizemos isso com duas chamadas de `split`. Primeiro a linha foi dividida em palavras e, depois, tiramos a quinta palavra e dividimos novamente nos dois pontos para retirar os dois caracteres em que estavamos interessados.

Embora tenha funcionado, o código é realmente muito frágil, pois assume que as linhas serão bem formatadas. Se você acrescentasse uma checagem de erros suficiente (ou um grande bloco try/except) para garantir que seu programa nunca falhará quando receber linhas formatadas incorretamente, o programa aumentaria para 10-15 linhas de código, o que seria muito difícil de ler.

Podemos fazer isso de um modo muito mais simples com a expressão regular a seguir:

```
^From .* [0-9][0-9]:
```

A tradução dessa expressão regular é que estamos procurando por linhas que comecem com "From " (note o espaço), seguido por algum número de caracteres (".*"), seguido por um espaço, seguido por dois dígitos "[0-9][0-9]", seguido por dois pontos. Essa é a definição do tipo de linhas que estamos procurando.

A fim de retirar somente a hora utilizando `findall()`, colocamos os dois dígitos entre parênteses como segue:

```
^From .* ([0-9][0-9]):
```

Isso resulta no seguinte programa:

```
import re
hand = open('mbox-short.txt')
for line in hand:
    line = line.rstrip()
    x = re.findall('^From .* ([0-9][0-9]):', line)
    if len(x) > 0 : print x
```

Quando o programa roda, produz a seguinte saída:

```
['09']
['18']
['16']
['15']
...
```

10.4 Caractere de escape

Uma vez que usamos caracteres especiais em expressões regulares para casar o começo ou o fim de uma linha ou especificar curingas, precisamos de uma maneira de indicar que esses caracteres são "normais" e queremos casar o caractere real como um sinal de dólar ou acento circunflexo.

Podemos indicar que queremos simplesmente casar um caractere prefixando o caractere com uma barra invertida. Por exemplo, podemos encontrar quantidades de dinheiro com a seguinte expressão regular.

```
import re
x = 'We just received $10.00 for cookies.'
y = re.findall('\$[0-9.]+',x)
```

Já que prefixamos o caractere dólar com uma barra invertida, ele realmente casará com o dólar na string de entrada ao invés de casar com o "fim da linha", e o resto da expressão regular casa um ou mais dígitos ou o caractere ponto. *Note:* Dentro de colchetes, caracteres não são "especiais". Então quando dizemos "[0-9.]", realmente significa dígitos ou um ponto. Fora dos colchetes, um ponto é o caractere "curinga" para casar qualquer caractere. Dentro dos colchetes, o ponto é um ponto.

10.5 Resumo

Embora isso tenha sido apenas uma visão superficial de expressões regulares, conseguimos aprender um pouco sobre a linguagem de expressões regulares. Elas são strings de busca com caracteres especiais que informam o que você quer ao sistema de expressão regular, que por sua vez define o "casamento" e o que é extraído das strings que casaram. Aqui temos alguns desses caracteres especiais e sequências de caracteres:

^
Corresponde ao início da linha.

$
Corresponde ao final da linha.

.
Corresponde a qualquer caractere (um curinga).

\s
Corresponde a um espaço em branco.

\S
Corresponde a um caractere sem espaço em branco (oposto do \s).

*
Aplica-se ao caractere imediatamente anterior e corresponde a zero ou mais do(s) caractere(s) anterior(es).

*?
Aplica-se ao caractere imediatamente anterior e corresponde a zero ou mais do(s) caractere(s) anterior(es) em "modo não ganancioso".

+
Aplica-se ao caractere imediatamente anterior e corresponde a um ou mais do(s) caractere(s) anterior(es).

+?
Aplica-se ao caractere imediatamente anterior e corresponde a um ou mais do(s) caractere(s) anterior(es) em "modo não ganancioso".

[aeiou]
Corresponde a um único caractere contanto que esteja no conjunto especificado. Nesse exemplo, corresponderia a "a", "e", "i", "o", ou "u", mas a nenhum outro caractere.

[a-z0-9]
Você pode especificar intervalos de caracteres usando o sinal de subtração. Esse exemplo é um único caractere que deve ser uma letra minúscula ou um dígito.

[^A-Za-z]
Quando o primeiro caractere é um acento circunflexo, ele inverte a lógica. Esse exemplo, corresponde um único caractere que é qualquer coisa *exceto* uma letra maiúscula ou minúscula.

( )
Quando adicionamos parênteses a uma expressão regular, eles são ignorados para efeito de correspondência, mas permite extrair um subconjunto específico da string correspondente ao invés de toda string quando usamos `findall()`.

\b
Corresponde a uma string vazia, mas somente no começo ou no final de uma palavra.

\B
Corresponde a uma string vazia, mas não no começo ou no final de uma palavra.

\d
Corresponde a qualquer dígito decimal; equivalente ao conjunto [0-9].

\D
Corresponde a qualquer caractere não-dígito; equivalente ao conjunto [^0-9].

10.6 Seção bônus para usuários de Unix

O suporte para pesquisar arquivos utilizando expressões regulares está presente no sistema operacional Unix desde 1960 e está disponível em quase todas as linguagens de programação de uma forma ou de outra.

Na realidade, é um programa de linha de comando integrado ao Unix chamado **grep** (Generalized Regular Expression Parser - Analisador generalizado de expressões regulares) que faz o mesmo que os exemplos `search()` nesse capítulo. Então, se você tem um sistema Macintosh ou Linux pode tentar os seguintes comandos em sua janela de linha de comando.

```
$ grep '^From:' mbox-short.txt
From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
```

Isso diz ao `grep` para mostrar as linhas que comecem com a string "From:" no arquivo `mbox-short.txt`. Se você experimentar um pouco o comando `grep` e ler sua documentação encontrará algumas pequenas diferenças entre as expressões regulares suportadas em Python e as expressões regulares suportadas pelo `grep`. Por exemplo, `grep` não suporta o caractere sem espaço em branco "\S" então você precisará usar uma notação um pouco mais complexa "[^]", que simplesmente significa que corresponde a um caractere que é qualquer coisa a não ser um espaço.

10.7 Depuração

Python possui uma documentação embutida simples e rudimentar que pode ser bastante útil se você precisar de uma ajuda para se lembrar do nome exato de um método em particular. Essa documentação pode ser vista em seu interpretador Python no modo interativo.

Você pode acionar um sistema de ajuda interativa usando `help()`.

```
>>> help()

Welcome to Python 2.6!  This is the online help utility.

If this is your first time using Python, you should definitely check out
the tutorial on the Internet at http://docs.python.org/tutorial/.

Enter the name of any module, keyword, or topic to get help on writing
Python programs and using Python modules.  To quit this help utility and
return to the interpreter, just type "quit".
```

```
To get a list of available modules, keywords, or topics, type "modules",
"keywords", or "topics".  Each module also comes with a one-line summary
of what it does; to list the modules whose summaries contain a given word
such as "spam", type "modules spam".

help> modules
```

Se você sabe que módulo deseja usar, pode usar o comando `dir()` para encontrar os métodos no módulo da seguinte forma:

```
>>> import re
>>> dir(re)
[.. 'compile', 'copy_reg', 'error', 'escape', 'findall',
'finditer', 'match', 'purge', 'search', 'split', 'sre_compile',
'sre_parse', 'sub', 'subn', 'sys', 'template']
```

Você também pode pegar um pequeno pedaço de documentação de um método usando o comando dir.

```
>>> help (re.search)
Help on function search in module re:

search(pattern, string, flags=0)
    Scan through string looking for a match to the pattern, returning
    a match object, or None if no match was found.
>>>
```

A documentação embutida não é muito extensa, mas pode ser muito útil quando você está com pressa e não tem acesso a um navegador web ou um site de buscas.

10.8 Glossário

código frágil: Código que funciona quando a entrada de dados está em um formato específico mas é propenso a quebrar se houver algum desvio em relação ao formato correto. Chamamos isso de "código frágil" porque é fácil de quebrar.

casamento ganancioso: A ideia de que os caracteres "+" and "*" em uma expressão regular se expandem para casar a maior string possível.

grep: Um comando disponível na maioria dos sistemas Unix que busca através de arquivos de texto à procura de linhas que casam com expressões regulares. O nome do comando significa "Generalized Regular Expression Parser"(Analisador Generalizado de Expressões Regulares).

expressão regular: Uma linguagem para expressar strings de pesquisa mais complexas. Uma expressão regular pode conter caracteres especiais que indicam que a busca somente corresponderá no início ou no final de uma linha ou muitas outras capacidades semelhantes.

curinga: Um caractere especial que busca por qualquer caractere. Em expressões regulares, o caractere wildcard é o "ponto"

10.9 Exercícios

Exercício 10.1 Escreva um programa simples para simular a operação do comando `grep` no Unix. Peça para o usuário entrar com uma expressão regular e conte o número de linhas que casam com a expressão regular:

```
$ python grep.py
Enter a regular expression: ^Author
mbox.txt had 1798 lines that matched ^Author

$ python grep.py
Enter a regular expression: ^X-
mbox.txt had 14368 lines that matched ^X-

$ python grep.py
Enter a regular expression: java$
mbox.txt had 4218 lines that matched java$
```

Exercício 10.2 Escreva um programa que procure as linhas do formulário

```
Nova Revisão: 39772
```

e extraia o número de cada uma das linhas usando expressão regular e o método `findall()`. Calcule e exiba a média dos números.

```
Enter file:mbox.txt
38549.7949721

Enter file:mbox-short.txt
39756.9259259
```

Capítulo 11

Programas em redes

Enquanto muitos dos exemplos usados neste livro tem focado na leitura de arquivos e procura por dados neles, existem muitas fontes de informação diferentes quando se leva em conta a Internet.

Nesse capítulo fingiremos ser um navegador web a obter páginas web usando o Protocolo de Transferência de Hipertexto (*HyperText Transport Protocol* – HTTP). Feito isso, faremos uma leitura por dados da página web e os analisaremos.

11.1 Protocolo de Transferência de Hipertexto - HTTP

O protocolo de rede que impulsiona a web é, na verdade, bem simples e existe um suporte embutido no Python chamado `sockets` que faz com seja muito fácil estabelecer conexões de rede e obter dados através desses sockets com um programa Python.

Um **socket** é bem parecido com um arquivo, exceto que um único socket provê uma conexão de duas vias entre dois programas. Você pode tanto ler quanto escrever pelo mesmo socket. Se você escrever alguma coisa em um socket, a escrita é enviada para a aplicação na outra ponta do socket. Se você ler a partir de um socket, você está recebendo os dados que a outra aplicação enviou.

Mas se você tentar ler um socket enquanto o programa na outra ponta do socket não enviar nenhum dado—você tem que sentar e esperar. Se os programas de ambos os lados do socket simplesmente esperarem por dados sem enviarem nada, eles continuarão esperando até que alguém envie algum dado..

Então, uma parte importante dos programas que se comunicam através da Internet tem algum tipo de protocolo. Um protocolo é um conjunto de regras precisas que determinam quem inicia, como será a comunidação, e então quais são as respostas para a mensagem enviada, e quem envia a próxima, e assim por diante. De certa

forma as aplicações, cada uma em uma ponta do socket, estão dançando e tem que garantir que uma não vai pisar no pé da outra.

Existem muitos documentos que descrevem estes protocolos de rede. O Protocolo de Transferência de Hipertexto é descrito no seguinte documento:

```
http://www.w3.org/Protocols/rfc2616/rfc2616.txt
```

Este é um longo e complexo documento de 176 páginas, com muitos detalhes. Se você achá-lo interessante, fique à vontade para lê-lo na integra. Mas se você der uma olhada pela página 36 do RFC2616, irá encontrar a sintaxe para a requisição GET. Para requisitar um documento de um servidor web, faremos uma conexão com o servidor `www.py4inf.com` na porta 80, e então enviamos uma linha com o seguinte formato:

```
GET http://www.py4inf.com/code/romeo.txt HTTP/1.0
```

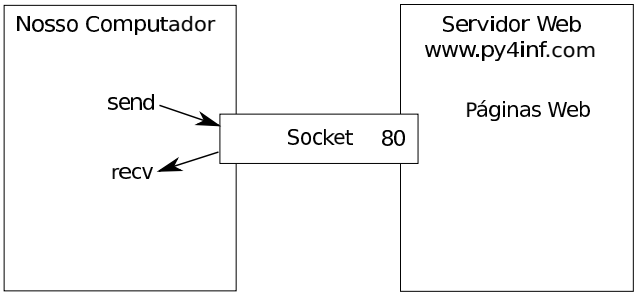
onde o segundo parâmetro é a página web que estamos solicitando, e então enviamos também uma linha em branco. O servidor web irá responder com algumas informações de cabeçalho sobre o documento e uma linha em branco seguida do conteúdo do documento.

11.2 O Navegador Web Mais Simples do Mundo

Talvez, a maneira mais simples de mostrar como o protocolo HTTP funciona é escrever um programa Python bem simples que faz a conexão com um servidor web e segue as regras do protocolo HTTP para solicitar um documento e exibir o que o servidor envia de volta.

```
import socket

mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('www.py4inf.com', 80))
mysock.send('GET http://www.py4inf.com/code/romeo.txt HTTP/1.0\n\n')

while True:
    data = mysock.recv(512)
    if ( len(data) < 1 ) :
        break
    print data

mysock.close()
```

Primeiro, o programa estabelece a conexão na porta 80 do servidor `www.py4inf.com`. Como nosso programa faz o papel de um "navegador web", o protocolo HTTP informa que nós temos que enviar um comando GET seguido de uma linha em branco.



Uma vez enviada a linha em branco, escrevemos um loop que recebe do socket, dados em pedaços de 512 caracteres e imprime os dados até que não exista mais dados para ler (por exemplo, a recv() retorna uma string vazia).

O programa produz a seguinte saída:

```
HTTP/1.1 200 OK
Date: Sun, 14 Mar 2010 23:52:41 GMT
Server: Apache
Last-Modified: Tue, 29 Dec 2009 01:31:22 GMT
ETag: "143c1b33-a7-4b395bea"
Accept-Ranges: bytes
Content-Length: 167
Connection: close
Content-Type: text/plain

But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

A saída começa com os cabeçalhos que o servidor web envia para descrever o documento. Por exemplo, o cabeçalho `Content-Type` indica que o documento é um documento em texto plano (`text/plain`).

Depois que o servidor nos enviar os cabeçalhos, ele adiciona uma linha em branco para indicar o final dos cabeçalhos, e então, envia realmente os dados do arquivo `romeo.txt`.

Esse exemplo mostra como fazer uma conexão de rede de baixo nível com sockets. Sockets podem ser usados para se comunicar com um servidor web ou com um servidor de e-mail ou muitos outros tipos de servidores. Tudo que é preciso é encontrar o documento que descreve o protocolo e escrever o código para enviar e receber os dados de acordo com o protocolo.

Contudo, como o protocolo que nós usamos mais comumente é o protocolo web HTTP, o Python tem uma biblioteca especificamente desenvolvida para ter suporte ao protocolo HTTP. E assim, obter documentos e dados através da web.

11.3 Obtendo uma imagem através do HTTP

No exemplo acima, nós pegamos um arquivo em texto plano que tinha novas linhas dentro do arquivo e nós simplesmente copiamos os dados para a tela a medida que o programa era executado. Nós podemos usar um programa similar para obter uma imagem através da web usando o HTTP. Ao invés de copiar os dados para a tela, a medida que o programa é executado, nós acumulamos os dados em uma string, retiramos os cabeçalhos, e então salvamos os dados da imagem em um arquivo. Como a seguir:

```
import socket
import time

mysock = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
mysock.connect(('www.py4inf.com', 80))
mysock.send('GET http://www.py4inf.com/cover.jpg HTTP/1.0\n\n')


count = 0
picture = "";
while True:
    data = mysock.recv(5120)
    if ( len(data) < 1 ) : break
    # time.sleep(0.25)
    count = count + len(data)
    print len(data),count
    picture = picture + data

mysock.close()

# Look for the end of the header (2 CRLF)
pos = picture.find("\r\n\r\n");
print 'Header length',pos
print picture[:pos]

# Skip past the header and save the picture data
picture = picture[pos+4:]
fhand = open("stuff.jpg","wb")
fhand.write(picture);
fhand.close()
```

Quando o programa é executado, ele produz a seguinte saída:

```
$ python urljpeg.py
2920 2920
1460 4380
1460 5840
1460 7300
...
1460 62780
1460 64240
2920 67160
1460 68620
1681 70301
```

```
Header length 240
HTTP/1.1 200 OK
Date: Sat, 02 Nov 2013 02:15:07 GMT
Server: Apache
Last-Modified: Sat, 02 Nov 2013 02:01:26 GMT
ETag: "19c141-111a9-4ea280f8354b8"
Accept-Ranges: bytes
Content-Length: 70057
Connection: close
Content-Type: image/jpeg
```

Você pode ver que para esta url, o cabeçalho `Content-Type` indica que o corpo do documento é uma imagem (`image/jpeg`). Uma vez terminado o programa, você pode ver os dados da imagem abrindo o arquivo `stuff.jpg` com um visualizador de imagens.

Durante a execução do programa, você pode ver que não temos 5120 caracteres para cada vez que chamamos o método `recv()`. Nós pegamos tantos caracteres quantos foram transferidos através da rede, do servidor web para nós, no momento que chamamos `recv()`. Neste exemplo, pegamos 1460 ou 2920 caracteres a cada vez que requisitamos até chegar a 5120 caracteres de dados.

Os seus resultados podem ser diferentes, dependendo da velocidade de sua rede. Note também que na última chamada de `recv()`, nós pegamos 1681 bytes, que é o final do fluxo (stream), e na chamada seguinte da `recv()` nós recebemos uma string vazia (zero-length). Que nos informa que o servidor chamou `close()` no seu final de socket e não existe mais dados para enviar.

Nós podemos reduzir nossas sucessivas chamadas a `recv()` descomentando , removendo o caractere cerquilha, da chamada de `time.sleep()`. Desta forma, nós esperamos um quarto de segundo depois de cada chamada, e assim, o servidor pode "se antecipar" à nós e enviar mais dados antes de nós chamarmos `recv()` novamente. Com esse "atraso", o programa é executado como a seguir:

```
$ python urljpeg.py
1460 1460
5120 6580
5120 11700
...
5120 62900
5120 68020
2281 70301
Header length 240
HTTP/1.1 200 OK
Date: Sat, 02 Nov 2013 02:22:04 GMT
Server: Apache
Last-Modified: Sat, 02 Nov 2013 02:01:26 GMT
ETag: "19c141-111a9-4ea280f8354b8"
```

```
Accept-Ranges: bytes
Content-Length: 70057
Connection: close
Content-Type: image/jpeg
```

Agora, ao invés de uma primeira e última chamada a `recv()`, nós agora pegamos 5120 caracteres a cada vez que pedimos novos dados.

Existe um buffer entre o servidor, fazendo solicitações `send()` e nossa aplicação fazendo solicitações `recv()`. Quando nós executamos o programa com o "atraso"estando ativo, em algum momento o servidor preenche o buffer no socket e é forçado a fazer uma pausa até que nosso programa comece a esvaziar o buffer. A pausa, tanto do envio quanto do recebimento da aplicação, é chamada "flow control" (controle de fluxo).

11.4 Obtendo páginas web com `urllib`

Embora nós possamos manualmente enviar e receber dados pelo HTTP usando a biblioteca socket, existe uma maneira muito mais simples de realizar essa tarefa comum em Python pelo uso da biblioteca `urllib`.

Usando a `urllib`, você pode tratar uma página web de maneira muito parecida a um arquivo. Você simplesmente indica qual página web você gostaria de obter e a `urllib` lida com todo o protocolo HTTP e detalhes sobre cabeçalhos.

O código equivalente para ler o arquivo `romeo.txt` a partir da web usando a `urllib` é como o seguinte:

```
import urllib

fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')
for line in fhand:
   print line.strip()
```

Uma vez que a página web tenha sido aberta com `urllib.urlopen`, nós podemos tratá-la como um arquivo e fazer a leitura usando um loop `for`.

Quando o programa é executado, nós apenas vemos na saída o conteúdo do arquivo. Os cabeçalhos continuam sendo enviados, mas o código da `urllib` consome os cabeçalhos e apenas retorna os dados para nós.

```
But soft what light through yonder window breaks
It is the east and Juliet is the sun
Arise fair sun and kill the envious moon
Who is already sick and pale with grief
```

Como exemplo, nós podemos escrever um programa para obter os dados de `romeo.txt` e calcular a frequência de cada palavra existente dentro do arquivo, como a seguir:

```
import urllib

counts = dict()
fhand = urllib.urlopen('http://www.py4inf.com/code/romeo.txt')
for line in fhand:
    words = line.split()
    for word in words:
        counts[word] = counts.get(word,0) + 1
print counts
```

Novamente, uma vez que nós abrimos a página web, podemos fazer a leitura como um arquivo local.

11.5 Analizando o HTML e varrendo a web

Um dos usos comuns da capacidade da `urllib` em Python é **varrer** a web. Varrer a web é quando nós escrevemos um programa que finge ser um navegador web e obtêm páginas, e então examina os dados nessas páginas a procura de padrões.

Como um exemplo, um mecanismo de busca como o Google irá olhar os fontes de uma página web e extrair os links para outras páginas e obter essas páginas, extrair os links para outras páginas e obter essas páginas, extrair links e assim por diante. Usando essa técnica, o Google **mapeia** seu caminho através de quase todas as páginas na web.

O Google também usa a frequência de links das páginas que ele encontra para uma página em particular de maneira a medir o quão "importante" uma página é, e em que posição a página deve aparecer em seus resultados de pesquisa.

11.6 Analisando o HTML através do uso de expressões regulares

Uma maneira simples de analisar o HTML é usar expressões regulares, para repetidamente, buscar e extrair substrings que coincidam com um padrão em particular.

Aqui está uma página web simples:

```
<h1>The First Page</h1>
<p>
If you like, you can switch to the
<a href="http://www.dr-chuck.com/page2.htm">
Second Page</a>.
</p>
```

Nós podemos construir uma expressão regular para identificar e extrair os valores dos links do texto abaixo, como a seguir:

```
href="http://.+?"
```

Nossa expressão regular procura por strings que iniciam com “href=”http://”, seguida de um ou mais caracteres (“.+?”), seguida por outra aspas. O ponto de interrogação adicionado ao “.+?” indica que a expressão é para coincidir com um padrão de forma “não gananciosa”, ao invés de uma maneira “gananciosa”. Um padrão não ganancioso tenta encontrar a *menor* string correspondente possível e a gananciosa tenta encontrar a *maior* string correspondente possível.

Nós adicionamos parênteses a nossa expressão regular para indicar qual parte de nossa string correspondente nós gostaríamos de extrair, e foi produzido o seguinte programa:

```
import urllib
import re

url = raw_input('Enter - ')
html = urllib.urlopen(url).read()
links = re.findall('href="(http://.*?)"', html)
for link in links:
    print link
```

O método de expressão regular `findall` irá retornar para nós uma lista de todas as strings que coincidem com nossa expressão regular, retornando apenas o texto do link entre as aspas duplas.

Quando nós executamos o programa, nós temos a seguinte saída:

```
python urlregex.py
Enter - http://www.dr-chuck.com/page1.htm
http://www.dr-chuck.com/page2.htm

python urlregex.py
Enter - http://www.py4inf.com/book.htm
http://www.greenteapress.com/thinkpython/thinkpython.html
http://allendowney.com/
http://www.py4inf.com/code
http://www.lib.umich.edu/espresso-book-machine
http://www.py4inf.com/py4inf-slides.zip
```

As expressões regulares funcionam muito bem quando o seu HTML está bem formatado e previsível. Mas como existem muitas páginas HTML “quebradas” por aí, a solução usando expressões regulares pode perder alguns links válidos ou terminar com dados ruins.

Isso pode ser resolvido utilizando uma robusta biblioteca de análise de HTML.

11.7 Analisando o HTML com o uso da BeautifulSoup

Existem várias bibliotecas Python que podem ajudar você a analisar o HTML e extrair dados das páginas. Cada uma das bibliotecas tem suas vantagens e desvantagens e você pode escolher uma com base em suas necessidades.

Como exemplo, iremos simplesmente analisar alguma entrada HTML e extrair os links usando a biblioteca **BeautifulSoup**. Você pode baixar e instalar o código BeautifulSoup de:

`http://www.crummy.com/software/`

Você pode baixar e fazer a "instalação" da biblioteca BeautifulSoup ou pode simplesmente colocar o arquivo `BeautifulSoup.py` no mesmo diretório que está a sua aplicação.

Ainda que o HTML se pareça com XML[1] e algumas páginas são cuidadosamente construídas para ser um XML, a maioria do HTML é, geralmente, quebrado. O que faz com que um analisador XML rejeite toda a página HTML por concluir que ela está impropriamente formada. A BeautifulSoup tolera muitas imperfeições HTML e ainda permite que você extraia facilmente os dados que você precisa.

Nós iremos usar a `urllib` para ler a página e então usar a `BeautifulSoup` para extrair os atributos `href` das tags de ancoragem (`a`).

```
import urllib
from BeautifulSoup import *

url = raw_input('Enter - ')
html = urllib.urlopen(url).read()
soup = BeautifulSoup(html)

# Retrieve all of the anchor tags
tags = soup('a')
for tag in tags:
   print tag.get('href', None)
```

O programa pede um endereço web, e então abre a página web, lê os dados e passa os dados para o analisador BeautifulSoup, e então obtém todas as tags de ancoragem e imprime o atributo `href` de cada tag.

Quando o programa é executado, ele se parece como a seguir:

```
python urllinks.py
Enter - http://www.dr-chuck.com/page1.htm
http://www.dr-chuck.com/page2.htm

python urllinks.py
Enter - http://www.py4inf.com/book.htm
http://www.greenteapress.com/thinkpython/thinkpython.html
http://allendowney.com/
http://www.si502.com/
http://www.lib.umich.edu/espresso-book-machine
http://www.py4inf.com/code
http://www.pythonlearn.com/
```

Você pode usar a BeautifulSoup para buscar várias partes de cada tag como a seguir:

[1] O formato XML será descrito no próximo capítulo.

```
import urllib
from BeautifulSoup import *

url = raw_input('Enter - ')
html = urllib.urlopen(url).read()
soup = BeautifulSoup(html)

# Retrieve all of the anchor tags
tags = soup('a')
for tag in tags:
   # Look at the parts of a tag
   print 'TAG:',tag
   print 'URL:',tag.get('href', None)
   print 'Content:',tag.contents[0]
   print 'Attrs:',tag.attrs
```

Isso produz a seguinte saída:

```
python urllink2.py
Enter - http://www.dr-chuck.com/page1.htm
TAG: <a href="http://www.dr-chuck.com/page2.htm">
Second Page</a>
URL: http://www.dr-chuck.com/page2.htm
Content: [u'\nSecond Page']
Attrs: [(u'href', u'http://www.dr-chuck.com/page2.htm')]
```

Estes exemplos apenas começam a mostrar o poder da BeautifulSoup, quando se refere a análise de HTML. Veja a documentação e exemplos em `http://www.crummy.com/software/BeautifulSoup/` para mais detalhes.

11.8 Lendo arquivos binários usando a `urllib`

Algumas vezes, você quer obter um arquivo não texto (ou binário), como um arquivo de imagem ou vídeo. Os dados nesses arquivos geralmente não são úteis para serem impressos, mas você pode facilmente fazer uma cópia da URL para um arquivo local em seu disco rígido usando a `urllib`.

O padrão é abrir a URL e usar `read` para baixar o conteúdo completo do documento para dentro de uma variável string (`img`), e então escrever essa informação em um arquivo local, como a seguir:

```
img = urllib.urlopen('http://www.py4inf.com/cover.jpg').read()
fhand = open('cover.jpg', 'w')
fhand.write(img)
fhand.close()
```

Esse programa lê todos os dados, de uma vez, através da rede e os armazena dentro da variável `img`, na memória de seu, computador. Então abre o arquivo `cover.jpg` e escreve os dados para o seu disco. Isso irá funcionar se o tamanho do arquivo for menor que o tamanho da memória de seu computador.

Contudo, se ele for um arquivo de áudio ou vídeo grande, esse programa pode falhar ou pelo menos rodar de forma extremamente vagarosa quando seu computador ficar sem memória. Para evitar este tipo de problema, nós podemos obter os dados em blocos (ou buffers) e então escrever cada bloco no disco antes de obter o próximo bloco. Desta forma o programa pode ler um arquivo de qualquer tamanho sem usar toda a memória que você tem em seu computador.

```
import urllib

img = urllib.urlopen('http://www.py4inf.com/cover.jpg')
fhand = open('cover.jpg', 'w')
size = 0
while True:
    info = img.read(100000)
    if len(info) < 1 : break
    size = size + len(info)
    fhand.write(info)

print size,'characters copied.'
fhand.close()
```

Neste exemplo, nós lemos apenas 100,000 caracteres por vez e então escrevemos esses caracteres no arquivo `cover.jpg` antes de obter os próximos 100,000 caracteres de dados a partir da web.

Esse programa é executado como a seguir:

```
python curl2.py
568248 characters copied.
```

Se você tem um computador Unix ou Macintosh, você provavelmente tem um comando em seu sistema operacional que executa essa operação, como a seguir:

```
curl -O http://www.py4inf.com/cover.jpg
```

O comando `curl` é a abreviação para "copy URL", então esses dois exemplos são, inteligentemente, chamados de `curl1.py` e `curl2.py` em `www.py4inf.com/code`, já que eles implementam uma funcionalidade similar ao comando `curl`. Existe também um programa exemplo `curl3.py` que realiza essa tarefa de maneira um pouco mais efetiva, caso você queira realmente usar esse padrão em um programa que esteja escrevendo.

11.9 Glossário

BeautifulSoup: Uma biblioteca Python para análise de documentos HTML e extração de dados desses documentos HTML que faz compensações nas maiorias das imperfeições em um HTML que os navegadores geralmente ignoram. Você pode baixar o código da BeautifulSoup em `www.crummy.com`.

porta: Um número que geralmente indica qual aplicação você está contactando quando você faz uma conexão por socket com um servidor. Como um exemplo, o tráfego web, usualmente, usa a porta 80, enquanto o tráfego de e-mail usa a porta 25.

scraping: Quando um programa finge ser um navegador web, obtém uma página web, e então olha o conteúdo da página web. Geralmente os programas estão seguindo os links de uma página para encontrar a próxima página. Para que eles possam atravessar uma rede de páginas ou uma rede social.

socket: Uma conexão de rede entre duas aplicações. Onde as aplicações podem enviar e receber dados em ambas as direções.

spider: Um mecanismo que busca pela web obtendo uma página e então todas as páginas com ligações a partir dessa página e assim por diante até ele ter praticamente todas as páginas na Internet que ele usará para construir sua indexação de busca.

11.10 Exercícios

Exercício 11.1 Altere o programa socket `socket1.py` para pedir ao usuário a URL, e assim, ele possa ler qualquer página web. Você pode usar `split('/')` para quebrar a URL em partes de componentes para que você possa extrair o nome da máquina para a chamada `connect`. Adicionar uma checagem de erro usando `try` e `except` para lidar com a conexão, caso o usuário digite um formato de URL impróprio ou não existente.

Exercício 11.2 Altere seu programa socket para que ele conte o número de caracteres que ele tenha recebido e interrompa a exibição de qualquer texto após ele ter mostrado 3000 caracteres. O programa deve obter o documento por inteiro, contar o número total de caracteres e exibir a contagem do número de caracteres no final do documento.

Exercício 11.3 Use a `urllib` para replicar o exercício prévio de (1) obtenção de um documento a partir de uma URL, (2) exibindo até 3000 caracteres, e (3) desfazendo a contagem do total de caracteres no documento. Não se preocupe com os cabeçalhos neste exercício, apenas mostre os primeiros 3000 caracteres do conteúdo do documento.

Exercício 11.4 Altere o programa `urllinks.py` para que ele extraia e conte as tags parágrafo (p) de um documento HTML obtido e exiba a contagem de parágrafos como saída de seu programa. Não exiba o texto de parágrafo, apenas faça a contagem. Teste seu programa em várias páginas web pequenas. E também em algumas páginas web grandes.

Exercício 11.5 (Avançado) Altere o programa socket para que ele apenas mostre os dados após os cabeçalhos até que uma linha em branco tenha sido recebida.

Lembre que o `recv` está recebendo caracteres (nova linha e outros caracteres), não linhas.

Capítulo 12

Banco de Dados e Structured Query Language (SQL)

12.1 O que é um banco de dados?

Um **banco de dados** é um tipo de arquivo organizado para armazenamento de dados. A maioria dos bancos de dados são orgazanizados como um dicionário, no sentido de que eles realizam o mapeamento por chaves e valores. A grande diferença é que os bancos de dados estão em disco (ou outros dispositivos de armazenamentos permanentes), então eles continuam armazenando os dados mesmo depois que o programa termina. Porque um banco de dados é armazenado de forma permanente, isto permite armazenar muito mais dados que um dicionário, que é limitado ao tamanho da memória no computador.

Como um dicionário, um banco de dados é um software desenvolvido para manter a inserção e acesso aos dados de forma muito rápida, até para grandes volumes de dados. O banco de dados mantém sua performance através da construção de **indices** assim que o dado é adicionado, isto permite ao computador acessar rapidamente uma entrada em particular.

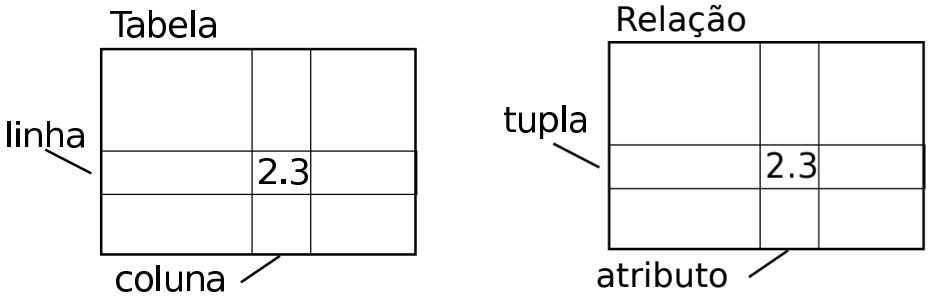
Existem diferentes tipos de sistemas de bancos de dados que são utilizados para diferentes propósitos, alguns destes são: Oracle, MySQL, Microsoft SQL Server, PostgreSQL, e SQLite. Focaremos no uso do SQLite neste livro pois é um banco de dados comum e já está integrado ao Python. O SQLite foi desenvolvido com o propósito de ser *embarcado* em outras aplicações para prover suporte a banco de dados junto à aplicação. Por exemplo, o navegador Firefox utiliza o SQLite internamente, assim como muitos outros produtos.

`http://sqlite.org/`

SQLite é adequado para alguns problemas de manipulação de dados que podemos ver na informática como a aplicação de indexação do Twitter que descrevemos neste capítulo.

12.2 Conceitos de bancos de dados

Quando você olha para um banco de dados pela primeira vez, parece uma planilha (como uma planilha de cálculo do LibreOffice) com múltiplas folhas. A estrutura de dados básica que compõem um banco de dados são: **tabelas**, **linhas**, e **colunas**.



Na descrição técnica de um banco de dados relacional o conceito de tabela, linha e coluna são referências formais para **relação**, **tupla**, e **atributo**, respectivamente. Usaremos os termos menos formais neste capítulo.

12.3 Plugin do Firefox de Gerenciamento do SQLite

O foco deste capítulo é o uso do Python para trabalhar com dados com o SQLite, muitas operações podem ser feitas de forma mais conveniente utilizando um *plugin* do Firefox, o **SQLite Database Manager** que está disponível gratuitamente através do *link*:

```
https://addons.mozilla.org/en-us/firefox/addon/sqlite-manager/
```

Utilizando o navegador você pode facilmente criar tabelas, inserir, editar ou executar consultas SQL nos dados da base de dados.

De certa forma, o gerenciador de banco de dados é similar a um editor de texto quando trabalha com arquivos de texto. Quando você quer fazer uma ou mais operações com um arquivo de texto, você pode simplesmente abrir o arquivo em um editor de texto e fazer as alterações que desejar. Quando você tem muitas alterações para fazer, normalmente você pode escrever um simples programa em Python para executar esta tarefa. Você encontrará os mesmos padrões quando for trabalhar com banco de dados. Você fará operações em um gerenciador de banco de dados e as operações mais complexas serão mais convenientes se forem feitas com Python.

12.4 Criando uma tabela em um banco de dados

Bancos de dados precisam de estruturas mais bem definidas do que listas ou dicionários em Python[1].

Quando criamos uma **tabela** em um banco de dados, precisamos informar ao banco de dados previamente o nome de cada **coluna** na tabela e o tipo de dados que planejamos armazenar em cada **coluna**. Quando o sistema de banco de dados conhece o tipo de dado em cada coluna, ele pode definir a forma mais eficiente de armazenar e consultar o dado baseado no tipo do dado.

Você pode visualizar os diversos tipos de dados que são suportados pelo SQLite através do seguinte endereço:

`http://www.sqlite.org/datatypes.html`

Definir a estrutura dos seus tipos de dados pode parecer inconveniente no começo, mas a recompensa é o acesso rápido aos dados mesmo quando o banco de dados contém um grande número de informações.

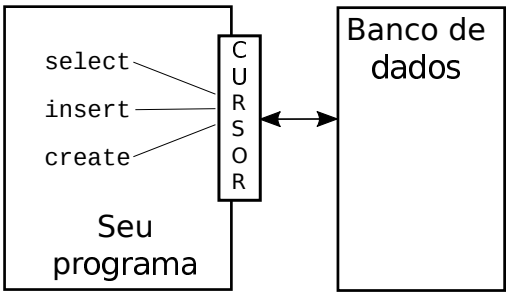
O seguinte código cria um arquivo de banco de dados com uma tabela, chamada `Tracks`, contendo duas colunas:

```
import sqlite3

conn = sqlite3.connect('music.sqlite3')
cur = conn.cursor()

cur.execute('DROP TABLE IF EXISTS Tracks ')
cur.execute('CREATE TABLE Tracks (title TEXT, plays INTEGER)')

conn.close()
```

A operação `connect` cria uma "conexão" com o banco de dados armazenado no arquivo `music.sqlite3` no diretório corrente. Se o arquivo não existir, este será criado. O motivo para isto ser chamado de "conexão" é que algumas vezes o banco de dados está em um "servidor de banco de dados" separado da aplicação propriamente dita. Em nossos exemplos o banco de dados está armazenado localmente em um arquivo no mesmo diretório que o código Python está sendo executado.

Um **cursor** é como um identificador de arquivo que podemos utilizar para realizar operações sobre as informações armazenadas em um banco de dados. Ao chamar a função `cursor()`, conceitualmente, é similar ao chamar a função `open()` quando estamos trabalhando com arquivos de texto.

[1] Atualmente o SQLite permite uma maior flexibilidade em relação aos tipos de dados que são armazenados em uma coluna, mas vamos manter os tipos de dados restritos neste capítulo, assim os mesmos conceitos aprendidos aqui podem ser aplicados a outros sistemas de banco de dados como MySQL.



Uma vez que temos o cursor, podemos começar a executar comandos no conteúdo armazenado no banco de dados utilizando o método `execute()`.

Os comandos de um banco de dados são expressos em uma linguagem especial que foi padronizada por diferentes fornecedores de bancos de dados, que nos permite aprender uma única linguagem. A linguagem dos bancos de dados é chamada de **Structured Query Language**[2] ou referenciada pelo acrônimo **SQL** `http://en.wikipedia.org/wiki/SQL`

Em nossos exemplos, estamos executando dois comandos SQL no banco de dados que criamos. Convencionaremos que os comandos SQL serão mostrados em maiúsculas e as partes que não são palavras reservadas do SQL (como os nomes das tabelas e colunas) serão mostrados em minúsculas.

O primeiro comando SQL remove a tabela `Tracks` do banco de dados se ela existir. Este padrão nos permite executar o mesmo programa para criar a tabela `Tracks` repetidas vezes sem que cause erro. Perceba que o comando `DROP TABLE` remove a tabela e todo o seu conteúdo do banco de dados (i.e., não é possível desfazer esta operação)

```
cur.execute('DROP TABLE IF EXISTS Tracks ')
```

O segundo comando cria a tabela `Tracks` com uma coluna chamada `title` com o tipo texto e uma coluna chamada `plays` com o tipo inteiro.

```
cur.execute('CREATE TABLE Tracks (title TEXT, plays INTEGER)')
```

Agora que criamos a tabela `Tracks`, podemos inserir algum dado dentro dela utilizando a operação SQL `INSERT`. Novamente, estamos estabelecendo uma conexão com o banco de dados e obtendo o `cursor`. E então executamos o comando SQL utilizando o cursor.

O comando SQL `INSERT` indica qual tabela estamos utilizando, e em seguida, cria uma nova linha listando quais campos utilizaremos para incluir `(title, plays)` seguido pelo comando `VALUES` com os valores que desejamos adicionar na nova linha. Especificamos os valores utilizando pontos de interrogação `(?, ?)` para indicar que os valores serão passados como tuplas `( 'My Way', 15)` como um segundo parâmetro da chamada `execute()`.

[2] Em Português, pode ser chamada de Linguagem de Consulta Estruturada

```
import sqlite3

conn = sqlite3.connect('music.sqlite3')
cur = conn.cursor()

cur.execute('INSERT INTO Tracks (title, plays) VALUES ( ?, ? )',
    ( 'Thunderstruck', 20 ) )
cur.execute('INSERT INTO Tracks (title, plays) VALUES ( ?, ? )',
    ( 'My Way', 15 ) )
conn.commit()

print 'Tracks:'
cur.execute('SELECT title, plays FROM Tracks')
for row in cur :
   print row

cur.execute('DELETE FROM Tracks WHERE plays < 100')
conn.commit()

cur.close()
```

Primeiro nós adicionamos com `INSERT` duas linhas na nossa tabela e usaremos `commit()` para forçar a escrita da informação no arquivo do banco de dados.

Faixas

título	tocadas
Thunderstruck	20
My Way	15

Depois usamos o comando `SELECT` para buscar a linha que acabamos de inserir na tabela. Com o comando `SELECT`, indicamos que coluna gostaríamos `(title, plays)` e de qual tabela queremos buscar a informação. Depois de confirmar a execução do comando `SELECT`, o cursor pode ser utilizado como repetição através de um comando `for`. Por questões de eficiência, o cursor não lê toda a informação da base de dados quando executamos o comando `SELECT`. Ao invés disto, a informação é lida sob demanda enquanto iteramos através da linha com o comando `for`.

A saída do programa fica da seguinte forma:

```
Tracks:
(u'Thunderstruck', 20)
(u'My Way', 15)
```

A iteração do `for` encontrou duas linhas, e cada linha é uma tupla em Python com o primeiro valor como `title` e o segundo como o número de `plays`. Não se preocupe com o fato de que *strings* são mostrados com o caractere `u'` no começo. Isto é uma indicação que a *string* estão em **Unicode**, o que indica que são capazes de armazenar um conjunto de caractere não-Latin.

No final do programa, executamos o comando SQL `DELETE` para remover as linhas que acabamos de criar, assim podemos executar o programa repetidas vezes. O `DELETE` pode ser utilizado com a condição `WHERE` que permite selecionar através de uma expressão o critério permitindo pesquisar no banco de dados somente as linhas que correspondem com a expressão utilizada. Neste exemplo a expressão construida se aplica em todas as linhas, para que possamos executar o programa outras vezes. Depois de executar o `DELETE` chamamos o `commit()` para forçar que o dado seja removido do banco de dados.

12.5 Resumo de Structured Query Language (SQL)

Estamos utilizando SQL junto com os exemplos de Python e até agora cobrimos muitos comandos SQL básicos. Nesta seção, vamos olhar a linguagem SQL com mais atenção e apresentaremos uma visão geral da sintaxe do SQL.

Existem diferentes fornecedores de bancos de dados, a linguagem SQL foi padronizada, desta forma podemos nos comunicar de maneira portável entre os diferentes sistemas de banco de dados dos diferentes fornecedores.

Basicamente um banco de dados relacional é composto por tabelas, linhas e colunas. As colunas geralmente possuem tipos, como textos, números ou informação de data. Quando criamos uma tabela, indicamos os nomes e tipos das colunas:

```
CREATE TABLE Tracks (title TEXT, plays INTEGER)
```

Para inserir uma linha em uma tabela, utilizamos o comando SQL `INSERT`:

```
INSERT INTO Tracks (title, plays) VALUES ('My Way', 15)
```

A declaração do `INSERT` especifica o nome da tabela, e então, uma lista dos campos/colunas que gostaríamos de definir na nova linha, e por fim, através do campo `VALUES` passamos uma lista de valores correspondentes a cada campo.

O comando `SELECT` é utilizado para buscar as linhas e colunas de um banco de dados. A declaração do `SELECT` permite que você especifique qual coluna gostaria de buscar, bem como utilizando a condição do `WHERE`, permite selecionar qual linha gostaríamos de visualizar. Isto também possibilita o uso de uma condição opcional, `ORDER BY`, para ordenar as linhas retornadas.

```
SELECT * FROM Tracks WHERE title = 'My Way'
```

O uso do `*` indica que o banco de dados deve retornar todas as colunas para cada linha que casa com a condição `WHERE`.

Atenção, diferente de Python, a condição `WHERE`, em SQL, utiliza o sinal de igual simples (=), para indicar uma condição de igualdade, ao invés de um sinal duplo (==) <, >, <=, >=, !=,

assim como é possível utilizar as condições `AND` e `OR` e parênteses para construir expressões lógicas.

Você pode pedir que as linhas retornadas sejam ordenadas por um dos campos como apresentados no exemplo a seguir:

```
SELECT title,plays FROM Tracks ORDER BY title
```

Para remover uma linha, é preciso combinar a condição `WHERE` com a condição `DELETE`. O `WHERE` irá determinar quais linhas serão removidas:

```
DELETE FROM Tracks WHERE title = 'My Way'
```

É possível alterar/atualizar uma ou mais colunas e suas linhas de uma tabela utilizando a condição SQL `UPDATE`, da seguinte forma:

```
UPDATE Tracks SET plays = 16 WHERE title = 'My Way'
```

A condição `UPDATE` especifica uma tabela e depois uma lista de campos e valores que serão alterados após o comando `SET`, e utilizando uma condição `WHERE`, opcional, é possível selecionar as linhas que serão atualizadas. Uma condição `UPDATE` irá mudar todas as linhas que casam com a condição `WHERE`. Se a condição `WHERE` não for especificada, o `UPDATE` será aplicado em todas as linhas da tabela.

Os quatro comandos básicos de SQL (INSERT, SELECT, UPDTE e DELETE) permitem as quatro operações básicas necessárias para criação e manutenção das informações em um banco de dados.

12.6 Rastreando o Twitter utilizando um banco de dados

Nesta seção, criaremos um programa simples para rastreamento que navegará através de contas de usuários do Twitter e construirá um banco de dados referentes a estes usuários. *Nota: Tenha muito cuidado ao executar este programa. Você não irá querer extrair muitas informações ou executar o programa por muito tempo e acabar tendo sua conta do Twitter bloqueada.*

Um dos problemas, em qualquer tipo de programas de rastreamento, é que precisa ser capaz de ser interrompido e reiniciado muitas vezes e você não quer perder informações que você já tenha recuperado até agora. Não quer sempre reiniciar a recuperação dos dados desde o começo, então armazenamos as informações tão logo seja recuperada, assim o programa poderá reiniciar a busca do ponto onde parou.

Vamos começar recuperando os amigos de uma pessoa no Twitter e seus status, iterando na lista de amigos, e adicionando cada um ao banco de dados para que possa ser recuperado no futuro. Depois de listar os amigos de uma pessoa, verificamos na nossa base de dados e coletamos os amigos de um dos amigos da

primeira pessoa. Vamos fazendo isto repetidas vezes, escolhendo umas das pessoas "não visitadas", recuperando sua lista de amigos, e adicionando amigos que não tenhamos visto anteriormente a nossa lista, para visitar futuramente.

Também rastrearemos quantas vezes vimos um amigo em particular na nossa base para ter uma ideia da sua "popularidade".

Armazenando nossa lista de contas conhecidas, no banco de dados no disco do nosso computador, e se já recuperamos a conta ou não, e quanto esta conta é popular, podemos parar e recomeçar nosso programa quantas vezes quisermos.

Este programa é um pouco complexo. É baseado em um exercício apresentado anteriormente neste livro, que utiliza a API do Twitter.

O seguinte código apresenta o programa que realiza o rastreamento no Twitter:

```
import urllib
import twurl
import json
import sqlite3

TWITTER_URL = 'https://api.twitter.com/1.1/friends/list.json'

conn = sqlite3.connect('spider.sqlite3')
cur = conn.cursor()

cur.execute('''
CREATE TABLE IF NOT EXISTS Twitter
(name TEXT, retrieved INTEGER, friends INTEGER)''')

while True:
    acct = raw_input('Enter a Twitter account, or quit: ')
    if ( acct == 'quit' ) : break
    if ( len(acct) < 1 ) :
        cur.execute('SELECT name FROM Twitter WHERE retrieved = 0 LIMIT 1')
        try:
            acct = cur.fetchone()[0]
        except:
            print 'No unretrieved Twitter accounts found'
            continue

    url = twurl.augment(TWITTER_URL,
                {'screen_name': acct, 'count': '20'} )
    print 'Retrieving', url
    connection = urllib.urlopen(url)
    data = connection.read()
    headers = connection.info().dict
    # print 'Remaining', headers['x-rate-limit-remaining']
    js = json.loads(data)
    # print json.dumps(js, indent=4)

    cur.execute('UPDATE Twitter SET retrieved=1 WHERE name = ?', (acct, ) )

    countnew = 0
```

```
    countold = 0
    for u in js['users'] :
        friend = u['screen_name']
        print friend
        cur.execute('SELECT friends FROM Twitter WHERE name = ? LIMIT 1',
            (friend, ) )
        try:
            count = cur.fetchone()[0]
            cur.execute('UPDATE Twitter SET friends = ? WHERE name = ?',
                (count+1, friend) )
            countold = countold + 1
        except:
            cur.execute('''INSERT INTO Twitter (name, retrieved, friends)
                VALUES ( ?, 0, 1 )''', ( friend, ) )
            countnew = countnew + 1
    print 'New accounts=',countnew,' revisited=',countold
    conn.commit()

cur.close()
```

Nossa base de dados está armazenada no arquivo `spider.sqlite3` e possui uma tabela chamada `Twitter`. Cada linha na tabela `Twitter` tem uma coluna para o nome da conta, se já recuperamos os amigos desta conta, e quantas vezes esta conta foi "seguida".

Na repetição principal do programa, pedimos ao usuário uma conta de Twitter ou "quit" para sair do programa. Se o usuário informar um usuário do Twitter, o programa começa a recuperar a lista de amigos e os status para aquele usuário e adiciona cada amigo na base de dados, se ainda não existir. Se o amigo já está na lista, nós adicionamos "1" no campo `friends` da base de dados.

Se o usuário pressionar `enter`, pesquisamos na base a próxima conta que não rastreamos ainda, e então rastreamos os amigos e status com aquela conta e adicionamos na base de dados ou atualizamos, incrementando seu contador de `friends`.

Uma vez que rastreamos a lista de amigos e status, iteramos entre todas os ítens `user` retornados no JSON e rastreamos o `screen_name` para cada usuário. Então utilizamos a declaração `SELECT` para ver se já armazenamos este `screen_name` em particular na base e recuperamos o contador de amigos (`friends`), se este registro existir.

```
    countnew = 0
    countold = 0
    for u in js['users'] :
        friend = u['screen_name']
        print friend
        cur.execute('SELECT friends FROM Twitter WHERE name = ? LIMIT 1',
            (friend, ) )
        try:
            count = cur.fetchone()[0]
            cur.execute('UPDATE Twitter SET friends = ? WHERE name = ?',
                (count+1, friend) )
            countold = countold + 1
```

```
        except:
            cur.execute('''INSERT INTO Twitter (name, retrieved, friends)
                VALUES ( ?, 0, 1 )''', ( friend, ) )
            countnew = countnew + 1
    print 'New accounts=',countnew,' revisited=',countold
    conn.commit()
```

Uma vez que o cursor tenha executado o `SELECT`, nós devemos recuperar as linhas. Podemos fazer isto com uma declaração de `for`, mas uma vez que estamos recuperando uma linha (`LIMIT 1`), podemos utilizar o método `fetchone()` para buscar a primeira (e única) linha que é o resultado da operação `SELECT`. Sendo o retorno `fetchone()` uma linha como uma **tupla** (ainda que haja somente um campo), pegamos o primeiro valor da tupla utilizando índice `[0]` para pegar o contador de amigos atual dentro da variável `count`.

Se a busca for bem sucedida, utilizamos a declação `UPDATE` com a cláusula `WHERE` para adicionar 1 na coluna `friends` para a linha que corresponde com a conta do amigo. Note que existem dois espaços reservados (i.e., pontos de interrogações) no SQL, e o segundo parâmetro para o `execute()` é uma tupla que armazena o valor para substituir no SQL no lugar dos pontos de interrogações.

Se o bloco `try` falhar, é provavelmente por que nenhum resultado corresponde a cláusula em `WHERE name = ?` do SELECT. Então no block `except`, utilizamos a declaração `INSERT` para adicionar o `screen_name` do amigo a tabela com a indicação que ainda não rastreamos o `screen_name` e setamos o contador de amigos com 0 (zero).

Assim, a primeira vez que o programa é executado e informamos uma conta do Twitter, a saída do programa é a seguinte:

```
Enter a Twitter account, or quit: drchuck
Retrieving http://api.twitter.com/1.1/friends ...
New accounts= 20  revisited= 0
Enter a Twitter account, or quit: quit
```

Como esta é a primeira vez que executamos o programa, o banco de dados está vazio e criamos o banco no arquivo `spider.sqlite3`, adicionamos a tabela chamada `Twitter` na base de dados. Então nós rastreamos alguns amigos e os adicionamos a base, uma vez que ela está vazia.

Neste ponto podemos escrever um *dumper* simples para olhar o que está no nosso arquivo `spider.sqlite3`:

```
import sqlite3

conn = sqlite3.connect('spider.sqlite3')
cur = conn.cursor()
cur.execute('SELECT * FROM Twitter')
count = 0
for row in cur :
   print row
```

```
   count = count + 1
print count, 'rows.'
cur.close()
```

Este programa abre o banco de dados e seleciona todas as colunas de todas as linhas na tabela `Twitter`, depois itera em cada linha e imprime o valor dentro de cada uma.

Se executarmos este programa depois da primeira execução do nosso rastreador *spider* do Twitter, sua saída será como a seguinte:

```
(u'opencontent', 0, 1)
(u'lhawthorn', 0, 1)
(u'steve_coppin', 0, 1)
(u'davidkocher', 0, 1)
(u'hrheingold', 0, 1)
...
20 rows.
```

Veremos uma linha para cada `screen_name`, que não tenhamos recuperado o dado daquele `screen_name`, e todos tem um amigo.

Agora nosso banco de dados reflete quais amigos estão relacionados com a nossa primeira conta do Twitter (**drchuck**) utilizada para rastreamento. Podemos executar o programa novamente e mandar rastrear a próxima conta "não processada" e recuperar os amigos, simplesmente pressionando `enter` ao invés de informar uma conta do Twitter, conforme o exemplo a seguir:

```
Enter a Twitter account, or quit:
Retrieving http://api.twitter.com/1.1/friends ...
New accounts= 18  revisited= 2
Enter a Twitter account, or quit:
Retrieving http://api.twitter.com/1.1/friends ...
New accounts= 17  revisited= 3
Enter a Twitter account, or quit: quit
```

Uma vez que pressionamos `enter` (i.e., não especificamos uma conta do Twitter), o seguinte código é executado:

```
    if ( len(acct) < 1 ) :
        cur.execute('SELECT name FROM Twitter WHERE retrieved = 0 LIMIT 1')
        try:
            acct = cur.fetchone()[0]
        except:
            print 'No unretrieved twitter accounts found'
            continue
```

Utilizamos a declaração SQL `SELECT` para recuperar o nome do primeiro (`LIMIT 1`) usuário que ainda tem seu "recuperamos este usuário" com o valor setado em zero. Também utilizamos o padrão `fetchone()[0]` dentro de um bloco try/except para extrair também um `screen_name` do dado recuperado ou apresentamos uma mensagem de erro e iteramos novamente.

Se tivermos sucesso ao recuperar um `screen_name` não processado, vamos extrair seus dados da seguinte maneira:

```
url = twurl.augment(TWITTER_URL, {'screen_name': acct, 'count': '20'} )
print 'Retrieving', url
connection = urllib.urlopen(url)
data = connection.read()
js = json.loads(data)

cur.execute('UPDATE Twitter SET retrieved=1 WHERE name = ?', (acct, ) )
```

Ao recuperar os dados com sucesso, utilizaremos a declaração `UPDATE` para setar a coluna `retrieved` para 1 para indicar que completamos a extração dos amigos relacionados com esta conta. Isto no permite recuperar o mesmo dado diversas vezes e nos permite prosseguir através da lista de amigos no Twitter.

Se executarmos o programa novamente, e pressionarmos `enter` duas vezes seguidas para recuperar os próximos amigos do amigo e depois executarmos o programa de *dumping*, ele nos mostrará a seguinte saída:

```
(u'opencontent', 1, 1)
(u'lhawthorn', 1, 1)
(u'steve_coppin', 0, 1)
(u'davidkocher', 0, 1)
(u'hrheingold', 0, 1)
...
(u'cnxorg', 0, 2)
(u'knoop', 0, 1)
(u'kthanos', 0, 2)
(u'LectureTools', 0, 1)
...
55 rows.
```

Podemos ver que gravamos de forma apropriada que visitamos os usuários `lhawthorn` e `opencontent`. E que as contas `cnxorg` e `kthanos` já tem dois seguidores. Desde que tenhamos recuperado os amigos de três pessoas (`drchuck`, `opencontent`, e `lhawthorn`) nossa tabela tem agora 55 linhas de amigos para recuperar.

Cada vez que executamos o programa e pressionamos `enter` ele pegará a próxima conta não visitada (e.g., a próxima conta será `steve_coppin`), recuperar seus amigos, marcá-los como recuperados, e para cada um dos amigos de `steve_coppin` também adicionaremos eles para no fim da base de dados e atualizaremos seus amigos que já estiverem na base de dados.

Assim que os dados do programa estejam armazenados no disco em um banco de dados, o rastreamento pode ser suspenso e reiniciado tantas vezes quanto quiser, sem a perda de informações.

12.7 Modelagem de dados básica

O verdadeiro poder de um banco de dados relacional é quando criamos múltiplas tabelas e criamos ligações entre elas. Decidir como dividir os dados da sua aplicação em diferentes tabelas e estabelecer a relação entre estas tabelas é o que chamamos de **modelagem de dados**. O documento que mostra a estrutura das tabelas e suas relações é chamado de **modelo de dados**.

Modelagem de dados é uma habilidade relativamente sofisticada e nesta seção nós iremos somente introduzir os conceitos mais básicos da modelagem de dados relacionais. Para maiores detalhes sobre modelagem de dados você pode começar com:

`http://en.wikipedia.org/wiki/Relational_model`

Digamos que para a nossa aplicação de rastreamento do Twitter, ao invés de só contar os amigos das pessoas, nós queiramos manter uma lista de todas as relações de entrada, então poderemos encontrar uma lista de todos que seguem uma pessoa em particular.

Já que todos, potencialmente, terão tantas contas que o sigam, nós não podemos simplesmente adicionar uma coluna para nossa tabela `Twitter`. Então criamos uma nova tabela que mantém o controle dos pares de amigos. A seguir temos uma forma simples de criar tal tabela:

```
CREATE TABLE Pals (from_friend TEXT, to_friend TEXT)
```

Toda vez que encontrarmos uma pessoa que `drchuck` está seguindo, nós iremos inserir uma linha da seguinte forma:

```
INSERT INTO Pals (from_friend,to_friend) VALUES ('drchuck', 'lhawthorn')
```

Como estamos processando 20 amigos da conta do *Twitter* do `drchuck`, vamos inserir 20 registros com "drchuck" como primeiro parâmetro e assim acaberemos duplicando a *string* muitas vezes no banco de dados.

Esta duplicação de dados, viola uma das melhores práticas da **normatização de banco de dados** que basicamente afirma que nunca devemos colocar o mesmo dado mais de uma vez em um banco de dados. Se precisarmos inserir um dado mais de uma vez, criamos uma referência numérica **key** (chave) para o dado, e utilizamos a chave para referenciar o dado.

Na prática, uma *string* ocupa muito mais espaço do que um inteiro, no disco e na memória do nosso computador, e leva mais tempo do processador para comparar e ordenar. Se tivermos somente algumas centenas de entradas, a base de dados e o tempo de processamento dificilmente importarão. Mas se tivermos um milhão de pessoas na nossa base de dados e uma possibilidade de 100 milhões de conexões de amigos, é importante permitir examinar os dados o mais rápido possível.

Nós armazenaremos nossas contas do Twitter em uma tabela chamada `People` ao invés de utilizar a tabela `Twitter` utilizada no exemplo anterior. A tabela `People`

tem uma coluna adicional para armazenar uma chave associada a linha para este usuário.

Podemos criar a tabela `People` com esta coluna `id` adicional com o seguinte comando:

```
CREATE TABLE People
    (id INTEGER PRIMARY KEY, name TEXT UNIQUE, retrieved INTEGER)
```

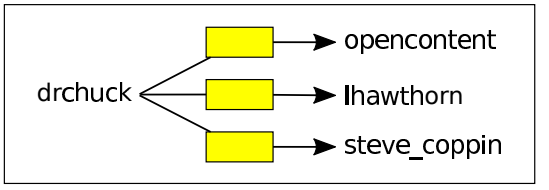
Perceba que nós não estamos mais mantendo uma conta de amigo em cada linha da tabela `People`. Quando selecionamos `INTEGER PRIMARY KEY` como o tipo da nossa coluna `id`, estamos indicando que gostaríamos que o SQLite gerencie esta coluna e defina uma chave numérica única automagicamente para cada linha que inserirmos. Também adicionamos uma palavra-chave `UNIQUE` para indicar que não permitiremos ao SQLite inserir duas linhas com o mesmo valor para `name`.

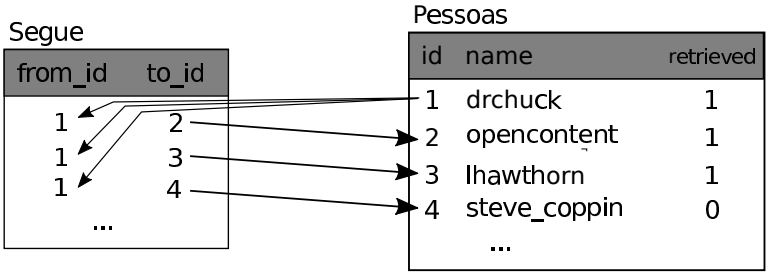
Agora, ao invés de criar a tabela `Pals` acima, criaremos uma tabela chamada `Follows` com duas colunas com o tipo inteiro `from_id` e `to_id` e associaremos na tabela onde a *combinação* de `from_id` e `to_id` devem ser únicos nesta tabela (i.e., não podemos inserir linhas duplicadas) na nossa base de dados.

```
CREATE TABLE Follows
    (from_id INTEGER, to_id INTEGER, UNIQUE(from_id, to_id) )
```

Quando adicionamos a condição `UNIQUE` a nossa tabela, estamos definindo um conjunto de regras e pedindo a base de dados para cumprir estas regras quando tentarmos inserir algum registro. Estamos criando estas regras como uma conveniência no nosso programa, como veremos a seguir. As regras nos impedem de cometer enganos e facilitam na escrita dos nossos códigos.

Em essência, criando a tabela `Follows`, estamos modelando uma "relação" onde uma pessoa "segue" outro alguém e representamos isto com um par de números indicando que (a) as pessoas estão conectadas e (b) a direção do relacionamento.





12.8 Programando com múltiplas tabelas

Agora nós iremos refazer o programa de rastreio do Twitter utilizando duas tabelas, as chaves primárias, e as chaves de referências estão descritas anteriormente. Abaixo está o código da nova versão do programa:

```
import urllib
import twurl
import json
import sqlite3

TWITTER_URL = 'https://api.twitter.com/1.1/friends/list.json'

conn = sqlite3.connect('friends.sqlitesqlite3')
cur = conn.cursor()

cur.execute('''CREATE TABLE IF NOT EXISTS People
    (id INTEGER PRIMARY KEY, name TEXT UNIQUE, retrieved INTEGER)''')
cur.execute('''CREATE TABLE IF NOT EXISTS Follows
    (from_id INTEGER, to_id INTEGER, UNIQUE(from_id, to_id))''')

while True:
    acct = raw_input('Enter a Twitter account, or quit: ')
    if ( acct == 'quit' ) : break
    if ( len(acct) < 1 ) :
        cur.execute('''SELECT id, name FROM People
            WHERE retrieved = 0 LIMIT 1''')
        try:
            (id, acct) = cur.fetchone()
        except:
            print 'No unretrieved Twitter accounts found'
            continue
    else:
        cur.execute('SELECT id FROM People WHERE name = ? LIMIT 1',
            (acct, ) )
        try:
            id = cur.fetchone()[0]
        except:
            cur.execute('''INSERT OR IGNORE INTO People (name, retrieved)
                VALUES ( ?, 0)''', ( acct, ) )
            conn.commit()
            if cur.rowcount != 1 :
                print 'Error inserting account:',acct
                continue
            id = cur.lastrowid

    url = twurl.augment(TWITTER_URL,
       {'screen_name': acct, 'count': '20'} )
    print 'Retrieving account', acct
    connection = urllib.urlopen(url)
    data = connection.read()
    headers = connection.info().dict
    print 'Remaining', headers['x-rate-limit-remaining']

    js = json.loads(data)
```

```
    # print json.dumps(js, indent=4)

    cur.execute('UPDATE People SET retrieved=1 WHERE name = ?', (acct, ) )

    countnew = 0
    countold = 0
    for u in js['users'] :
        friend = u['screen_name']
        print friend
        cur.execute('SELECT id FROM People WHERE name = ? LIMIT 1',
            (friend, ) )
        try:
            friend_id = cur.fetchone()[0]
            countold = countold + 1
        except:
            cur.execute('''INSERT OR IGNORE INTO People (name, retrieved)
                VALUES ( ?, 0)''', ( friend, ) )
            conn.commit()
            if cur.rowcount != 1 :
                print 'Error inserting account:',friend
                continue
            friend_id = cur.lastrowid
            countnew = countnew + 1
        cur.execute('''INSERT OR IGNORE INTO Follows (from_id, to_id)
            VALUES (?, ?)''', (id, friend_id) )
    print 'New accounts=',countnew,' revisited=',countold
    conn.commit()

cur.close()
```

Este programa está começando a ficar um pouco complicado, mas ilustra os padrões que precisamos para utilizar quando estamos usando chaves inteiras para conectar as tabelas. Os padrões básicos são:

1. Criar tabelas com chaves primárias e restrições.
2. Quando temos uma chave lógica para uma pessoa (i.e., conta) e precisamos do valor de `id` para a pessoa, dependendo se a pessoa já está na tabela `People` ou não, também precisaremos de: (1) olhar para a pessoa na tabela `People` e recuperar o valor de `id` da pessoa, ou (2) adicionar a pessoa na tabela `People` e pegar o valor de `id` para a nova linha recém adicionada.
3. Inserir a linha que captura a relação com "segue".

Vamos tratar cada um dos itens acima, em partes.

12.8.1 Restrições em uma tabela

Da forma como projetamos a estrutura da tabela, podemos informar ao banco de dados que gostaríamos de reforçar algumas regras. Estas regras nos ajudam a não cometer enganos e a não inserir dados incorretos nas nossas tabelas. Quando criamos nossas tabelas:

```
cur.execute('''CREATE TABLE IF NOT EXISTS People
    (id INTEGER PRIMARY KEY, name TEXT UNIQUE, retrieved INTEGER)''')
cur.execute('''CREATE TABLE IF NOT EXISTS Follows
    (from_id INTEGER, to_id INTEGER, UNIQUE(from_id, to_id))''')
```

Indicamos que a coluna `name` na tabela `People` deve ser `UNIQUE`. Também indicaremos que a combinação dos dois números em cada linha da tabela `Follows` devem ser únicos. Estas restrições nos mantém longe da possibilidade de cometer enganos como adicionar a mesma relação mais de uma vez.

Podemos obter vantagens destas restrições conforme mostra o seguinte código:

```
cur.execute('''INSERT OR IGNORE INTO People (name, retrieved)
    VALUES ( ?, 0)''', ( friend, ) )
```

Adicionamos a condição `OR IGNORE` a declaração `INSERT` para indicar que se este `INSERT` em particular pode causar uma violação para a regra "`name` deve ser único", assim o banco de dados tem permissão de ignorar o `INSERT`.

De forma similar, o seguinte código garante que não adicionemos a mesma relação `Follows` duas vezes.

```
cur.execute('''INSERT OR IGNORE INTO Follows
    (from_id, to_id) VALUES (?, ?)''', (id, friend_id) )
```

Novamente, nós simplesmente dizemos para o banco de dados ignorar nossa tentativa de inserir `INSERT` se isto violar a regra de exclusividade que especificamos para a linha `Follows`.

12.8.2 Restaurar e/ou inserir um registro

Quando solicitamos ao usuário uma conta do Twitter, se a conta existir, precisamos verificar o valor do `id`. Se a conta não existir ainda na tabela `People`, devemos inserir o registro e pegar o valor do `id` da linha inserida.

Isto é um padrão muito comum e é feito duas vezes no programa acima. Este código mostra como verificamos o `id` da conta de um amigo, quando extraímos um `screen_name` de um nó de `user` recuperado do JSON do Twitter.

Ao longo do tempo será cada vez mais provável que a conta já esteja registrada no banco de dados, então primeiro checamos para ver se o registro existe em `People` utilizando uma declaração de `SELECT`.

Se tudo estiver certo[3] dentro da seção `try`, recuperamos o registro usando `fetchone()` e depois recuperar o primeiro (e somente o primeiro) elemento da tupla que retornou e a armazenamos em `friend_id`.

Se o `SELECT` falhar, o código de `fetchone()[0]` falhará e o controle irá mudar para a seção `except`.

[3]Em geral, quando uma sentença inicia com "se tudo estiver certo" você verá que o código precisa utilizar a condição *try/except*.

```
        friend = u['screen_name']
        cur.execute('SELECT id FROM People WHERE name = ? LIMIT 1',
            (friend, ) )
        try:
            friend_id = cur.fetchone()[0]
            countold = countold + 1
        except:
            cur.execute('''INSERT OR IGNORE INTO People (name, retrieved)
                VALUES ( ?, 0)''', ( friend, ) )
            conn.commit()
            if cur.rowcount != 1 :
                print 'Error inserting account:',friend
                continue
            friend_id = cur.lastrowid
            countnew = countnew + 1
```

Se terminar no código do `except`, isto significa que aquela linha não foi encontrada, então devemos inserir o registro. Usamos o `INSERT OR IGNORE` somente para evitar erros e depois chamamos `commit()` para forçar que a base de dados seja atualizada. Depois que a escrita esteja completa, nós podemos checar, com `cur.rowcount`, para ver quantas linhas foram afetadas. Uma vez que estamos tentando inserir uma simples linha, se o número de linhas afetadas é alguma coisa diferente de 1, isto é um erro.

Se o `INSERT` for executado com sucesso, nós podemos verificar, através do `cur.lastrowid` para descobrir qual valor o banco de dados associou na coluna `id` na nossa nova linha.

12.8.3 Armazenando a conexão do amigo

Uma vez que sabemos o valor da chave para o usuário do Twitter e o amigo extraído do JSON, simplesmente inserimos os dois números dentro da tabela `Follows` com o seguinte código:

```
cur.execute('INSERT OR IGNORE INTO Follows (from_id, to_id) VALUES (?, ?)',
    (id, friend_id) )
```

Note que deixamos o banco de dados cuidar para nós de realizar a "inserção-dupla" da conexão criando a tabela com a restrição única e depois adicionando `OR IGNORE` a nossa condição de `INSERT`.

Esta é um exemplo de execução deste programa:

```
Enter a Twitter account, or quit:
No unretrieved Twitter accounts found
Enter a Twitter account, or quit: drchuck
Retrieving http://api.twitter.com/1.1/friends ...
New accounts= 20  revisited= 0
Enter a Twitter account, or quit:
Retrieving http://api.twitter.com/1.1/friends ...
New accounts= 17  revisited= 3
```

```
Enter a Twitter account, or quit:
Retrieving http://api.twitter.com/1.1/friends ...
New accounts= 17  revisited= 3
Enter a Twitter account, or quit: quit
```

Nós iniciamos com a conta `drchuck` e depois deixamos o programa pegar automaticamente as próximas duas contas para recuperar e adicionar à nossa base de dados.

Abaixo estão as primeiras linhas das tabelas `People` e `Follows` depois que esta execução é finalizada:

```
People:
(1, u'drchuck', 1)
(2, u'opencontent', 1)
(3, u'lhawthorn', 1)
(4, u'steve_coppin', 0)
(5, u'davidkocher', 0)
55 rows.
Follows:
(1, 2)
(1, 3)
(1, 4)
(1, 5)
(1, 6)
60 rows.
```

Você também pode ver os campos `id`, `name` e `visited` na tabela `People` e os números dos finais das conexões na tabela `Follows`. Na tabela `People`, podemos ver que as primeiras três pessoas foram visitadas e seus dados foram recuperados. Os dados na tabela `Follows` indica que `drchuck` (usuário 1) é amigo de todas as pessoas mostradas nas primeiras cincos linhas. Isto mostra que o primeiro dado que recuperamos e armazenamos foram dos amigos do `drchuck`. Se você mostrou mais linhas da tabela `Follows`, você poderá ver os amigos dos usuários 2 e 3.

12.9 Três tipos de chaves

Agora que iniciamos a construção de um modelo de dados colocando nossos dados dentro de múltiplas tabelas e linhas conectadas nestas tabelas utilizando **chaves**, precisamos olhar para algumas terminologias sobre as chaves. Existem genericamente, três tipos de chaves que são utilizadas em um banco de dados.

- Uma **chave lógica** é uma chave que o "mundo real" pode usar para consultar um registro. Na nossa modelagem, o campo `name` é uma chave lógica. Ele é o nome que usamos para consultar um registro de usuário diversas vezes no nosso programa, usando o campo `name`. Você perceberá que faz sentido adicionar a restrição de `UNIQUE` para uma chave lógica. Uma vez que é através de chaves lógicas que consultamos uma linha do mundo exterior, faz pouco sentido permitir que múltiplas linhas tenham o mesmo valor em uma tabela.

- Um **chave primária** é usualmente um número que é associado automaticamente por um banco de dados. Geralmente não terá significado fora do programa e só é utilizada para conectar as linhas de diferentes tabelas. Quando queremos verificar uma linha em uma tabela, normalmente buscamos pela linha utilizando a chave primária, é a forma mais rápida de encontrar uma linha. Uma vez que chaves primárias são números inteiros, eles ocupam pouco espaço e podem ser comparados ou ordenados rapidamente. No nosso modelo, o campo `id` é um exemplo de chave primária.

- Uma **chave estrangeira** é normalmente um número que aponta para a chave primária associada a uma linha em outra tabela. Um exemplo de chave estrangeira no nosso modelo é o campo `from_id`.

Nós estamos utilizando uma convenção de sempre chamar uma chave primária de um campo `id` e adicionando o sufixo `_id` para qualquer campo que seja uma chave estrangeira.

12.10 Utilizando o JOIN para recuperar informações

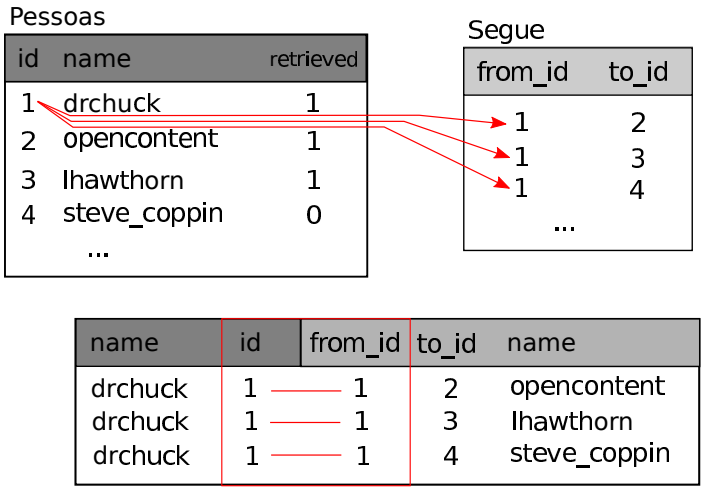
Agora que seguimos as regras de normalização de bancos de dados e temos os dados separados em duas tabelas, associadas através de chaves primárias e chaves estrangeiras, precisamos ser capazes de construir uma chamada de `SELECT` que reagrupa os dados em toda as tabelas.

SQL utiliza a cláusula `JOIN` para reconectar estas tabelas. Na cláusula `JOIN` você especifica o campo que serão utilizados para reconectar as linhas entre as tabelas.

Abaixo, um exemplo de um `SELECT` com um `JOIN`:

```
SELECT * FROM Follows JOIN People
    ON Follows.from_id = People.id WHERE People.id = 1
```

O `JOIN` indica os campos que utilizamos para selecionar registros, cruzando ambas as tabelas `Follows` e `People`. A cláusula `ON` indica como as duas tabelas devem ser unidas: Junta as linhas da tabela `Follows` às da tabela `People` onde o campo `from_id` em `Follows` tem o mesmo valor do campo `id` na tabela `People`.



O resultado do `JOIN` cria uma super "meta-linha" que tem os dois campos da tabela `People` que casam com os campos da tabela `Follows`. Onde existir mais de uma ocorrência entre o campo `id` e o `from_id` da tabela `People`, então o `JOIN` cria uma "meta-linha" para *cada* par de linhas que correspondem, duplicando os dados conforme for necessário.

O seguinte código demonstra o dado que nós teremos no banco de dados após o executar o programa coletor de dados (acima) diversas vezes.

```
import sqlite3

conn = sqlite3.connect('spider.sqlite3')
cur = conn.cursor()

cur.execute('SELECT * FROM People')
count = 0
print 'People:'
for row in cur :
   if count < 5: print row
   count = count + 1
print count, 'rows.'

cur.execute('SELECT * FROM Follows')
count = 0
print 'Follows:'
for row in cur :
   if count < 5: print row
   count = count + 1
print count, 'rows.'

cur.execute('''SELECT * FROM Follows JOIN People
    ON Follows.to_id = People.id WHERE Follows.from_id = 2''')
count = 0
print 'Connections for id=2:'
for row in cur :
   if count < 5: print row
   count = count + 1
print count, 'rows.'
```

```
cur.close()
```

Neste programa, primeiro descarregamos a tabela `People` e `Follows` e depois descarregamos um subconjunto de dados das tabelas juntas.

Aqui temos a saída do programa:

```
python twjoin.py
People:
(1, u'drchuck', 1)
(2, u'opencontent', 1)
(3, u'lhawthorn', 1)
(4, u'steve_coppin', 0)
(5, u'davidkocher', 0)
55 rows.
Follows:
(1, 2)
(1, 3)
(1, 4)
(1, 5)
(1, 6)
60 rows.
Connections for id=2:
(2, 1, 1, u'drchuck', 1)
(2, 28, 28, u'cnxorg', 0)
(2, 30, 30, u'kthanos', 0)
(2, 102, 102, u'SomethingGirl', 0)
(2, 103, 103, u'ja_Pac', 0)
20 rows.
```

Você vê as colunas das tabelas `People` e `Follows` e por último, o conjunto de linhas que é o resultado do `SELECT` com o `JOIN`.

No último `SELECT`, nós estamos procurando por contas que tem amigos de "conteúdo aberto" (i.e., `People.id=2`).

Em cada uma das "meta-linhas" da última seleção, as primeiras duas colunas são da tabela `Follows` seguidas pelas colunas três até cinco da tabela `People`. Você também pode ver que a segunda coluna (`Follows.to_id`) relaciona a terceira coluna (`People.id`) em cada uma das "meta-linhas" que foram unidas.

12.11 Sumário

Este capítulo cobriu os fundamentos para o uso, básico, de banco de dados no Python. É muito mais complicado escrever código para usar um banco de dados para armazenar informações do que dicionários ou arquivos com Python, então existem poucas razões para se utilizar um banco de dados, a menos que a sua aplicação realmente precise das capacidades de um banco de dados. As situações onde um banco de dados podem ser muito úteis são: (1) quando sua aplicação precisa realizar pequenas atualizações com um conjunto grande de dados, (2) quando

seus dados são tão grandes que não podem ser armazenados em um dicionário e você precisa acessar estas informações repetidas vezes, ou (3) quando você tem um processo de execução demorada e você quer ser capaz de parar e recomeçar e manter os dados entre as pesquisas.

Você pode construir um banco de dados simples com uma única tabela para atender muitas aplicações, mas a maiorias dos problemas vão necessitar de várias tabelas e conexões/relações entre as linhas em diferentes tabelas. Quando você começar a fazer relações entre as tabelas, é importante fazer algum planejamento e seguir as regras de normatização de banco de dados para fazer um melhor uso das capacidades dos bancos de dados. Uma vez que a principal motivação para utilizar um banco de dados é que você tenha um grande conjunto de dados para tratar, é importante modelar os dados eficientemente, assim seus programas poderão rodar tão rápido quanto for possível.

12.12 Depuração

Algo comum quando se está desenvolvendo um programa em Python para se conectar em um banco de dados SQLite, é executar um programa para checar os resultados utilizando o Navegador SQLite. O navegador permitirá que rapidamente verifique se o seu programa está funcionando corretamente.

Você deve ter cuidado, porque o SQLite cuida para que dois programas não façam modificações nos dados ao mesmo tempo. Por exemplo, se você abrir um banco de dados no navegador e faz alterações no banco de dados, enquanto não pressionar o botão "salvar", o navegador "trava" o arquivo do banco de dados e impede qualquer outro programa de acessar o arquivo. Desta forma seu programa em Python não conseguirá acessar o arquivo se ele estiver travado.

Então, uma solução é garantir que fechou o navegador ou utilizar o menu **Arquivo** para fechar o banco de dados no navegador antes de tentar acessar o banco de dados através do Python, evitando problemas no seu código porque o banco de dados está travado.

12.13 Glossário

atributo: Um dos valores dentro de uma tupla. Comumente chamado de "coluna" ou "campo".

restrição: Quando ordenamos a um banco de dados para reforçar uma regra em um campo ou em uma linha na tabela. Uma restrição comum é insistir que não pode haver valore duplicados em um campo em particular (i.e., todos os valores tem que ser únicos).

cursor: Um cursor permite execução de um comando SQL em um banco de dados e recuperar informações de um banco de dados. Um cursor é similar a um *socket* ou identificador de arquivos para uma conexão de rede e arquivos, respectivamente.

navegador de banco de dados: um conjunto de *software* que permite se conectar diretamente a um banco de dados e manipulá-lo diretamente, sem escrever um programa.

chave estrangeira: Uma chave numérica que aponta para uma chave primária de uma linha em outra tabela. Chaves estrangeiras estabelecem relações entre linhas armazenadas em diferentes tabelas.

índice: Dados adicionais que um banco de dados mantém, como linhas e inserções dentro de uma tabela para realizar consultas mais rápido.

chave lógica: Uma chave que o "mundo externo" utiliza para consultar uma linha em particular. Por exemplo em uma tabela de contas de usuários, o e-mail de uma pessoa pode ser um bom candidato a chave lógica para a informação de usuário.

normatização: Projetar um modelo de dados para que nenhum dado seja replicado. Armazenamos cada item de dados em um lugar no banco de dados e referenciamos isso em qualquer lugar utilizando chave estrangeira.

chave primária: Uma chave numérica associada a cada linha que é usada para referenciar uma linha em uma tabela de outra tabela. Normalmente o banco de dados é configurado para automaticamente associar chaves primárias assim que linhas são inseridas.

relação: Uma área dentro do banco de dados que contém tuplas e atributos. Tipicamente chamada de "tabela".

tupla: Uma entrada em um banco de dados que é um conjunto de atributos. Tipicamente chamado de "linha".

Capítulo 13

Visualizando dados

Até agora, aprendemos a linguagem Python, como utilizá-la para trabalhar com redes e banco de dados para manipular dados.

Neste capítulo, serão apresentadas três aplicações completas que utilizarão todos estes conceitos para gerenciar e visualizar dados. Você pode utilizar estas aplicações como exemplo de código que podem ajudar na solução de problemas reais.

Cada uma das aplicações é um arquivo ZIP que você pode fazer download, extrair para o seu computador e executar.

13.1 Construindo um mapa no Google a partir de dados geocodificados

Neste projeto, utilizaremos a API de geocodificação do Google para obter algumas localizações geográficas informadas pelos usuários de nomes de universidades e colocar os dados em um mapa no Google.

Para iniciar, faça o download da aplicação em:

`www.py4inf.com/code/geodata.zip`

O primeiro problema a ser resolvido é que a API gratuita de geocodificação do Google limita o número de requisições por dia. Se você tiver muitos dados, você pode precisar parar e reiniciar o processo de busca muitas vezes. Nós podemos quebrar o problema em duas fases.

Na primeira fase nós faremos uma "'pesquisa" nos dados do arquivo **where.data** e então ler uma linha por vez, retornando a informação geocodificada do Google e armazenar em um banco de dados **geodata.sqlite**. Antes de efetuar a pesquisa no Google, utilizando a API, para cada localização informada pelo usuário, nós vamos checar para ver se já existe este dado para a localização informada. O banco de dados está funcionando como um "cache" local dos nossos dados de localização para ter certeza que nunca buscaremos no Google duas vezes pelo mesmo dado.

Você pode reiniciar o processo a qualquer hora deletando o arquivo **geodata.sqlite**.

Execute o programa **geoload.py**. Este programa fará a leitura das linhas do arquivo **where.data** e para cada linha checar se o dado já existe no banco de dados. Se nós não tivermos o dado para a localização, será utilizada a API para retornar o dado e armazená-lo no banco de dados.

Aqui está uma simples execução após a coleta de alguns dados no banco de dados:

```
Found in database  Northeastern University
Found in database  University of Hong Kong, ...
Found in database  Technion
Found in database  Viswakarma Institute, Pune, India
Found in database  UMD
Found in database  Tufts University

Resolving Monash University
Retrieving http://maps.googleapis.com/maps/api/
    geocode/json?sensor=false&address=Monash+University
Retrieved 2063 characters {    "results" : [
{u'status': u'OK', u'results': ... }

Resolving Kokshetau Institute of Economics and Management
Retrieving http://maps.googleapis.com/maps/api/
    geocode/json?sensor=false&address=Kokshetau+Inst ...
Retrieved 1749 characters {    "results" : [
{u'status': u'OK', u'results': ... }
...
```

As primeiras cinco execuções já estão no banco de dados e então serão ignoradas. O programa encontra o ponto onde parou e então continua o trabalho recuperando novas informações.

O programa **geoload.py** pode ser parado a qualquer hora, há um contador que você pode utilizar para limitar o número de chamadas para a API de geocodificação a

cada execução. Dado o arquivo **where.data**, que possui algumas centenas de itens, você não consegue ultrapassar o limite diário, mas pode fazer várias execuções em vários dias diferentes para ir pegando aos poucos todos os dados que você precisa.

Uma vez que você tiver alguns dados carregados em **geodata.sqlite**, você pode visualizá-los utilizando o programa **geodump.py**. Este programa lê o banco de dados e escreve no arquivo **where.js** com a localização de latitude e longitude em um formato de código JavaScript.

Segue uma execução do programa **geodump.py**:

```
Northeastern University, ... Boston, MA 02115, USA 42.3396998 -71.08975
Bradley University, 1501 ... Peoria, IL 61625, USA 40.6963857 -89.6160811
...
Technion, Viazman 87, Kesalsaba, 32000, Israel 32.7775 35.0216667
Monash University Clayton ... VIC 3800, Australia -37.9152113 145.134682
Kokshetau, Kazakhstan 53.2833333 69.3833333
...
12 records written to where.js
Open where.html to view the data in a browser
```

O arquivo **where.html** consiste de um HTML e um JavaScript para visualizar um mapa Google. Ele lê os dados mais recentes em **where.js** para pegar os dados a serem visualizados. Aqui está um formato do arquivo **where.js**.

```
myData = [
[42.3396998,-71.08975, 'Northeastern Uni ... Boston, MA 02115'],
[40.6963857,-89.6160811, 'Bradley University, ... Peoria, IL 61625, USA'],
[32.7775,35.0216667, 'Technion, Viazman 87, Kesalsaba, 32000, Israel'],
   ...
];
```

Esta é uma variável JavaScript que contém uma lista de listas. A sintaxe para representação de listas em JavaScript é muito similar ao Python, sendo assim, deve ser familiar para você também.

Simplemente abra o arquivo **where.html** em um browser para ver as localizações. Você pode passar o mouse por cima de cada um dos pontos do mapa para encontrar a localização que a API de geocodificação retornou para uma entrada do usuário. Se você não puder ver qualquer dado quando abrir o arquivo **where.html**, você pode querer checar o console do desenvolvedor (JavaScript) de seu browser e ver se encontra algum erro.

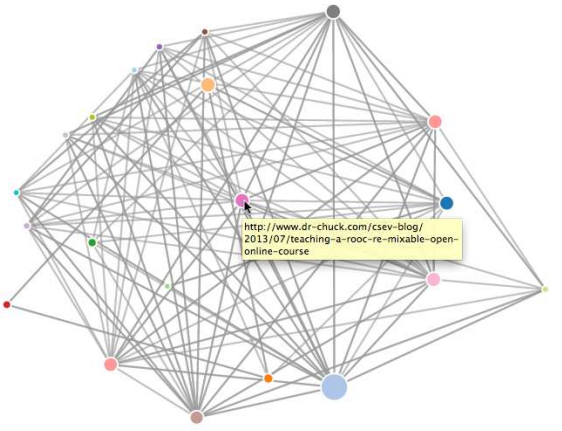
13.2 Visualizando redes e interconexões

Nesta aplicação, vamos realizar algumas das funções de um motor de busca. Primeiramente, vamos extrair um pequeno pedaço da web e rodar uma versão simplificada do algoritmo de page rank do Google para determinar quais páginas estão altamente conectadas, e então visualizar o page rank e a conectividade de nosso

pequeno pedaço da web. Utilizaremos a biblioteca de visualização JavaScript D3 `http://d3js.org/` para produzir a saída da visualização.

Você pode fazer download e extrair esta aplicação de:

`www.py4inf.com/code/pagerank.zip`



O primeiro programa (**spider.py**) vasculha páginas web e grava uma série de páginas no banco de dados (**spider.sqlite**), gravando as ligações entre as páginas. Você pode reiniciar o processo a qualquer momento, deletando o arquivo **spider.sqlite** e reexecutando o **spider.py**.

```
Enter web url or enter: http://www.dr-chuck.com/
['http://www.dr-chuck.com']
How many pages:2
1 http://www.dr-chuck.com/ 12
2 http://www.dr-chuck.com/csev-blog/ 57
How many pages:
```

Neste exemplo de execução, pedimos ao programa para extrair e retornar duas páginas. Se você reiniciar o programa e pedir a ele para obter mais páginas, não irá pegar novamente as mesmas páginas que já estão no banco de dados. Após o restart ele vai sortear randomicamente páginas e começar de lá. Assim, cada execução sucessiva do **spider.py** é um aditivo.

```
Enter web url or enter: http://www.dr-chuck.com/
['http://www.dr-chuck.com']
How many pages:3
3 http://www.dr-chuck.com/csev-blog 57
4 http://www.dr-chuck.com/dr-chuck/resume/speaking.htm 1
5 http://www.dr-chuck.com/dr-chuck/resume/index.htm 13
How many pages:
```

Você pode ter múltiplos pontos de start no mesmo banco de dados, dentro do programa, eles são chamados “webs”. O programa escolhe randomicamente um dos links que ainda não foi visitado através de toda a web como sendo a próxima página a ser visitada.

Se você quer visualizar o conteúdo do arquivo **spider.sqlite**, você pode rodar o programa **spdump.py**, como segue:

```
(5, None, 1.0, 3, u'http://www.dr-chuck.com/csev-blog')
(3, None, 1.0, 4, u'http://www.dr-chuck.com/dr-chuck/resume/speaking.htm')
(1, None, 1.0, 2, u'http://www.dr-chuck.com/csev-blog/')
(1, None, 1.0, 5, u'http://www.dr-chuck.com/dr-chuck/resume/index.htm')
4 rows.
```

Isto mostra o número de links visitados, o antigo page rank, o novo page rank, o id da página, e a url da página. O programa **spdump.py** mostra somente páginas que tem pelo menos um link que já foi visitado.

Uma vez que você tem algumas páginas no banco de dados, você pode rodar o page rank nas páginas usando o programa **sprank.py**. Você apenas diz quantas iterações de páginas devem ser executadas.

```
How many iterations:2
1 0.546848992536
2 0.226714939664
[(1, 0.559), (2, 0.659), (3, 0.985), (4, 2.135), (5, 0.659)]
```

Você pode analisar o banco de dados novamente para ver se o page rank foi atualizado:

```
(5, 1.0, 0.985, 3, u'http://www.dr-chuck.com/csev-blog')
(3, 1.0, 2.135, 4, u'http://www.dr-chuck.com/dr-chuck/resume/speaking.htm')
(1, 1.0, 0.659, 2, u'http://www.dr-chuck.com/csev-blog/')
(1, 1.0, 0.659, 5, u'http://www.dr-chuck.com/dr-chuck/resume/index.htm')
4 rows.
```

Você pode rodar o **sprank.py** quantas vezes quiser, isto irá apenas refinar o page rank cada vez que você executar. Pode até mesmo rodar o **sprank.py** um pequeno número de vezes e então recuperar mais algumas páginas com o **spider.py** e então rodar o **sprank.py** para recuperar os valores do page rank. Um motor de pesquisa geralmente roda os programas de recuperação e rankeamento ao mesmo tempo.

Se você quiser reiniciar os cálculos de page rank sem fazer a extração das páginas web novamente, você pode usar o **spreset.py** e então reiniciar o **sprank.py**.

```
How many iterations:50
1 0.546848992536
2 0.226714939664
3 0.0659516187242
4 0.0244199333
5 0.0102096489546
6 0.00610244329379
...
42 0.000109076928206
43 9.91987599002e-05
44 9.02151706798e-05
45 8.20451504471e-05
46 7.46150183837e-05
47 6.7857770908e-05
```

```
48 6.17124694224e-05
49 5.61236959327e-05
50 5.10410499467e-05
[(512, 0.0296), (1, 12.79), (2, 28.93), (3, 6.808), (4, 13.46)]
```

Para cada iteração do algoritmo de page rank, ele imprime a média de modificações no page rank por página. A rede inicia-se desbalanceada e então os valores do page rank individual mudam com velocidade entre as iterações. Mas em poucas iterações, o page rank coverge. Você deve rodar o **prank.py** por tempo suficiente para que os valores de page rank possam convergir.

Se você quiser visualizar as primeiras páginas no rank, rode o **spjson.py** para ler o banco de dados e escrever os dados dos links com maior pontuação no formato JSON para serem vistos no web browser.

```
Creating JSON output on spider.json...
How many nodes? 30
Open force.html in a browser to view the visualization
```

Você pode visualizar este dado abrindo o arquivo **force.html** em seu web browser. Isto mostra um layout automático de nós e links. Você pode clicar e arrastar qualquer nó e também dar um duplo clique no nó para encontrar a URL que ele representa.

Se você rodar novamente algum script, lembre-se de rodar também o **spjson.py** e pressionar o botão de refresh no browser para ler os novos dados do arquivo **spider.json**.

13.3 Visualizando dados de e-mail

Até este ponto do livro, esperamos que você já tenha se familiarizado com os nossos arquivos de dados, **mbox-short.txt** e **mbox.txt**. Agora é hora de levar a nossa análise de e-mails para um próximo nível.

No mundo real, às vezes você tem que baixar dados de e-mails dos servidores. Isto pode levar um bom tempo e talvez o dado seja inconsistente, com erros de preenchimento, e precisem de muitas limpezas e ajustes. Nesta seção, nós trabalharemos com uma aplicação que é muito complexa e baixa aproximadamente um gigabyte de dados e faz a leitura.

Você pode fazer o download desta aplicação em:

`www.py4inf.com/code/gmane.zip`

Nós utilizaremos dados de um serviço de arquivamento de listas de e-mails gratuita chamado `www.gmane.org`. Este serviço é muito popular em projetos open source porque disponibiliza uma vasta coleção pesquisável de arquivos sobre a atividade de e-mail deles. Eles também tem uma política liberal em relação ao acesso de seus dados através de sua API. Não tem taxas de limites, mas pedem para que você não sobrecarregue os seus serviços e pegue apenas os dados que você precise. Você pode ler os termos e condições do gname nesta página:

`http://gmane.org/export.php`

É muito importante que você faça uso dos dados do gname.org com responsabilidade adicionando delays no acesso ao serviço deles e espalhando tarefas de longa duração em um longo período de tempo. Não abuse deste serviço gratuito, arruinando-o para o resto de nós.

Quando os dados de e-mail do Sakai foram vasculhados usando este software, produziu aproximadamente um Gigabyte de dados e executou muitas vezes durante vários dias. O arquivo **README.txt** no arquivo ZIP acima, pode ter instruções sobre como você pode fazer download de uma cópia preparada previamente através do arquivo **content.sqlite** para a maioria dos corpos dos e-mails Sakai, assim você não precisa coletar os dados por cinco dias apenas para rodar seus programas. Se você baixar conteúdos pré preparados, você ainda pode rodar o processo de coleta para pegar mensagens mais recentes.

O primeiro passo é fazer uma coleta no repositório gname. A URL base está fixa dentro do arquivo **gmane.py** e está apontando para a lista de desenvolvedores do Sakai. Você pode coletar outro repositório apenas mudando a url base. Certifique-se de deletar o arquivo **content.sqlite** se você trocar a url base.

O arquivo **gmane.py** funciona como um cache spider responsável, executando lentamente e retornando uma mensagem de e-mail por segundo, não sendo bloqueado

desta forma pelo site gname. Ele armazena todos os seus dados em um banco de dados e pode ser interrompido e reiniciado quantas vezes forem necessárias. Pode levar muitas horas para baixar todos os dados. Você pode precisar reiniciar muitas vezes.

Aqui está uma execução do **gmane.py** retornando as últimas cinco mensagens da lista de desenvolvedores do Sakai:

```
How many messages:10
http://download.gmane.org/gmane.comp.cms.sakai.devel/51410/51411 9460
    nealcaidin@sakaifoundation.org 2013-04-05 re: [building ...
http://download.gmane.org/gmane.comp.cms.sakai.devel/51411/51412 3379
    samuelgutierrezjimenez@gmail.com 2013-04-06 re: [building ...
http://download.gmane.org/gmane.comp.cms.sakai.devel/51412/51413 9903
    da1@vt.edu 2013-04-05 [building sakai] melete 2.9 oracle ...
http://download.gmane.org/gmane.comp.cms.sakai.devel/51413/51414 349265
    m.shedid@elraed-it.com 2013-04-07 [building sakai] ...
http://download.gmane.org/gmane.comp.cms.sakai.devel/51414/51415 3481
    samuelgutierrezjimenez@gmail.com 2013-04-07 re: ...
http://download.gmane.org/gmane.comp.cms.sakai.devel/51415/51416 0

Does not start with From
```

O programa escaneia o **content.sqlite** e inicia a execução a partir da primeira mensagem que ainda não foi processada. Ele continua coletando até chegar ao número desejado de mensagens ou atingir uma página que possuir uma mensagem fora do padrão.

Às vezes o `gmane.org` perde alguma mensagem. Seja pelos administradores que deletaram uma mensagem ou então porque se perdeu mesmo. Se seu programa para, e parece que ele perdeu alguma mensagem, vá para o administrador SQLite e adicione uma linha com o id que está faltando, deixando todos os outros campos em branco e reinicie o **gmane.py**. Isto irá liberar o seu programa para continuar a execução. Estas mensagens vazias serão ignoradas na próxima fase do processo.

Uma coisa legal é que uma vez que você coletou todas as mensagens e tem elas dentro do **content.sqlite**, você pode rodar o **gmane.py** novamente para pegar as últimas mensagens novas enviadas para a lista.

Os dados do **content.sqlite** são um pouco crus, com um modelo de dados ineficiente e não comprimido. Isto é intencional, pois permite a você olhar o conteúdo do **content.sqlite** no SQLite Manager para depurar problemas que ocorreram no processo de coleta. Seria uma má ideia rodar qualquer tipo de consulta neste banco de dados, pois poderia demorar.

O segundo processo é rodar o programa **gmodel.py**. Este programa lê os dados crus do **content.sqlite** e produz uma versão bem modelada dos dados no arquivo **index.sqlite**. Este arquivo será muito menor (quase 10 vezes menor) do que o **index.sqlite**, porque comprime o corpo e o cabeçalho do texto.

Cada vez que o **gmodel.py** roda, ele deleta e refaz o arquivo **index.sqlite**, permitindo a você ajustar os parâmetros e editar as tabelas de mapeamento no **con-**

tent.sqlite para ajustar o processo de limpeza dos dados. Esta é uma amostra da execução do **gmodel.py**. Ele imprime uma linha por vez, 250 mensagens de e-mail são processadas, assim você pode ver algum progresso acontecendo, como este programa pode rodar por um longo tempo, é muito fácil conseguir um arquivo de e-mail de um Gigabyte de dados.

```
Loaded allsenders 1588 and mapping 28 dns mapping 1
1 2005-12-08T23:34:30-06:00 ggolden22@mac.com
251 2005-12-22T10:03:20-08:00 tpamsler@ucdavis.edu
501 2006-01-12T11:17:34-05:00 lance@indiana.edu
751 2006-01-24T11:13:28-08:00 vrajgopalan@ucmerced.edu
...
```

O programa **gmodel.py** trata um número de tarefas de limpeza de dados.

Nomes de domínios são truncados para dois níveis para .com, .org, .edu, e .net. Outros nomes de domínios são truncados para três níveis. Assim si.umich.edu se transforma em umich.edu e caret.cam.ac.uk se transforma em cam.ac.uk. Endereços de e-mail também são forçados para minúsculo, seguem alguns endereços do @gmane.org, por exemplo:

```
arwhyte-63aXycvo3TyHXe+LvDLADg@public.gmane.org
```

são convertidos para endereços reais a qualquer hora que ocorrer um encontro real de endereço de e-mail em qualquer lugar no corpo da mensagem.

No banco de dados **content.sqlite**, há duas tabelas que permitem a você mapear ambos os nomes de domínio e os endereços de e-mail individuais que mudam ao longo do tempo da lista. Por exemplo, Steve Githens usou os seguintes endereços de e-mail conforme foi mudando de emprego na lista de desenvolvedores Sakai:

```
s-githens@northwestern.edu
sgithens@cam.ac.uk
swgithen@mtu.edu
```

Nós podemos adicionar duas entradas na tabela de mapeamento em **content.sqlite**, assim o **gmodel.py** irá mapear todos os três para um único endereço:

```
s-githens@northwestern.edu ->  swgithen@mtu.edu
sgithens@cam.ac.uk -> swgithen@mtu.edu
```

Você pode criar entradas similares na tabela de mapeamento DNS se possuir múltiplos nomes DNS mapeados para um DNS simples. O seguinte mapeamento foi adicionado para os dados Sakai:

```
iupui.edu -> indiana.edu
```

assim todas as contas de vários campus da Universidade de Indiana são atreladas juntas.

Você pode reexecutar o **gmodel.py** de novo e de novo conforme olhar para os dados e adicionar mapeamentos para tornar os dados mais limpos. Quando você

acabar, terá uma agradável versão indexada dos e-mails em **index.sqlite**. Este é o arquivo para se fazer análise dos dados. Com este arquivo, a análise dos dados será muito rápida.

A primeira, análise simples dos dados é para determinar ”quem enviou mais e-mails?”e ”qual organização enviou mais e-mails?”Isto é feito usando o **gbasic.py**:

```
How many to dump? 5
Loaded messages= 51330 subjects= 25033 senders= 1584

Top 5 Email list participants
steve.swinsburg@gmail.com 2657
azeckoski@unicon.net 1742
ieb@tfd.co.uk 1591
csev@umich.edu 1304
david.horwitz@uct.ac.za 1184

Top 5 Email list organizations
gmail.com 7339
umich.edu 6243
uct.ac.za 2451
indiana.edu 2258
unicon.net 2055
```

Observe quão mais rápido o **gbasic.py** executa quando comparado ao **gmane.py** ou até mesmo ao **gmodel.py**. Eles todos trabalham no mesmo dado, mas o **gbasic.py** está usando dados normalizados e comprimidos em **index.sqlite**. Se você tiver muitos dados para gerenciar, um processo multi-passos assim como estes nesta aplicação podem levar mais tempo para serem desenvolvidos, mas irão economizar muito tempo quando você começar a explorar e visualizar os seus dados.

Você pode produzir uma simples visualização da frequência das palavras nas linhas do arquivo **gword.py**:

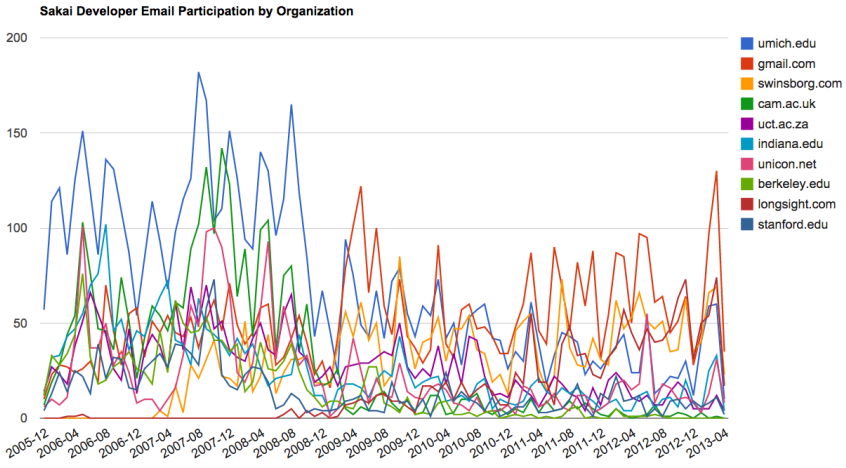
```
Range of counts: 33229 129
Output written to gword.js
```

Isto produz o arquivo **gword.js**, o qual você pode visualizar usando o **gword.htm** para produzir uma nuvem de palavras similares àquelas no início desta seção.

Uma segunda visualização foi produzida pelo **gline.py**. Ele computa a participação das organizações por e-mail ao longo do tempo.

```
Loaded messages= 51330 subjects= 25033 senders= 1584
Top 10 Oranizations
['gmail.com', 'umich.edu', 'uct.ac.za', 'indiana.edu',
'unicon.net', 'tfd.co.uk', 'berkeley.edu', 'longsight.com',
'stanford.edu', 'ox.ac.uk']
Output written to gline.js
```

Sua saída é escrita para **gline.js**, o qual é visualizada usando **gline.htm**.



Esta é uma aplicação relativamente complexa e sofisticada e têm características para se fazer uma coleta real de dados, limpeza e visualização.

Capítulo 14

Automação de tarefas comuns no seu computador

Temos lido dados de arquivos, redes, serviços e banco de dados. Python também pode navegar através de todas as pastas e diretórios no seu computador e ler os arquivos também.

Neste capítulo, nós iremos escrever programas que analisam o seu computador e executam algumas operações em seus arquivos. Arquivos são organizados em diretórios (também chamados de pastas). Scripts Python simples podem fazer o trabalho de tarefas simples que devem ser feitas em centenas ou milhares de arquivos espalhados por uma árvore de diretórios ou todo o seu computador.

Para navegar através de todos os diretórios e arquivos em uma árvore nós utilizamos `os.walk` e um laço de repetição `for`. Isto é similar ao comando `open` e nos permite escrever um laço de repetição para ler o conteúdo de um arquivo, `socket` nos permite escrever um laço de repetição para ler o conteúdo de uma conexão e `urllib` nos permite abrir um documento web e navegar por meio de um laço de repetição no seu conteúdo.

14.1 Nomes e caminhos de arquivos

Todo programa em execução tem um "diretório atual" que é o diretório padrão para a maioria das operações. Por exemplo, quando você abre um arquivo para leitura, o Python o procura no diretório atual.

O módulo `os` disponibiliza funções para trabalhar com arquivos e diretórios (`os` do inglês "operating system"que significa sistema operacional). `os.getcwd` retorna o nome do diretório atual:

```
>>> import os
>>> cwd = os.getcwd()
```

```
>>> print cwd
/Users/csev
```

`cwd` significa **diretório atual de trabalho**. O resultado neste exemplo é `/Users/csev`, que é o diretório home do usuário chamado `csev`.

Uma string `cwd` que identifica um arquivo é chamado de path. Um **caminho relativo** é relativo ao diretório atual (corrente); Um **caminho absoluto** tem inicio no diretório raiz do sistema de arquivos.

Os caminhos que temos visto até agora são nomes de arquivos simples, por isso são relativos ao diretório atual. Para encontrar o caminho absoluto de um arquivo, você pode usar `os.path.abspath`:

```
>>> os.path.abspath('memo.txt')
'/Users/csev/memo.txt'
```

`os.path.exists` verifica se um determinado arquivo existe:

```
>>> os.path.exists('memo.txt')
True
```

Se existir, `os.path.isdir` verifica se é um diretório:

```
>>> os.path.isdir('memo.txt')
False
>>> os.path.isdir('music')
True
```

Da mesma forma, `os.path.isfile` verifica se é um arquivo.

`os.listdir` retorna uma lista com os arquivos (e outros diretórios) do diretório informado:

```
>>> os.listdir(cwd)
['musicas', 'fotos', 'memo.txt']
```

14.2 Exemplo: Limpando um diretório de fotos

Há algum tempo atrás, desenvolvi um pequeno software tipo Flickr que recebe fotos do meu celular e armazena essas fotos no meu servidor. E escrevi isto antes do Flickr existir e continuo usando por que eu quero manter copias das minhas imagens originais para sempre.

Eu também gostaria de enviar uma simples descrição numa mensagem MMS ou como um título de uma mensagem de e-mail. Eu armazenei essas mensagens em um arquivo de texto no mesmo diretório do arquivo das imagens. Eu criei uma estrutura de diretórios baseada no mês, ano, dia e hora no qual a foto foi tirada, abaixo um exemplo de nomenclatura para uma foto e sua descrição:

```
./2006/03/24-03-06_2018002.jpg
./2006/03/24-03-06_2018002.txt
```

Após sete anos, eu tenho muitas fotos e legendas. Ao longo dos anos como eu troquei de celular, algumas vezes, meu código para extrair a legenda para uma mensagem quebrou e adicionou um bando de dados inúteis no meu servidor ao invés de legenda.

Eu queria passar por estes arquivos e descobrir quais dos arquivos texto eram realmente legendas e quais eram lixo e, em seguida, apagar os arquivos que eram lixo. A primeira coisa a fazer foi conseguir um simples inventário dos arquivos texto que eu tinha em uma das subpastas usando o seguinte programa:

```
import os
count = 0
for (dirname, dirs, files) in os.walk('.'):
   for filename in files:
       if filename.endswith('.txt') :
           count = count + 1
print 'Files:', count

python txtcount.py
Files: 1917
```

O segredo para um código tão pequeno é a utilização da biblioteca `os.walk` do Python. Quando nós chamamos `os.walk` e inicializamos um diretório, ele ”caminha”através de todos os diretórios e subdiretórios recursivamente. O caractere . indica para iniciar no diretório corrente e navegar para baixo. Assim que encontra cada diretório, temos três valores em uma tupla no corpo do laço de repetição `for`. O primeiro valor é o diretório corrente, o segundo é uma lista de sub-diretórios e o terceiro valor é uma lista de arquivos no diretório corrente.

Nós não temos que procurar explicitamente dentro de cada diretório por que nós podemos contar com `os.walk` para visitar eventualmente todas as pastas mas, nós queremos procurar em cada arquivo, então, escrevemos um simples laço de repetição `for` para examinar cada um dos arquivos no diretório corrente. Vamos verificar se cada arquivo termina com ”.txt”e depois contar o número de arquivos através de toda a árvore de diretórios que terminam com o sufixo ”.txt”.

Uma vez que nós temos uma noção da quantidade de arquivos terminados com ”.txt”, a próxima coisa a se fazer é tentar determinar automaticamente no Python quais arquivos são maus e quais são bons. Para isto, escreveremos um programa simples para imprimir os arquivos e seus tamanhos.

```
import os
from os.path import join
for (dirname, dirs, files) in os.walk('.'):
   for filename in files:
       if filename.endswith('.txt') :
           thefile = os.path.join(dirname,filename)
           print os.path.getsize(thefile), thefile
```

Agora, em vez de apenas contar os arquivos, criamos um nome de arquivo concatenando o nome do diretório com o nome do arquivo dentro do diretório usando

`os.path.join`. É importante usar o `os.path.join` para concatenar a sequência de caracteres por que no Windows usamos a barra invertida para construir os caminhos de arquivos e no Linux ou Apple nós usamos a barra (/) para construir o caminho do arquivo. O `os.path.join` conhece essas diferenças e sabe qual sistema esta rodando dessa forma, faz a concatenação mais adequada considerando o sistema. Então, o mesmo código em Python roda tando no Windows quanto em sistemas tipo Unix.

Uma vez que temos o nome completo do arquivo com o caminho do diretório, nós usamos o utilitário `os.path.getsize` para pegar e imprimir o tamanho, produzindo a seguinte saída.

```
python txtsize.py
...
18 ./2006/03/24-03-06_2303002.txt
22 ./2006/03/25-03-06_1340001.txt
22 ./2006/03/25-03-06_2034001.txt
...
2565 ./2005/09/28-09-05_1043004.txt
2565 ./2005/09/28-09-05_1141002.txt
...
2578 ./2006/03/27-03-06_1618001.txt
2578 ./2006/03/28-03-06_2109001.txt
2578 ./2006/03/29-03-06_1355001.txt
...
```

Analisando a saída, nós percebemos que alguns arquivos são bem pequenos e muitos dos arquivos são bem grandes e com o mesmo tamanho (2578 e 2565). Quando observamos alguns desses arquivos maiores manualmente, parece que os arquivos grandes são nada mais que HTML genérico idênticos que vinham de e-mails enviados para meu sistema a partir do meu próprio telefone:

```
<html>
        <head>
                <title>T-Mobile</title>
...
```

Espiando o conteúdo destes arquivos, parece que não há informações importantes, então provavelmente podemos eliminá-los.

Mas antes de excluir os arquivos, vamos escrever um programa para procurar por arquivos que possuem mais de uma linha e exibir o conteúdo do arquivo. Não vamos nos incomodar mostrando os arquivos que são exatamente 2578 ou 2565 caracteres, pois sabemos que estes não têm informações úteis.

Assim podemos escrever o seguinte programa:

```
import os
from os.path import join
for (dirname, dirs, files) in os.walk('.'):
   for filename in files:
       if filename.endswith('.txt') :
```

```
thefile = os.path.join(dirname,filename)
size = os.path.getsize(thefile)
if size == 2578 or size == 2565:
    continue
fhand = open(thefile,'r')
lines = list()
for line in fhand:
    lines.append(line)
fhand.close()
if len(lines) > 1:
     print len(lines), thefile
     print lines[:4]
```

Nós usamos um `continue` para ignorar arquivos com dois ”Maus tamanhos”, então, abrimos o resto dos arquivos e lemos as linhas do arquivo em uma lista Python, se o arquivo tiver mais que uma linha nós imprimimos a quantidade de linhas e as primeiras três linhas do arquivo.

Parece que filtrando esses dois tamanhos de arquivo ruins, e supondo que todos os arquivos de uma linha estão corretos, nós temos abaixo alguns dados bastante limpos:

```
python txtcheck.py
3 ./2004/03/22-03-04_2015.txt
['Little horse rider\r\n', '\r\n', '\r']
2 ./2004/11/30-11-04_1834001.txt
['Testing 123.\n', '\n']
3 ./2007/09/15-09-07_074202_03.txt
['\r\n', '\r\n', 'Sent from my iPhone\r\n']
3 ./2007/09/19-09-07_124857_01.txt
['\r\n', '\r\n', 'Sent from my iPhone\r\n']
3 ./2007/09/20-09-07_115617_01.txt
...
```

Mas existe um ou mais padrões chatos de arquivo: duas linhas brancas seguidas por uma linha que diz ”Sent from my iPhone”que são exceção em meus dados. Então, fizemos a seguinte mudança no programa para lidar com esses arquivos também.

```
lines = list()
for line in fhand:
    lines.append(line)
if len(lines) == 3 and lines[2].startswith('Sent from my iPhone'):
    continue
if len(lines) > 1:
     print len(lines), thefile
     print lines[:4]
```

Nós simplesmente verificamos se temos um arquivo com três linhas, e se a terceira linha inicia-se com o texto específico, então nós o pulamos. Agora quando rodamos o programa, vemos apenas quatro arquivos multi-linha restantes e todos esses arquivos parecem fazer sentido:

```
python txtcheck2.py
3 ./2004/03/22-03-04_2015.txt
['Little horse rider\r\n', '\r\n', '\r']
2 ./2004/11/30-11-04_1834001.txt
['Testing 123.\n', '\n']
2 ./2006/03/17-03-06_1806001.txt
['On the road again...\r\n', '\r\n']
2 ./2006/03/24-03-06_1740001.txt
['On the road again...\r\n', '\r\n']
```

Se você olhar para o padrão global deste programa, nós refinamos sucessivamente como aceitamos ou rejeitamos arquivos e uma vez encontrado um padrão que era ”ruim”nós usamos `continue` para ignorar os maus arquivos para que pudéssemos refinar nosso código para encontrar mais padrões que eram ruins.

Agora estamos nos preparando para excluir os arquivos, nós vamos inverter a lógica e ao invés de imprimirmos os bons arquivos, vamos imprimir os maus arquivos que estamos prestes a excluir.

```
import os
from os.path import join
for (dirname, dirs, files) in os.walk('.'):
   for filename in files:
       if filename.endswith('.txt') :
           thefile = os.path.join(dirname,filename)
           size = os.path.getsize(thefile)
           if size == 2578 or size == 2565:
               print 'T-Mobile:',thefile
               continue
           fhand = open(thefile,'r')
           lines = list()
           for line in fhand:
               lines.append(line)
           fhand.close()
           if len(lines) == 3 and lines[2].startswith('Sent from my iPhone'):
               print 'iPhone:', thefile
               continue
```

Podemos ver agora uma lista de possíveis arquivos que queremos apagar e por quê esses arquivos são eleitos a exclusão. O Programa produz a seguinte saída:

```
python txtcheck3.py

...
T-Mobile: ./2006/05/31-05-06_1540001.txt
T-Mobile: ./2006/05/31-05-06_1648001.txt
iPhone: ./2007/09/15-09-07_074202_03.txt
iPhone: ./2007/09/15-09-07_144641_01.txt
iPhone: ./2007/09/19-09-07_124857_01.txt
...
```

Podemos verificar pontualmente esses arquivos para nos certificar que não inserimos um bug em nosso programa ou talvez na nossa lógica, pegando arquivos que

não queríamos. Uma vez satisfeitos de que esta é a lista de arquivos que queremos excluir, faremos a seguinte mudança no programa:

```
            if size == 2578 or size == 2565:
                print 'T-Mobile:',thefile
                os.remove(thefile)
                continue
...
            if len(lines) == 3 and lines[2].startswith('Sent from my iPhone'):
                print 'iPhone:', thefile
                os.remove(thefile)
                continue
```

Nesta versão do programa, iremos fazer ambos, imprimir o arquivo e remover os arquivos ruins com `os.remove`

```
python txtdelete.py
T-Mobile: ./2005/01/02-01-05_1356001.txt
T-Mobile: ./2005/01/02-01-05_1858001.txt
...
```

Apenas por diversão, rodamos o programa uma segunda vez e o programa não irá produzir nenhuma saída desde que os arquivos ruins não existam.

Se rodar novamente `txtcount.py` podemos ver que removemos 899 arquivos ruins:

```
python txtcount.py
Files: 1018
```

Nesta seção, temos seguido uma sequência onde usamos o Python primeiro para navegar através dos diretórios e arquivos procurando padrões. Usamos o Python devagar para ajudar a determinar como faríamos para limpar nosso diretório. Uma vez descoberto quais arquivos são bons e quais não são, nós usamos o Python para excluir os arquivos e executar a limpeza.

O problema que você precisa resolver pode ser bastante simples precisando procurar pelos nomes dos arquivos, ou talvez você precise ler cada arquivo, procurando por padrões dentro dos mesmos, às vezes você precisa ler o conteúdo dos arquivos fazendo alguma mudança em alguns deles, seguindo algum tipo de critério. Todos estes são bastante simples uma vez que você entenda como tt os.walk e outros utilitários `os` podem ser usados.

14.3 Argumentos de linha de comando

Nos capítulos anteriores tivemos uma série de programas que solicitavam por um nome de arquivo usando `raw_input` e então, liam os dados de um arquivo e processavam os dados, como a seguir:

```
nome = raw_input('Informe o arquivo:')
handle = open(nome, 'r')
texto = handle.read()
...
```

Nós podemos simplificar este programa um pouco pegando o nome do arquivo a partir de um comando quando iniciamos o Python. Até agora nós simplesmente executamos nossos programas em Python e respondemos a solicitação como segue:

```
python words.py
Informe o arquivo: mbox-short.txt
...
```

Nós podemos colocar strings adicionais depois do nome do arquivo Python na linha de comando e acessá-los de dentro de um programa Python. Eles são chamados **argumentos de linha de comando**. Aqui está um simples programa que demonstra a leitura de argumentos a partir de uma linha de comando:

```
import sys
print 'Contagem:', len(sys.argv)
print 'Tipo:', type(sys.argv)
for arg in sys.argv:
   print 'Argumento:', arg
```

Os conteúdos de `sys.argv` são uma lista de strings onde a primeira string contém o nome do programa Python e as outras são argumentos na linha de comando após o nome do arquivo Python.

O seguinte mostra nosso programa lendo uma série de argumentos de linha de comando de uma linha de comando:

```
python argtest.py ola alguem
Contagem: 3
Tipo: <type 'list'>
Argumento: argtest.py
Argumento: ola
Argumento: alguem
```

Há três argumentos que são passados ao nosso programa como uma lista de três elementos. O primeiro elemento da lista é o nome do arquivo (argtest.py) e os outros são os dois argumentos de linha de comando após o nome do arquivo.

Nós podemos reescrever nosso programa para ler o arquivo, obtendo o nome do arquivo a partir do argumento de linha de comando, como segue:

```
import sys

name = sys.argv[1]
handle = open(name, 'r')
text = handle.read()
print name, 'is', len(text), 'bytes'
```

Nós pegamos o segundo argumento da linha de comando, que contém o nome do arquivo (pulando o nome do programa na entrada `[0]`). Nós abrimos o arquivo e lemos seu conteúdo, como segue:

```
python argfile.py mbox-short.txt
mbox-short.txt is 94626 bytes
```

Usar argumentos de linha de comando como entrada, torna o seu programa Python fácil de se reutilizar, especialmente quando você somente precisa passar uma ou duas strings.

14.4 Pipes

A maioria dos sistemas operacionais oferecem uma interface de linha de comando, conhecido também como **shell**. Shells normalmente normalmente disponibilizam comandos para navegar entre arquivos do sistema e executar aplicações. Por exemplo, no Unix, você pode mudar de diretório com `cd`, mostrar na tela o conteúdo de um diretório com `ls` e rodar um web browser digitando (por exemplo) `firefox`.

Qualquer programa que consiga rodar a partir do shell também pode ser executado a partir do Python usando um **pipe**. Um pipe é um objeto que representa um processo em execução.

Por exemplo, o comando Unix [1] `ls -l` normalmente mostra o conteúdo do diretório corrente (no modo detalhado). Você pode rodar `ls` com `os.open`:

```
>>> cmd = 'ls -l'
>>> fp = os.popen(cmd)
```

Um argumento é uma string que contém um comando shell. O valor de retorno é um ponteiro para um arquivo que se comporta exatamente como um arquivo aberto. Você pode ler a saída do processo `ls` uma linha de cada vez com o comando `readline` ou obter tudo de uma vez com o comando `read`:

```
>>> res = fp.read()
```

Quando terminar, você fecha o pipe como se fosse um arquivo:

```
>>> stat = fp.close()
>>> print stat
None
```

O valor de retorno é o status final do processo `ls`; `None` significa que ele terminou normalmente (sem erros).

[1] Ao usar pipes para interagir com comandos do sistema operacional como `ls`, é importante saber qual sistema operacional você está usando e executar somente comandos pipe que são suportados pelo seu sistema operacional.

14.5 Glossário

absolute path: Uma string que descreve onde um arquivo ou diretório é armazenado, começando desde o "topo da árvore de diretórios" de modo que ele pode ser usado para acessar o arquivo ou diretório, independentemente do diretório de trabalho corrente.

checksum: Ver também **hashing**. O termo "checksum" vem da necessidade de se verificar se os dados corromperam durante o envio pelo rede ou quando gravados em um meio de backup. Quando os dados são gravados ou enviados, o sistema emissor calcula o checksum e também o envia. Quando o dado foi completamente lido ou recebido, o sistema receptor calcula novamente o checksum com base nos dados recebidos e os compara com o checksum recebido. Se os checksum's não corresponderem, devemos assumir que os dados estão corrompidos, uma vez que já finalizou a transmissão. checksum

command-line argument: Parâmetros na linha de comando após o nome do arquivo Python.

current working directory: O diretório corrente no qual você está. Você pode mudar seu diretório de trabalho usando o comando `cd`, disponível na maioria dos sistemas operacionais em sua interface de linha de comando. Quando você abre um arquivo em Python usando apenas o nome do arquivo, sem o caminho, o arquivo deve estar no diretório de trabalho atual, onde está executando o programa.

hashing: Leitura através de uma grande quantidade de dados, produzindo um checksum global para os dados. As melhores funções hash produzem muito poucas "colisões", que é quando você passa diferentes dados para a função hash e recebe de volta o mesmo hash. MD5, SHA1 e SHA256 são exemplos de funções hash mais usadas.

pipe: Um pipe é uma conexão com um programa em execução. Usando um pipe, você pode escrever um programa para enviar os dados para outro programa ou receber dados a partir desse programa. Um pipe é semelhante a um **socket**, com exceção de que o pipe só pode ser usado para conectar programas em execução no mesmo computador (ou seja, não através de uma rede). pipe

relative path: Uma string que descreve onde um arquivo ou diretório é armazenado em relação ao diretório de trabalho atual.

shell: Uma interface de linha de comando para um sistema operacional. Também chamado em alguns sistemas operacionais de "terminal". Nesta interface, você digita um comando com parâmetros em uma única linha e pressiona "enter"para executar o comando.

walk: Um termo que usamos para descrever a noção de visitar uma árvore inteira de diretórios e sub-diretórios, até que tenhamos visitado todos eles. Nós chamamos isso de "caminhar pela árvore de diretórios".

14.6 Exercícios

Exercício 14.1 Numa grande coleção de arquivos MP3, pode existir mais de uma cópia de um mesmo som, armazenado em diferentes diretórios ou com diferentes nomes de arquivo. O objetivo deste exercício é procurar por essas duplicatas.

1. Escreva um programa que caminhe no diretório e em todos os seus subdiretórios, procurando por todos os arquivos com o sufixo `.mp3` e liste o par de arquivos com o mesmo tamanho. Dica: Use um dicionário onde a chave seja o tamanho do arquivo do `os.path.getsize` e o valor seja o nome do caminho concatenado com o nome do arquivo. Conforme você for encontrando cada arquivo, verifique se já tem um arquivo que tem o mesmo tamanho do arquivo atual. Se assim for, você tem um arquivo duplicado, então imprima o tamanho e os nomes dos dois arquivos (um a partir do hash e o outro a partir do arquivo que você está olhando no momento).

2. Adaptar o programa anterior para procurar arquivos com conteúdo duplicado usando um hash ou um **checksum**. Por exemplo, MD5 (Message-Digest algorithm 5) recebe uma "mensagem" grande e retorna um "checksum" de 128 bits. A probabilidade de que dois arquivos com diferentes conteúdos retornem o mesmo checksum é muito pequena.

   Você pode ler sobre o MD5 em wikipedia.org/wiki/Md5. O seguinte trecho de código abre um arquivo, o lê, e calcula o seu checksum.

   ```
   import hashlib
   ...
             fhand = open(thefile,'r')
             data = fhand.read()
             fhand.close()
             checksum = hashlib.md5(data).hexdigest()
   ```

   Você deve criar um dicionário onde o checksum é a chave e o nome do arquivo é o valor. Quando você calcular um checksum e ele já existir no dicionário como uma chave, então você terá dois arquivos duplicados. Então imprima o arquivo existente no dicionário e o arquivo que você acabou de ler. Aqui estão algumas saídas de uma execução sob uma pasta com arquivos de imagens.

   ```
   ./2004/11/15-11-04_0923001.jpg ./2004/11/15-11-04_1016001.jpg
   ./2005/06/28-06-05_1500001.jpg ./2005/06/28-06-05_1502001.jpg
   ./2006/08/11-08-06_205948_01.jpg ./2006/08/12-08-06_155318_02.jpg
   ```

   Aparentemente, eu às vezes envio a mesma foto mais de uma vez ou faço uma cópia de uma foto de vez em quando sem excluir a original.

Apêndice A

Programando Python no Windows

Neste apêndice, demonstraremos os passos para você conseguir rodar Python no Windows. Existem muitos jeitos de se fazer e a ideia aqui é escolher um modo que simplifique o processo.

Primeiro, você precisa instalar um editor de programas. Você pode não querer usar o Notepad ou o editor Microsoft Word para editar programas Python. Programas devem ser arquivos texto simples, então você precisará de um bom editor de arquivos texto.

Nosso editor recomendado para Windows é o NotePad++, que pode ser instalado a partir daqui:

`https://notepad-plus-plus.org/`

Faça o download da versão mais recente do Python 2 a partir do site oficial `www.python.org`

`https://www.python.org/downloads/`

Uma vez que você instalou o Python, você deve ter uma nova pasta em seu computador, tal como `C:\Python27`.

Para criar um programa Python, execute o NotePad++ a partir do seu menu iniciar e salve o arquivo com o sufixo ".py". Para este exercício, coloque uma pasta na sua Área de Trabalho chamada `py4inf`. É melhor utilizar nomes de pasta curtos e não ter nenhum tipo de espaço, acento ou caractere especial, seja na pasta ou no nome do arquivo.

Vamos fazer o nosso primeiro programa Python:

```
print 'Hello Chuck'
```

Com exceção que você deve trocar para o seu nome. Salve o arquivo em: `Desktop\py4inf\prog1.py`.

Então abra a janela de linha de comando. Isto varia de acordo com a versão do Windows que você utiliza:

- Windows Vista e Windows 7: Pressione **Iniciar** e então na janela de pesquisa que se abre, digite a palavra `command` e pressione enter.
- Windows XP: Pressione **Iniciar**, e **Executar**, e então digite `cmd` na caixa de diálogo e pressione **OK**.

Você verá uma janela de texto com um prompt que te mostrará em qual pasta você se encontra.

Windows Vista and Windows-7: `C:\Users\csev`
Windows XP: `C:\Documents and Settings\csev`

Este é o seu "diretório do usuário". Agora nós precisamos caminhar para a pasta onde você salvou o seu programa Python utilizando os seguintes comandos:

```
C:\Users\csev\> cd Desktop
C:\Users\csev\Desktop> cd py4inf
```

Então digite

```
C:\Users\csev\Desktop\py4inf> dir
```

para listar os seus arquivos. Você verá o `prog1.py` quando você digitar o comando `dir`.

Para executar o seu programa, simplesmente digite o nome do seu arquivo no prompt de comando e pressione enter.

```
C:\Users\csev\Desktop\py4inf> prog1.py
Hello Chuck
C:\Users\csev\Desktop\py4inf>
```

Você pode editar o arquivo no NotePad++, salvar, e então voltar para a linha de comando e executar o seu programa de novo apenas digitando o nome do arquivo na linha de comando.

Se você estiver confuso na janela de comando, apenas feche e abra uma nova.

Dica: Você pode pressionar a "seta para cima" na linha de comando para rolar e executar o último comando executado anteriormente.

Você também deve olhar nas preferências do NotePad++ e configurar para expandir os caracteres tab para serem quatro espaços. Isto irá te ajudar bastante e não enfrentar erros de identação.

Você pode encontrar maiores informações sobre editar e executar programas Python em `www.py4inf.com`.

Apêndice B

Python Programming on Macintosh

Apêndice C

Programação Python no Macintosh

Neste apêndice, apresentaremos uma série de passos para que você possa executar o Python no Macintosh. Uma vez que Python já está incluso no Sistema Operacional Macintosh, só precisamos aprender como editar os arquivos Python e executar programas Python no terminal.

Existem várias abordagens que você pode adotar para edição e execução dos programas Python, e esta é somente umas das formas que encontramos, por ser muito simples.

Primeiro, você precisará instalar um editor de textos. Você não vai querer utilizar o TextEdit ou o Microsoft Word para editar os programas Python. Os arquivos de programas devem estar em texto-puro então você precisará de um editor que é bom em editar arquivos de texto.

Recomendamos para Macintosh o editor TextWrangler que pode ser baixado e instalado através do seguinte endereço:

`http://www.barebones.com/products/TextWrangler/`

Para criar um programa Python, execute **TextWrangler** a partir da sua pasta de **Aplicações**.

Vamos fazer nosso primeiro programa em Python:

```
print 'Hello Chuck'
```

A única alteração que você deve fazer é referente ao nome, troque **Chuck** pelo seu nome. Salve o arquivo em uma pasta chamada `py4inf` em seu Desktop. É melhor manter os nomes das suas pastas pequenos e sem espaços, seja nas pastas ou nos nomes dos arquivos. Uma vez que você tenha criado a pasta, salve o arquivo dentro dela `Desktop\py4inf\prog1.py`.

Então, execute o programa através do **Terminal**. A forma mais fácil de fazer isto é utilizando o Spotlight (a lupa) no lado superior direito da sua tela, e escreva "terminal", e execute a aplicação.

Você vai começar no seu diretório "home". Você pode ver o seu diretório corrente (que você se encontra) através digitando o comando `pwd` na janela do terminal

```
67-194-80-15:~ csev$ pwd
/Users/csev
67-194-80-15:~ csev$
```

Você deve estar na pasta que contém seu arquivo de programa Python para executá-lo. Utilize o comando `cd` para entrar em uma nova pasta, e depois o comando `ls` para listar os arquivos na pasta.

```
67-194-80-15:~ csev$ cd Desktop
67-194-80-15:Desktop csev$ cd py4inf
67-194-80-15:py4inf csev$ ls
prog1.py
67-194-80-15:py4inf csev$
```

Para executar o programa, digite o comando `python` seguido do nome do seu arquivo na linha de comando e pressione enter.

```
67-194-80-15:py4inf csev$ python prog1.py
Hello Chuck
67-194-80-15:py4inf csev$
```

Você pode editar o arquivo no TextWrangler, salvá-lo, e então voltar para a linha de comando e executar o programa novamente, digitando o nome do arquivo na linha de comando.

Se você ficar confuso com a linha de comando, apenas feche-a e abra uma nova janela.

Dica: Você também pode pressionar a "seta para cima" na linha de comando para executar um comando executado anteriormente.

Você também deve verificar as preferências do TextWrangler e definir para que o caractere `tab` seja substituido por quatro espaço. Isto evitará perder tempo procurando por erros de indentação.

Você também pode encontrar maiores informações sobre como editar e executar programas Python no endereço `www.py4inf.com`.

Apêndice D

Contribuições

Lista de Contribuidores para o "Python para Informáticos"

Bruce Shields por copiar as edições dos primeiros rascunhos Sarah Hegge, Steven Cherry, Sarah Kathleen Barbarow, Andrea Parker, Radaphat Chongthammakun, Megan Hixon, Kirby Urner, Sarah Kathleen Barbrow, Katie Kujala, Noah Botimer, Emily Alinder, Mark Thompson-Kular, James Perry, Eric Hofer, Eytan Adar, Peter Robinson, Deborah J. Nelson, Jonathan C. Anthony, Eden Rassette, Jeannette Schroeder, Justin Feezell, Chuanqi Li, Gerald Gordinier, Gavin Thomas Strassel, Ryan Clement, Alissa Talley, Caitlin Holman, Yong-Mi Kim, Karen Stover, Cherie Edmonds, Maria Seiferle, Romer Kristi D. Aranas (RK), Grant Boyer, Hedemarrie Dussan,

Prefácio de "Think Python"

A estranha história de "Think Python"

(Allen B. Downey)

Em Janeiro de 1999 estava me preparando para dar aulas para uma turma de Introdução à Programação em Java. Tinha ensinado por três vezes e estava ficando frustrado. O nível de reprovação na matéria estava muito alto e, mesmo para estudantes que tinham sido aprovados, o nível de aproveitamento foi muito baixo.

Um dos problemas que eu percebi, eram os livros. Eles eram muito grandes, com muitos detalhes desnecessários sobre Java, e orientação insuficiente sobre como programar. E todos sofriam do efeito alçapão: eles iniciavam fácil, continuavam gradualmente, e então em algum lugar em torno do Capítulo 5 o chão se desfazia. Os estudantes teriam novos assuntos, muito rápido, e eu perderia o resto do semestre juntando as peças.

Duas semanas antes do primeiro dia de aula, decidi escrever meu próprio livro.

Meus objetivos eram:

- Mantê-lo curto. É melhor para os estudantes lerem 10 páginas do que estudar 50 páginas.
- Ser cuidadoso com o vocabulário. Tentei minimizar os jargões e definir os termos na primeira vez que for utilizar.
- Evolução gradual. Para evitar o efeito alçapão, peguei os tópicos mais difíceis e dividi em séries de pequenos passos.
- Foco em programação, não na linguagem. Eu inclui um subconjunto mínimo de Java e deixei o resto de fora.

Eu precisava de um título, e por um capricho eu escolhi *Como Pensar como um Cientista da Computação*.

Minha primeira versão foi dura, mas funcionou. Os estudantes leram e entenderam o suficiente que eu pudesse dedicar as aulas nos tópicos difíceis, os tópicos interessantes e (mais importantes) deixando os estudantes praticarem.

Eu liberei o livro sob a Licença GNU Free Documentation, que permite aos usuários copiar, modificar e redistribuir o livro.

O que aconteceu depois disso foi a parte mais legal. Jeff Elkner, um professor de escola de ensino médio na Virgínia, adotou meu livro e adaptou para Python. Ele me enviou uma cópia da sua adaptação, e então tive a experiência de aprender Python lendo meu próprio livro.

Eu e Jeff revisamos o livro, incorporando um caso de estudo do Chris Meyers, e em 2001 nós liberamos *Como Pensar como um Cientista da Computação: Aprendendo com Python*, também sob a licença GNU Free Documentation. Pela Green Tea Press, publiquei o livro e comecei a vender cópias físicas pela Amazon.com e na livraria da Faculdade. Outros livros da Green Tea Press estão disponíveis no endereço `greenteapress.com`.

Em 2003 eu comecei a lecionar na faculdade de Olin e comecei a ensinar Python pela primeira vez. O contraste com Java foi impressionante. Os estudantes lutavam menos e aprendiam mais, trabalhavam com mais interesse nos projetos, e normalmente se divertiam mais.

Pelos últimos cinco anos eu continuei a desenvolver o livro, corrigindo erros, melhorando alguns dos exemplos e adicionando materiais, especialmente exercícios. Em 2008, comecei a trabalhar em uma nova revisão, ao mesmo tempo eu entrei em contato com um editor da Editora da Universidade de Cambridge que se interessou em publicar a próxima edição. Ótima oportunidade!

Eu espero que você aprecie trabalhar neste livro, e que ele ajude você a aprender a programar e pense, pelo menos um pouco, como um cientista da computação.

Reconhecimentos para “Think Python”

(Allen B. Downey)

Primeiramente e mais importante, eu gostaria de agradecer Jeff Elkner, que adaptou meu livro em Java para Python, que pegou este projeto e me introduziu no que se tornou a minha linguagem favorita.

Eu também quero agradecer Chris Meyers, que contribuiu para muitas seções para *Como Pensar como um Cientista da Computação*.

E eu gostaria de agradecer a Free Software Foundation por desenvolver a Licença GNU Free Documentation, que ajudou na minha colaboração entre Jeff e Chris possível.

Gostaria de agradecer aos editores da Lulu que trabalharam no *How to Think Like a Computer Scientist*.

Agradeço a todos os estudantes que trabalharam nas primeiras versões deste livro e todos os contribuidores (listados no apêndice) que enviaram correções e sugestões.

E agradeço a minha esposa, Lisa, pelo seu trabalho neste livro, e a Green Tea Press, por todo o resto.

Allen B. Downey
Needham MA

Allen Downey é professor associado do curso de Ciência da Computação na Faculdade de Engenharia Franklin W. Olin.

Lista de contribuidores para o “Think Python”

(Allen B. Downey)

Mais de 100 leitores atentos e dedicados tem enviado sugestões e correções nos últimos anos. Suas contribuições e entusiasmo por este projeto, foram de grande ajuda.

Para detalhes sobre a natureza das contribuições de cada uma destas pessoas, veja o texto the “Think Python”.

Lloyd Hugh Allen, Yvon Boulianne, Fred Bremmer, Jonah Cohen, Michael Conlon, Benoit Girard, Courtney Gleason e Katherine Smith, Lee Harr, James Kaylin, David Kershaw, Eddie Lam, Man-Yong Lee, David Mayo, Chris McAloon, Matthew J. Moelter, Simon Dicon Montford, John Ouzts, Kevin Parks, David Pool, Michael Schmitt, Robin Shaw, Paul Sleigh, Craig T. Snydal, Ian Thomas, Keith Verheyden, Peter Winstanley, Chris Wrobel, Moshe Zadka, Christoph Zwerschke, James Mayer, Hayden McAfee, Angel Arnal, Tauhidul Hoque

e Lex Berezhny, Dr. Michele Alzetta, Andy Mitchell, Kalin Harvey, Christopher P. Smith, David Hutchins, Gregor Lingl, Julie Peters, Florin Oprina, D. J. Webre, Ken, Ivo Wever, Curtis Yanko, Ben Logan, Jason Armstrong, Louis Cordier, Brian Cain, Rob Black, Jean-Philippe Rey da Ecole Centrale Paris, Jason Mader da George Washington University fez uma série Jan Gundtofte-Bruun, Abel David e Alexis Dinno, Charles Thayer, Roger Sperberg, Sam Bull, Andrew Cheung, C. Corey Capel, Alessandra, Wim Champagne, Douglas Wright, Jared Spindor, Lin Peiheng, Ray Hagtvedt, Torsten Hübsch, Inga Petuhhov, Arne Babenhauserheide, Mark E. Casida, Scott Tyler, Gordon Shephard, Andrew Turner, Adam Hobart, Daryl Hammond e Sarah Zimmerman, George Sass, Brian Bingham, Leah Engelbert-Fenton, Joe Funke, Chao-chao Chen, Jeff Paine, Lubos Pintes, Gregg Lind e Abigail Heithoff, Max Hailperin, Chotipat Pornavalai, Stanislaw Antol, Eric Pashman, Miguel Azevedo, Jianhua Liu, Nick King, Martin Zuther, Adam Zimmerman, Ratnakar Tiwari, Anurag Goel, Kelli Kratzer, Mark Griffiths, Roydan Ongie, Patryk Wolowiec, Mark Chonofsky, Russell Coleman, Wei Huang, Karen Barber, Nam Nguyen, Stéphane Morin, Fernando Tardio, e Paul Stoop.

Apêndice E

Detalhes sobre Direitos Autorais

Eu preferiria ter licenciado este livro sobre uma licença menos restritiva que a licença CC-BY-SA. Mas infelizmente existem algumas organizações sem escrúpulos que procuram por livros livres de licenças, e então, publicam e vendem virtualmente cópias idênticas destes livros em serviços que imprimem sob demanda, como a LuLu ou a CreateSpace. A CreatSpace tem (agradecidamente) adicionado uma política que dá aos atuais detendetores dos direitos autorais preferências sobre um não-dententor dos direitos autorais que tentar publicar um trabalho licenciado livremente. Infelizmente existem muitos serviços de impressão-por-demanda e muitos deles tem uma política que considere trabalhos assim como a CreateSpace.

Lamentavelmente eu adicionei o elemento NC a licença deste livro para me dar recursos em casos em que alguém tente clonar este livro e vendê-lo comercialmente. Infelizmente, adicionar o elemento NC, limita o uso deste material que eu gostaria de permitir. Então eu decidi adicionar esta seção ao documento para descrever situações específicas onde eu dou a permissão em casos específicos para uso do material deste livro, em situações que alguém pode considerar comercial.

- Se você está imprimindo um número limitado de cópias de todo o livro ou parte dele para uso em um curso (e.g., como um pacote de um curso), então você está permitido pela licença CC-BY deste material para este propósito.
- Se você é um professor em um universidade e você traduziu este livro para outro idioma, que não seja Inglês e ensina utilizando a versão traduzida deste livro, então você pode me contactar e eu vou conceder uma licença CC-BY-SA para este material respeitando a publicação da sua tradução. Em particular, você terá permissão de vender o resultado da sua tradução comercialmente.

Se você prentende traduzir este livro, você pode entrar em contato comigo e nós teremos certeza que você tem todo o material relacionado ao curso e então você pode traduzí-los também.

Obviamente, você é bem vindo para me contactar e pedir permissão se estas cláusulas forem insuficientes. Em todo o caso, permissão para reúso e mesclas a este material serão concedidas desde que fique claro os benefícios para os alunos e professores dos valores adicionados que acumularão como resultado do novo trabalho.

Charles Severance
www.dr-chuck.com
Ann Arbor, MI, USA
September 9, 2013

Índice Remissivo