Чтение из файлов

Глава 7

Python for Informatics: Exploring Information
www.pythonlearn.com

Пора поработать с данными!

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 -0500To: source@collab.sakaiproject.orgFrom: stephen.marquard@uct.ac.zaSubject: [sakai] svn commit: r39772 - content/branches/Details: http://source.sakaiproject.org/viewsvn/?view=rev&rev=39772
...

Разбор файлов

- Текстовый файл можно представить как последовательность строк

```
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
```

http://www.py4inf.com/code/mbox-short.txt

Открытие файла

- Прежде чем мы сможем прочитать содержимое файла, мы должны задать файл, с которым мы будем, и что мы хотим с ним сделать
- Это делается с помощью функции open()
- Функция open() возвращает дескриптор файла - переменную, используемую для выполнения операций над файлом
- Эта функция схожа с опцией "Файл -> Открыть" в текстовом процессоре

Использование функции open()

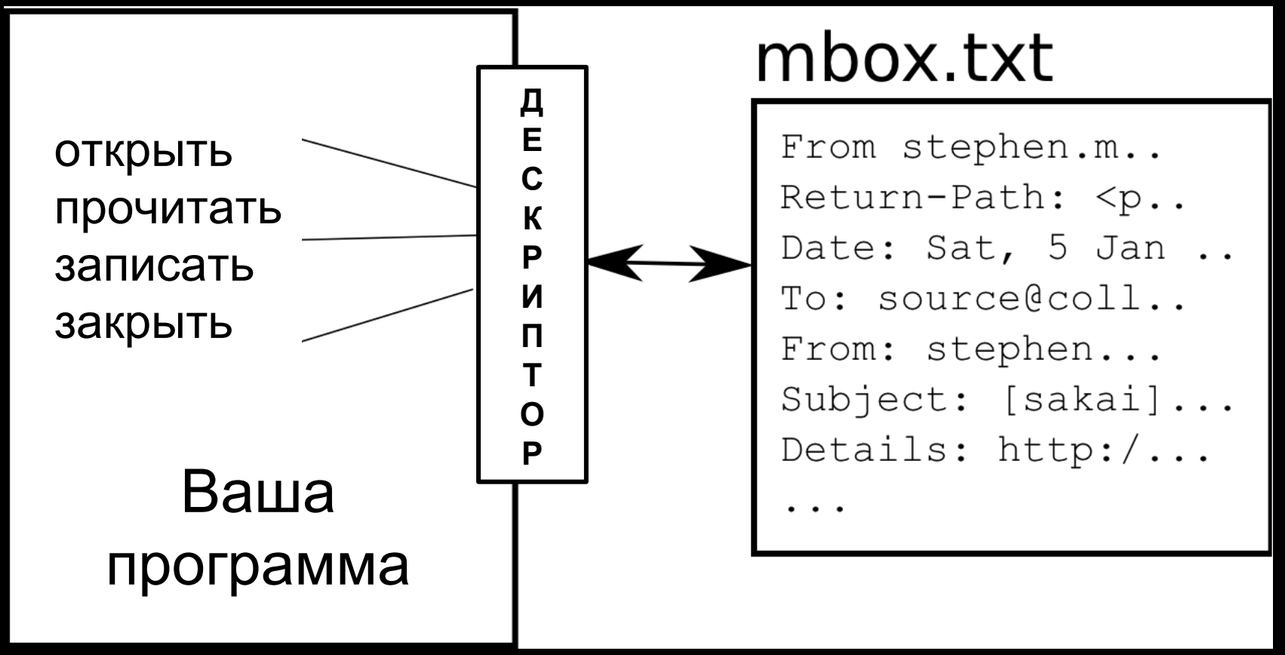
- handle = open(filename, mode)

```
fhand = open('mbox.txt', 'r')
```

  - Возвращает дескриптор файла, используемый для выполнения операций над файлом
  - Название файла является строкой
  - Указание режима необязательно. 'r' используется для чтения из файла, а 'w' - для записи в файл

Дескриптор файла

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



Если файл отсутствует

```
>>> 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'
```

Символ новой строки

- Для указания окончания строки используется символ новой строки
- В строках он отмечается как \n
- Символ новой строки считается как один знак, а не два

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

Обработка файлов

- Текстовый файл можно представить как последовательность строк

```
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
```

Обработка файлов

- В конце каждой строки текстового файла содержится символ новой строки

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

Дескриптор файла как последовательность

- Открытый для чтения дескриптор файла можно рассматривать как последовательность строк, где каждая строка в файле является строкой в последовательности
- Инструкцию for можно использовать для итерации по последовательности
- Не забудьте, что последовательность - это упорядоченный набор данных

```
xfile = open('mbox.txt')
for cheese in xfile:
        print cheese
```

Подсчет строк в файле

- Открыть файл только для чтения
- Цикл с for для чтения каждой строки
- Посчитать строки и распечатать число строк

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

```
$ python open.py
Количество строк: 132045
```

Чтение из *всего* файла

- Можно прочитать весь файл (включая разрывы строк) как одну строку

```
>>> fhand = open('mbox-short.txt')
>>> inp = fhand.read()
>>> print len(inp)94626
>>> print inp[:20]
From stephen.marquar
```

Поиск по файлу

- Если в наш цикл for добавить инструкцию if, можно распечатать строки, отвечающие некоторому заданному критерию

```
fhand = open('mbox-short.txt')
for line in fhand:
      if line.startswith('From:') :
            print line
```

ОЙ!

Откуда взялись все эти пустые строки?

From: stephen.marquard@uct.ac.za

From: louis@media.berkeley.edu

From: zqian@umich.edu

From: rjlowe@iupui.edu

...

ОЙ!

Откуда взялись все эти пустые строки?

- В конце каждой строки файла имеется символ разрыва строки
- Оператор печати добавляет к каждой строке разрыв строки

From: stephen.marquard@uct.ac.za\n
\n
From: louis@media.berkeley.edu\n
\n
From: zqian@umich.edu\n
\n
From: rjlowe@iupui.edu\n
\n
...

Поиск по файлу (исправленный)

- С помощью оператора rstrip() библиотеки строк мы можем избавиться от лишних пробелов справа
- Разрыв строки считается “пробелом” и удаляется

```
fhand = open('mbox-short.txt')
for line in fhand:
        line = line.rstrip()
        if line.startswith('From:') :
            print line
```

From: stephen.marquard@uct.ac.za
From: louis@media.berkeley.edu
From: zqian@umich.edu
From: rjlowe@iupui.edu
....

Пропуск строк с помощью continue

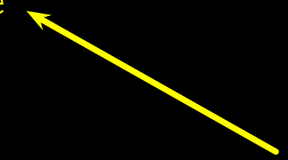
- С помощью оператора continue можно удобно пропустить строку

```
fhand = open('mbox-short.txt')
for line in fhand:
    line = line.rstrip()
    if not line.startswith('From:') :
        continue
    print line
```

Выбор строк с помощью in

- С помощью ключевого слова in мы можем выполнять поиск по строкам

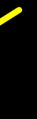
```
fhand = open('mbox-short.txt')
for line in fhand:
    line = line.rstrip()
    if not '@uct.ac.za' in line :
        continue
    print line
```



```
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...
```

Запрос названия файла

```
fname = raw_input('Введите название файла:  ')
fhand = open(fname)
count = 0
for line in fhand:
    if line.startswith('Subject:') :
        count = count + 1
print 'Найдено', count, 'строк с темой в', fname
```

Enter the file name: mbox.txt
There were 1797 subject lines in mbox.txt

Enter the file name: mbox-short.txt
There were 27 subject lines in mbox-short.
txt

Неверные названия файлов

```
fname = raw_input('Введите название файла:  ')
try:
    fhand = open(fname)
except:
    print 'Не удается открыть файл:', fname
    exit()
count = 0
for line in fhand:
    if line.startswith('Subject:') :
        count = count + 1
print 'Найдено', count, 'строк с темой в', fname
```

Введите название файла: mbox.txt
Найдено 1797 строк с темой в mbox.txt

Введите название файла: na na boo boo
Не удается открыть файл: na na boo boo

Обзор

- Вторичная память
- Открытие файла - дескриптор файла
- Структура файла - символ разрыва строки
- Чтение файла по строкам с помощью цикла for
- Поиск строк
- Чтение названий файлов
- Работа с неверными названиями файлов

Благодарность / Содействие

Первоначальная разработка: Чарльз Северанс, Школа информации Мичиганского университета

Здесь впишите дополнительных авторов и переводчиков...

...