In/Out: Modules, Files, Objects

Introduction to Python

Data Sciences Institute, University of Toronto
Instructor: A Mahfouz | TA: Kaylie Lau
July 2022

Contents:

Modules

What is a module?

A *module* is a file that contains Python function definitions and executable statements. Just as functions let us reuse code in a prigram, modules let us reuse code across multiple programs. Modules to do related tasks can be collected into a *package*. Later on, we will use packages to work with datasets. You may hear terms module, package, and library used interchangeably -- they all refer to reusable collections of code that make it easier to do certain tasks.

Working with modules

Python comes with several modules built in. To use one, we `import` it. Normally, `import` statements go at the very top of the file -- we're only putting them in the middle here for teaching purposes.

```
In [1]: import math
```

When using code in a module, we first reference the module, follwed by a period, then the value or function being used.

In [2]:
```
math.pi
```

Out[2]:

```
3.141592653589793
```

In [3]:
```
math.gcd(13984, 232, 18356)
```

Out[3]:

```
4
```

Not referencing the module produces an error.

```
In [4]: pi
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [4], in <cell line: 1>()
----> 1 pi

NameError: name 'pi' is not defined
```

```
In [5]: gcd(13984, 232, 18356)
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Input In [5], in <cell line: 1>()
----> 1 gcd(13984, 232, 18356)

NameError: name 'gcd' is not defined
```

If we only want a part of a module, we can import parts with a `from...import` statement. This is useful for large modules like `datetime`, where we may only want to use a few features.

In [6]:
```
from datetime import date
```

In [7]:
```
# we don't have to write datetime.date.today()
print(date.today())
```

```
2022-07-07
```

Working with Files

So far we have worked with data bundled with our notebooks. Most data, of course, is stored elsewhere: in databases and files stored locally and online. We need to be able to read data into a program to do analyses. We should also be able to write data out to files of our own.

Reading and writing files

Luckily, Python has a built-in function, `open()` for opening files. `open()` takes a string indicating the *file path*, or location, of the file to open, plus a one-character string indicating whether we should open the file in read-only, overwrite, or append mode.

`open()` mode	Description
`'r'`	Read-only. Produces an error if the file does not already exist.
`'w'`	Write. Creates a new file if one does not exist. If the file already exists, the current contents are deleted and overwritten.
`'a'`	Append. Adds to an existing file. If the file does not exist, it will be created.

Paths can be *absolute*, like `C:/Users/Owner/Documents/data/demo_data.csv`, or *relative*, like `data/demo_data.csv`. Relative paths are relative to the folder the Python file is in -- if our code is not in `Documents`, `data/demo_data.csv` won't work.

Opening files

To open a file, we use a `with` statement, which follows the pattern `with open('file_path') as file_variable_name:`, then an indented block of code to process the file. The `with` statement ensures that Python closes the file when we're done working with it.

In [8]:
```
with open('sample_data/california_housing_test.csv', 'r') as f:
    print(f)
```

```
<_io.TextIOWrapper name='sample_data/california_housing_test.csv' mode='r' encoding='cp1252'>
```

Reading files

Opening a file doesn't immediately get us the file's contents. To do that, we must use a read method.

- `read()` returns the full file contents, which can be overwhelming for larger files.
- `readline()` returns only the next line in the file. Python keeps track of where it is in the file.
- `readlines()` returns the full file as a list. Each item is one line in the file.

In [9]:
```
with open('sample_data/california_housing_test.csv', 'r') as f:
    for i in range(5):
        print(f.readline())
```

```
longitude,latitude,housing_median_age,total_rooms,total_bedrooms,population,households,median_inc
ome,median_house_value

-122.05,37.37,27,3885,661,1537,606,6.6085,344700

-118.3,34.26,43,1510,310,809,277,3.599,176500

-117.81,33.78,27,3589,507,1484,495,5.7934,270500

-118.36,33.82,28,67,15,49,11,6.1359,330000
```

Writing files

There are corresponding `write()` methods for files.

- `write()` writes a string to file.
- `writelines()` writes each item in an iterable to file, with no separating text in between.

In [10]:
```
provinces = ['BC', 'AB', 'SK', 'MB', 'ON', 'QC', 'NL', 'NB', 'NS', 'PE']
with open('provinces.txt', 'w') as province_file:
    province_file.writelines(provinces)
```

In [11]:
```
with open('provinces.txt', 'r') as province_file:
    print(province_file.read())
```

```
BCABSKMBONQCNLNBNSPE
```

Working with specific file formats

Python has built-in modules for working with specific file formats, like `csv` and `json`. We won't spend much time here, as we will soon be working with libraries that let us open, analyze, and write data in both formats.

In [12]:
```
import csv
```

In [ ]:
```
with open('sample_data/california_housing_test.csv', 'r') as f:
        contents = csv.reader(f)
    for row in contents:
        print(row)
```

Navigating folders

Being able to navigate the computer's file system enables us to work with entire folders' worth of files stored locally on a computer (or "locally" in an environment like Colab). Python's built-in `os` module lets us do just that.

In [14]:
```
import os
```

In [15]:
```
# get the path to the folder we're currently in
os.getcwd()
```

Out[15]:

```
'C:\\Users\\unive\\Documents\\GitHub\\dsi-python-workshop\\01-slides'
```

In [16]:
```
# see the contents of the current folder
os.listdir()
```

Out[16]:

```
['.ipynb_checkpoints',
 '01_intro_python.ipynb',
 '02_control_flow_iterables.ipynb',
 '03_modules_files_oop.ipynb',
 'plants.txt',
 'provinces.txt',
 'sample_data']
```

In [17]:
```
# loop over the files in the sample data folder
for i in os.listdir('sample_data'):
    print(i)
```

```
anscombe.json
california_housing_test.csv
```

Manipulating paths

The `os.path` submodule provides safe ways to manipulate paths. `os.path.join()` lets us create properly formatted paths from separate folder and file names, without worrying about getting slashes right. `os.path.exists()` lets us check for a file before trying to open or accidentally overwriting it.

In [18]:
```
# create full paths for sample data files

cwd = os.getcwd()
full_paths = []
for i in os.listdir('sample_data'):
    full_paths.append(os.path.join(cwd,
                                   'sample_data',
                                   i))
full_paths
```

Out[18]:

```
['C:\\Users\\unive\\Documents\\GitHub\\dsi-python-workshop\\01-slides\\sample_data\\anscombe.json
',
 'C:\\Users\\unive\\Documents\\GitHub\\dsi-python-workshop\\01-slides\\sample_data\\california_ho
using_test.csv']
```

In [19]:
```
# check if a file exists to avoid overwriting it
        text = 'Lavender is a small purple flower.'

if os.path.exists('plants.txt'):
    print('plants.txt already exists')
else:
    with open('plants.txt', 'w') as f:
        f.write(text)
```

```
plants.txt already exists
```

Object-Oriented Programming

(a very brief introduction)

What is object-oriented programming (OOP)?

Object-oriented programming is an approach to writing programs that organizes code into *objects* with data, or *attributes*, and *methods* (which we've seen before!). The Python types we have seen are all objects. The data science packages we'll encounter adopt this approach as well.

In Python, objects are modeled with *classes*. We can think of a class as a template that defines an object's attributes and methods.

For example, we can model basketball players as a class. To keep it simple, let's say each player has a position, team, and jersey number. Players can dribble, shoot, or pass the ball to another player.
In this case, the class *BasketballPlayer* would have the attributes *position*, *team*, and _jersey*number*, and the methods *dribble()*, *shoot()*, and _pass*ball(player)*. To create a new BasketballPlayer, we would have to somehow specify their position, team, and jersey number.

We can model abstract concepts as classes, too. A *Dataset* class might have attributes like values and size. Methods could include calculating summary statistics, like finding the mean or counting unique values. To create a new Dataset, we would have to specify the values, while the size could be calculated for us.
A *LinearModel* class might have attributes like coefficients and variables, and methods to fit the model to data or make predictions.

Creating a class

Let's turn *Dataset* into an actual class. Similar to how we use `def` to start a function definition, we use `class` to define a new class. Class names, by convention, start with a capital letter.

```
class Dataset:
    '''Data values and methods to summarize them''''
```

Then, we define the class's `__init__()` method, which is a special method that is called whenever we make a new instance of our class. It *initializes* the object in memory. The first parameter of every method is `self`, a reference to the object. `__init()__` should also take data to use to make an object -- in this case, values to store.

```
class Dataset:
    '''Data values and methods to summarize them'''

    def __init__(self, values):
        '''Create a new dataset with values as a list'''
```

We use `__init__()` to attach data, or attributes, to the individual object. This looks like variable assignments we've seen so far, except that every variable name starts with `self.`. Notice that attributes can be derived.

```
class Dataset:
    '''Data values and methods to summarize them'''

    def __init__(self, values):
        '''Create a new dataset with values as a list'''
        self.values = values
        self.size = len(values)
```

We can then define additional methods. Each method takes `self` as its first parameter. Attributes should always be referenced with the `self.` prefix.

```
def unique(self):
        '''Return a list of unique values'''
        return list(set(self.values))
```

A bare-bones Dataset class might look like this.

In [20]:
```
class Dataset:
    '''Data values and methods to summarize them'''

    def __init__(self, values):
        '''Create a new dataset with values as a list'''
        self.values = values
        self.size = len(values)

    def unique(self):
        '''Return a list of unique values'''
        return list(set(self.values))

    def mean(self):
        '''Return the mean of all numeric values'''
        numbers = []

        for i in self.values:
            if type(i) is int or type(i) is float:
                numbers.append(i)
        return sum(numbers)/len(numbers)
```

To create a Dataset instance, we do not call `__init__()` directly. Instead, we call `Dataset()`, passing in `values`. Once we have a Dataset instance, we can call Dataset methods just as we did built-in type methods.

In [21]:
```
grades = Dataset([90, 87, 70, 70, 87, 'A'])

print(grades.unique())
print(grades.size)
print(grades.mean())
```

```
['A', 90, 70, 87]
6
80.8
```

Why the double underscores?

`__init__()` is what Python calls a magic or dunder (for double underscore) method. These methods customize what Python does when other functions are called, like how `__init__()` defined `Dataset()`.
Other common magic methods we encounter in classes are `__repr__()`, which tells Python what should appear when we `print()` an object, and `__str__()`, which Python uses to convert an object to a string.

Why learn this?

It's possible to do a lot of analysis without ever designing our own classes. We do, however, encounter objects all the time. The hypothetical *Dataset* and *LinearModel* classes are simplified versions of real classes in packages like `numpy` and `scikit-learn`. Having an intuition for objects makes it easier to understand why code looks the way it does and helps demystify documentation, which often assumes the reader is familiar with OOP concepts.

References

- Bostroem, Bekolay, and Staneva (eds): "Software Carpentry: Programming with Python" Version 2016.06, June 2016, **https://github.com/swcarpentry/python-novice-inflammation**, 10.5281/zenodo.57492.
- Chapter 6, 10 and 14, Gries, Campbell, and Montojo, 2017, *Practical Programming: An Introduction to Computer Science Using Python 3.6*
- "Modules", Python Software Foundation, *Python Language Reference, version 3*. Available at **https://docs.python.org/3/tutorial/modules.html**.