Reading, writing files

Terence Parr
MSDS program
University of San Francisco

See notebook https://github.com/parrt/msds501/blob/master/notes/files.ipynb

UNIVERSITY OF SAN FRANCISCO

What are files?

- Both the disk and RAM are forms of memory
- RAM is much faster (but smaller) than the disk but RAM data disappears when the power goes out
- Disks, in contrast, are persistent.
- A *file* is simply a chunk of data on the disk identified by a filename
- File data is less convenient to access because we have to explicitly load the file into working memory first
- If a file is too big to fit into memory all at once, we have to process the data in chunks

File state

- Files must be opened and then closed when we're done
- Files are opened for reading or writing
- Files are opened with a mode: text or binary

```
f = open('foo.txt', mode='r')   # open for read text mode
read from f
f.close()                        # ok, we're done
```

- **f** is a file descriptor

Avoiding confusion

- The filename is a string that identifies a file on the disk. It can be fully qualified or relative to the current working directory

- The file object is not the filename and is also not the file itself on the disk. It's really just a descriptor and refers to the file content

- The contents of the file is different than the filename and the file (descriptor) object that Python gives us

The WITH statement

- The **with** statement helps us to automatically closing files

```
!head -5 data/prices.txt
```
```
0.605
0.600
0.594
0.592
0.600
```

```
with open("data/prices.txt") as f:
    contents = f.read()
lines = contents.split()
print(lines[0:3])
print(f.closed)
```
```
['0.605', '0.600', '0.594']
True
```

Most common programming pattern

- Load all file contents into a string

```python
with open('data/IntroIstanbul.txt') as f:
    contents = f.read() # read all content of the file
print(contents[0:200]) # print just the first 200 characters
```

```
The City and ITS People
Istanbul is one of the worlds most venerable cities. Part
of the citys allure is its setting, where Europe faces Asia across
```

2nd most common programming pattern

- Load all lines of a file or words of a file into a list

```python
with open("data/prices.txt") as f:
    contents = f.read()
lines = contents.split()
print(lines[0:3])
print(f.closed)
```

```
['0.605', '0.600', '0.594']
True
```

```python
with open('data/prices.txt') as f:
    prices = f.readlines()
prices[0:3]
```

```
['0.605\n', '0.600\n', '0.594\n']
```

```python
with open('data/IntroIstanbul.txt') as f:
    contents = f.read() # read all content of the file
words = contents.split()
print(words[0:100]) # print first 100 words
```

```
['The', 'City', 'and', 'ITS', 'People', 'Istanbul', 'is',
```

UNIVERSITY OF SAN FRANCISCO

Using Pandas to load CSV files

- If the text file is also a comma separated value file (CSV), the easiest way to load the data is with Pandas

```python
import pandas as pd
data = pd.read_csv('data/player-heights.csv')
data.head(5)
```

	Football height	Basketball height
0	6.33	6.08
1	6.50	6.58
2	6.50	6.25
3	6.25	6.58
4	6.50	6.25