Data Foundations: More File IO and Objects

Instructor: Anthony Rios

Outline

More File IO

- Loading a CSV
- Creating a CSV File
- XML
- JSON

More File IO

- Loading a CSV
- Creating a CSV File
- XML
- JSON

CSV Files

A CSV file (Comma Separated Values file) is a type of **plain text file** that uses specific structuring to arrange **tabular data**.

```
anthony@MacBook:~$ cat exampleCSV.csv
column 1 name,column 2 name, column 3 name
first row data 1,first row data 2,first row data 3
second row data 1,second row data 2,second row data 3
...
```

Other popular delimiters include **tab** (\t), **colon** (:), and **semi-colon** (;) characters.

CSV files are common export formats from Excel and relational databases (e.g., MySQL and MS SQL Server).

CSV Basics

```
>>> print(csv)
File "<stdin>", line 1, in <module>
NameError: name 'csv' not defined.

>>> import csv
>>> print(csv)
<module 'csv' from '/.../python3.6/csv.py'>

>>> myFile = open('myfile.csv')
>>> myCSV = csv.reader(myFile, delimiter=',')
```

Basic Format:

csv.reader(**FILE HANDLE**, delimiter=**Delimiter Character**)

Reading a CSV

anthony@MacBook:~$ cat mycsv.csv

```
name,department,birthday month
Sarah,IT,January
John,Marketing,November
```

example.py

```
import csv
myFile = open('mycsv.csv')
csvRead = csv.reader(myFile, delimiter=',')
for row in csvReader:
        print(row[1]) # prints the second column
myFile.close()
```

anthony@MacBook:~$ cat myfile.txt

```
dept.
IT
Marketing
```

Reading a CSV

anthony@MacBook:~$ cat mycsv.csv

```
name,department,birthday month
Sarah,IT,January
John,Marketing,November
```

example.py

```
csvRead = csv.reader(myFile, delimiter=',')
isHeader = True
for row in csvReader:
        if isHeader: # Ignore header
                isHeader = False
        else:
                print(row[1]) # prints the second column
```

anthony@MacBook:~$ cat myfile.txt

```
IT
Marketing
```

Exercise 1

Write code that reads the csv file "housing_prices.csv" and calculate/print the following:

- Calculate and print the sum of all house prices
- Calculate and print the average price
- Calculate and print the max price
- Print the name of the street that contains the most expensive house.


10 minutes

Creating a CSV File

example.py

```
import csv
myFile = open('new_csv.csv', 'w', newline = '')
csvWriter = csv.writer(myFile, delimiter=',')
csvWriter.writerow(['col 1', 'col 2', 'col 3'])
csvWriter.writerow(['a','b','c'])
myFile.close()
```

anthony@MacBook:~$ cat new_csv.csv

```
col 1,col2,col 3
a,b,c
```

Exercise 2

Given the following list of lists

myData = [['name','department','birthday month'], ['John Doe','Marketing','November'],['Jane Smith', 'IT', 'March']]

create a csv file that is delimited with the tab '\t' character using the csv.writer() method. Name the file "employee_birthday.csv". **The list is already in the jupyter notebook under Exercise 2. Simply run the sell containing the list.**.


3 minutes

Data Example

Assume we want to send, retrieve, and display the following information:

Person

Name: Chuck
Phone (international): +1 734 303 4456
Email: Hidden

XML

- Introduced in 1996
- e**X**tensible **M**arkup **L**anguage, is a a specification for creating custom **markup languages**.
  - A **markup language** is a system for annotating a document in a way that is **syntactically distinguishable** from the text.
- XML is a **meta-language**. That means that you use it to for creating your own languages.
- The primary purpose is to help **share data** across different computers.

XML

Person

Name: Chuck
Phone (international): +1 734 303 4456
Email: Hidden

anthony@MacBook:~$ cat myfile.xml

```
<person>
        <name>Chuck</name>
        <phone type="intl">
                +1 734 303 4456
        </phone>
        <email hide="yes" />
</person>
```

XML has **NO** predefined tags.

XML

anthony@MacBook:~$ cat myfile.xml

```
<person>
      <name>Chuck</name>
      <phone type="intl">
            +1 734 303 4456
      </phone>
      <email hide="yes" />
</person>
```

- **Start Tag** – <person>, <name>, <phone type="intl">
- **End Tag** – </person>, </name>, and </phone>
- **Content** – Chuck and +1 734 303 4456
- **Attributes** (Will be part of the start tag) – **type="intl"** and **hide="yes"**
- **Element/node** – consists of a **start** tag, **content** (optional), and an **end** tag.

XML

anthony@MacBook:~$ cat myfile.xml

```
<person>
      <name>Chuck</name>
      <phone type="intl">
            +1 734 303 4456
      </phone>
      <email hide="yes" />
</person>
```

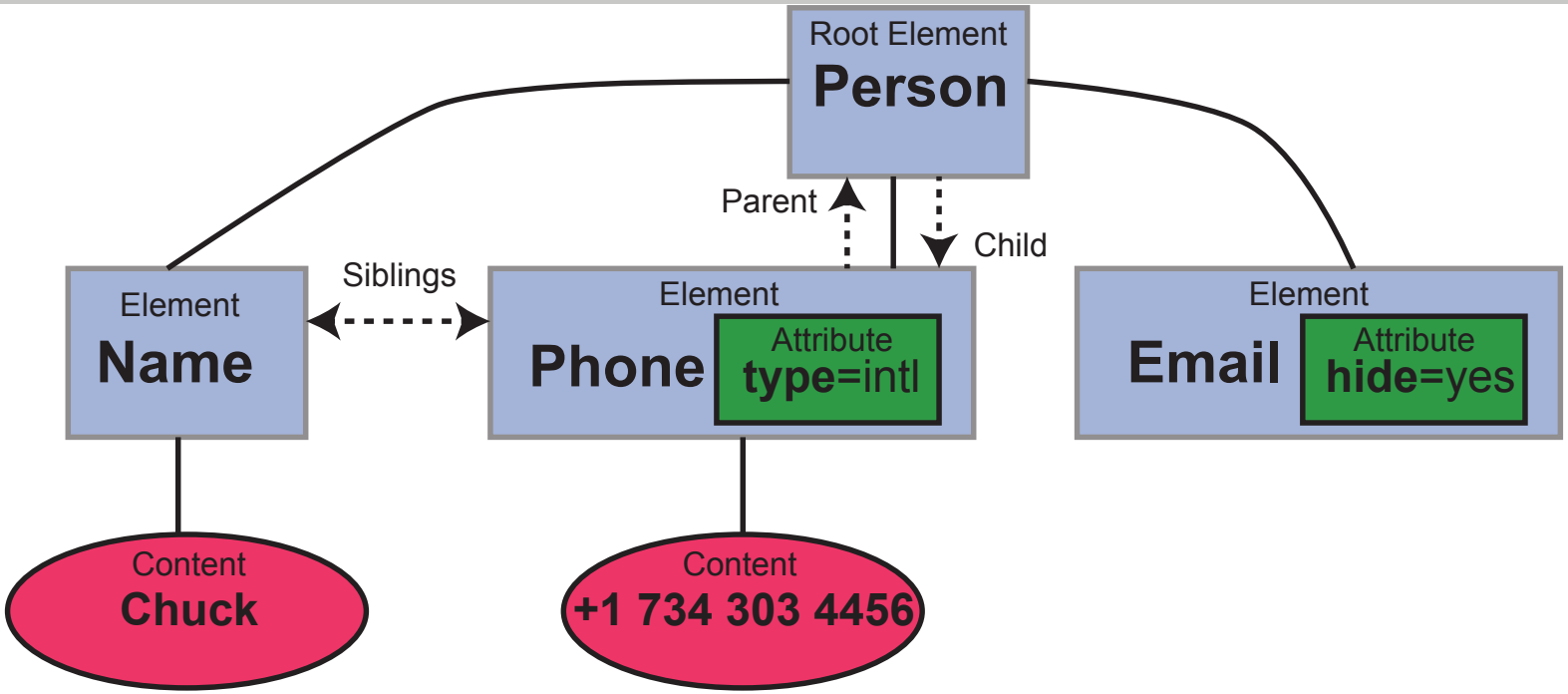
A tag is **empty** if it does not have any content <email></email>

Empty tags can also be written as <email/>

XML

anthony@MacBook:~$ cat myfile.xml

```
<person>
        <name>Chuck</name>
        <phone type="intl">
                +1 734 303 4456
        </phone>
        <email hide="yes" />
</person>
```



Parsing XML

example.py

```
import xml.etree.ElementTree as ET
data = '''<person> # ''' is used for long multi-line strings
                <name>Chuck</name>
                <phone type="intl">
                        +1 734 303 4456
                </phone>
                <email hide="yes" />
        </person>'''

tree = ET.fromstring(data) # converts a string of XML into a "tree" of nodes
# Find will return the first element that matches input parameter "name"
print('Name: {}'.format(tree.find('name').text))
# get returns a specific attribute of the element returned by find.
print('Attr: {}'.format(tree.find('email').get('hide')))
```

anthony@MacBook:~$ python example.py

```
Name: Chuck
Attr: yes
```

Looping through XML nodes/elements

anthony@MB:~$ cat myfile.xml

```
<stuff>
  <users>
    <user x="2">
      <id>001</id>
      <name>Chuck</name>
    </user>
    <user x="7">
      <id>009</id>
      <name>Brent</name>
    </user>
  </users>
</stuff>
```

anthony@MB:~$ python ex.py

```
User count: 2
Name Chuck
Id 001
Attribute 2
Name Brent
Id 009
Attribute 7
```

ex.py

```
import xml.etree.ElementTree as ET
to_open = open('myfile.xml') # Open XML file
input = to_open.read() # Read XML file into string
stuff = ET.fromstring(input)
# Returns all "user" subtrees in the XML file.
# findall takes and XPath expression as input
lst = stuff.findall('users/user')
# Count the number of elements/subtrees
# returned by findall
print('User count: {}'.format(len(lst))
for item in lst:
    print('Name {}'.format(item.find('name').text))
    print('Id {}'.format(item.find('id').text))
    print('Attribute {}'.format(item.get('x')))
```

https://www.w3schools.com/xml/xpath_syntax.asp

Exercise 3

A garden center has an XML (plant_catalog.xml) file that stores information, including price, for all plants they sell. The store is having a sale where everything is 20% off. Write a program that that prints the plant "COMMON" name, the current price, and the new sale price. An example of what the output should look like is shown below:

```
anthony@MB:~$ python ex.py
Bloodroot $2.44 to $1.95
Columbine $9.37 to $7.50
Marsh Marigold $6.81 to $5.45
...
```

Hint: You will need to use "string indexing".


10 minutes

JSON

`https://www.youtube.com/watch?v=7mj-p1Os6QA`

- First introduced in 1999
- JSON: **J**ava**S**cript **O**bject **N**otation
- JSON is a syntax for storing and exchanging data
- JSON is text, written with JavaScript Object Notation

JSON

- JSON is a **language-independent** data format

- JSON was derived from JavaScript, but as of 2017 **many programming languages** include code to generate and parse JSON-format data

JSON

Person

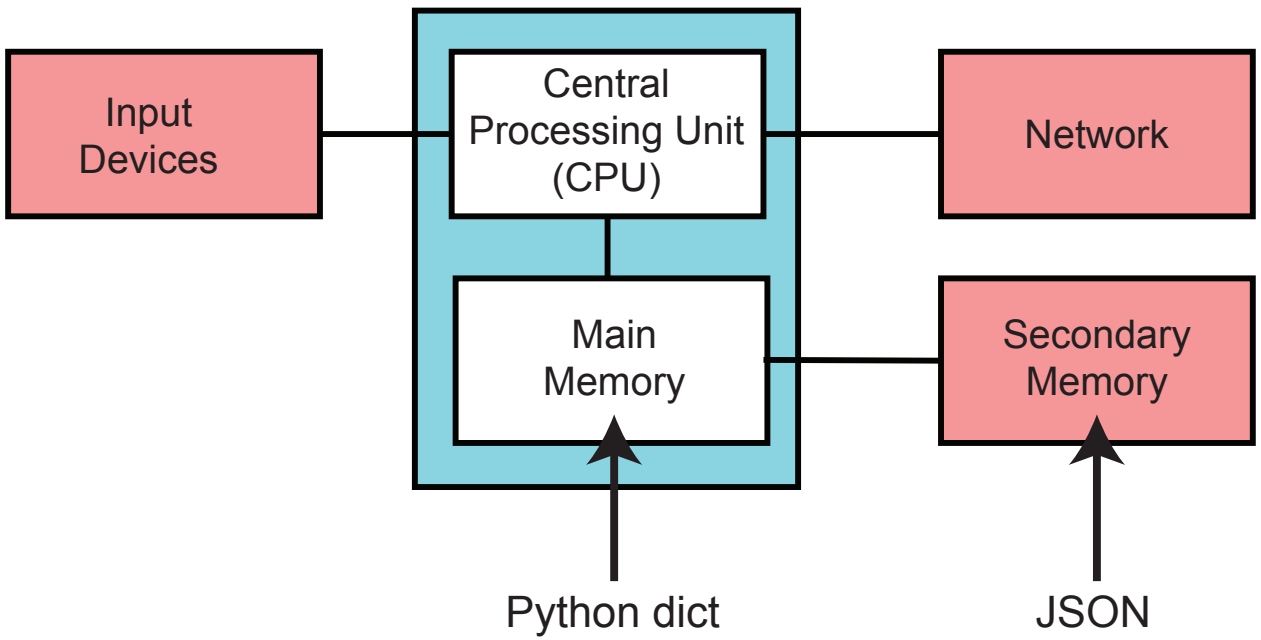
Name: Chuck
Phone (international): +1 734 303 4456
Email: Hidden

anthony@MacBook:~$ cat myfile.json

```
{
        "name" : "Chuck",
        "phone" : {
                "type" : "intl",
                "number" : "+1 734 303 4456"
        },
        "email" : {
                "hide" : "yes"
        }
}
```

JSON vs Python Dictionaries

- **JSON** is a **serialization format**. That is, JSON is a way of representing structured data in the form of a **string**.
- A **dictionary** is a **data structure**. That is, it is a way of storing data in memory that provides certain abilities to your code: in the case of dictionaries, those abilities include **rapid lookup** and **enumeration**.



JSON vs Python Dictionaries

JSON is built on **two structures**:

- A collection of **name(key)/value pairs**. In various languages, this is realized as an object, record, struct, dictionary, hash table, keyed list, or associative array.

- An **ordered list** of values. In most languages, this is realized as an array, vector, list, or sequence.

Python's **dicts** are an implementation of one of the structures JSON is inspired by, **key/value pairs**.

JSON vs XML

JSON is Like XML Because

- Both JSON and XML are "self describing" (human readable)
- Both JSON and XML are hierarchical (values within values)
- Both JSON and XML can be parsed and used by lots of programming languages

JSON is Unlike XML Because

- JSON doesn't use end tag
- JSON is shorter
- JSON is quicker to read and write
- JSON can use arrays

JSON vs XML

Why JSON is Better Than XML

- XML is much more difficult to parse than JSON.
- JSON resembles standard key/value data structures.

For WEB applications, JSON is faster and easier than XML:

Using XML

- Fetch an XML document
- loop through the document
- Extract values and store in variables

Using JSON

- Fetch a JSON string
- Parse JSON directly into variables

Parsing JSON

>>> myJSON = '{"name":"Anthony", "age":102,
"department":"ISCS"}'

>>> myJSON['name']
Traceboack (most recent call last):
File "<stdin>", line 1
TypeError: string indices must be integers

myJSON is a **string**, not a dictionary!

Parsing JSON

```
>>> import json # Load python JSON module

>>> myJSON = '{"name":"Anthony", "age":102,
"department":"ISCS"}'

>>> JSON_to_dict = json.loads(myJSON) # Load JSON from string

>>> JSON_to_dict['name']
'Anthony'
```

Loading JSON from a File

```
anthony@MacBook:~$ cat myfile.json
[{"name":"Anthony","age":102},
        {"name":"John","age":50}]
```

```
>>> import json
>>> myFile = open('myfile.json')
>>> data = json.load(myFile)  # Return JSON from file
>>> data # JSON object stored a list of dictionaries
{'name':'Anthony','age':102},{'name':'John','age':50}
>>> data[0] # First dict. in list
{'name':'Anthony',age':102}
>>> myFile.close()
```

Convert a Dictionary to a JSON String

```
>>> data = [{'name':'Anthony','age':102},{'name':'John','age':50}]

>>> import json

>>> JSONString = json.dumps(data) # Convert data to string

>>> JSONString
'[{"name":"Anthony","age":102},{"name":"John","age":50}]'
```

Saving JSON to a File

```
>>> data = [{'name':'Anthony','age':102},{'name':'John','age':50}]

>>> import json

>>> myFile = open('myjson.json','w')

>>> json.dump(data, myFile) # Save "data" to myjson.json

>>> myFile.close()
```

Exercise 4

Using the "exampleJSON.json" file, complete the following tasks:

- Load the file into a python dictionary.
- Change the email of item with the name "Anthony" to "anthony.rios@utsa.edu"
- Add a new person to the list with the name "
- Save the new dictionary to a JSON file "exampleJSON2.json"


10 minutes

Loads vs Load and Dumps vs Dump

What is the point of Loads and Dumps?

Why would we want to convert a python dict to a string when we can save directly to a file?

- "load" will load the **entire JSON object** into memory.
  - If we have a 10GB file (or bigger), loading the entire object is **NOT** feasible.
- Similarly, "dump" requires the entire python object to be loaded in memory.
  - For streaming data, we can **NOT** store all data in memory.
  - Example: Twitter data (Imagine collecting tweets with the hashtag #DataScience for 2 months)

JSON vs JSON Lines (JSONL)

anthony@MB:~$ cat myData.**json**

```
[ {"name":"Anthony", "age":102}
{"name":"John", "age":50}
{"name":"Jane", "age":75} ]
```

anthony@MB:~$ cat myData.**jsonl**

```
{"name":"Anthony", "age":102}
{"name":"John", "age":50}
{"name":"Jane", "age":75}
```

Reading a JSONL File

anthony@MB:~$ cat myData.**jsonl**

```
{"name":"Anthony", "age":102}
{"name":"John", "age":50}
{"name":"Jane", "age":75}
```

example.py

```
import json
myFile = open('myData.jsonl')
for line in myFile: # Loop over JSON file line-by-line
    lineData = json.loads(line.strip()) # Read 1 line at a time
    print("Name: {}".format(lineData["name"]))
myFile.close()
```

anthony@MB:~$ python example.py

```
Name: Anthony
Name: John
Name: Jane
```

Exercise 5

Write code to loop over the Twitter JSONL file "twitter.jsonl" and compute the following:

- Count and print the total number of tweets (1 JSON line = 1 tweet).
- Count and print the total number of users that are in the dataset (hint: data['user']['screen_name']).
- Print the screen name of the user who has the most tweets.


15 minutes

END

The End: Please feel free to post questions in the Slack.