Data Ingestion

Agenda

- Python Review
- Virtual Envs & Requirements
- Data Ingestion
  - Plain text
  - CSV
  - Excel
  - PDF
  - API
- Assignment: for the CodeAcademy submission
- Web Scraping (if we have time)

Python Review

Review

Integers
Floating Point

Dynamic Typing – no declarations

```
x = 5
y = 6.3
```

Names start with a letter, cAsE SeNsiTiVe.
Long names OK.

Review Character Strings

Dynamic typing – no declaration
No memory allocation
Immutable

```
s = "Good Afternoon"

len(s)                  # length of string
```

Review String Slicing

```
s = “Good Afternoon”

s[0]  evaluates to “G”

s[5:10] selects “After”     # string slicing

s[:10] selects “Good After”
s[5:]  selects “Afternoon”

s[-4:] selects “noon”    # last 4 characters
```

String Methods

String is a Class with data & subroutines:

```
t = s.upper()
pos = s.find("A")

first = "George"
last = "Washington"
name = first + " " + last
                    # string concatenation
```

Review Lists

Ordered sequence of items

Can be floats, ints, strings, Lists

a = [16, 25.3, “hello”, 45]

a[0] contains 16

a[-1] contains 45

a[0:2] is a list containing [16, 25.3]

Create a List

```
days = [ ]
days.append(“Monday”)
days.append(“Tuesday”)

years = range(2000, 2014)
years = xrange(2000, 2014)
```

List Methods

List is a Class with data & subroutines:

d.insert( )
d.remove( )
d.sort( )

Can concatenate lists with +

String split

```
s = “Princeton Plasma Physics Lab”

myList = s.split()          # returns a list of strings

print myList
      [ “Princeton”, “Plasma”, “Physics”, “Lab” ]

help(str.split)             # delimiters, etc.
```

Tuple

Designated by ( ) parenthesis

A List that can not be changed. Immutable.
No append.

Good for returning multiple values from a subroutine function.

Can extract slices.

Review math module

```
import math
dir(math)

math.sqrt(x)
math.sin(x)
math.cos(x)
```

```
from math import *
dir()

sqrt(x)
```

```
from math import pi
dir()

print pi
```

import a module

```
import math                # knows where to find it
```

```
import sys
sys.path.append("/u/efeibush/python")
import  cubic.py     # import your own code
```

```
if task == 3:
    import math      # imports can be anywhere
```

Review Defining a Function

Block of code separate from main.

Define the function before calling it.

```
def myAdd(a, b):        # define before calling
     return a + b

p = 25                  # main section of code
q = 30

r = myAdd(p, q)
```

Keyword Arguments

Provide default values for optional arguments.

```
def setLineAttributes(color="black",
    style="solid", thickness=1):
    ...

# Call function from main program
setLineAttributes(style="dotted")
setLineAttributes("red", thickness=2)
```

Looping with the range() function

```
for i in range(10):          # i gets 0 - 9
```

range() is limited to integers

numpy provides a range of floats

Summary

Integer, Float
String
List
Tuple

def function
Keywords: if elif else
while for in
import print

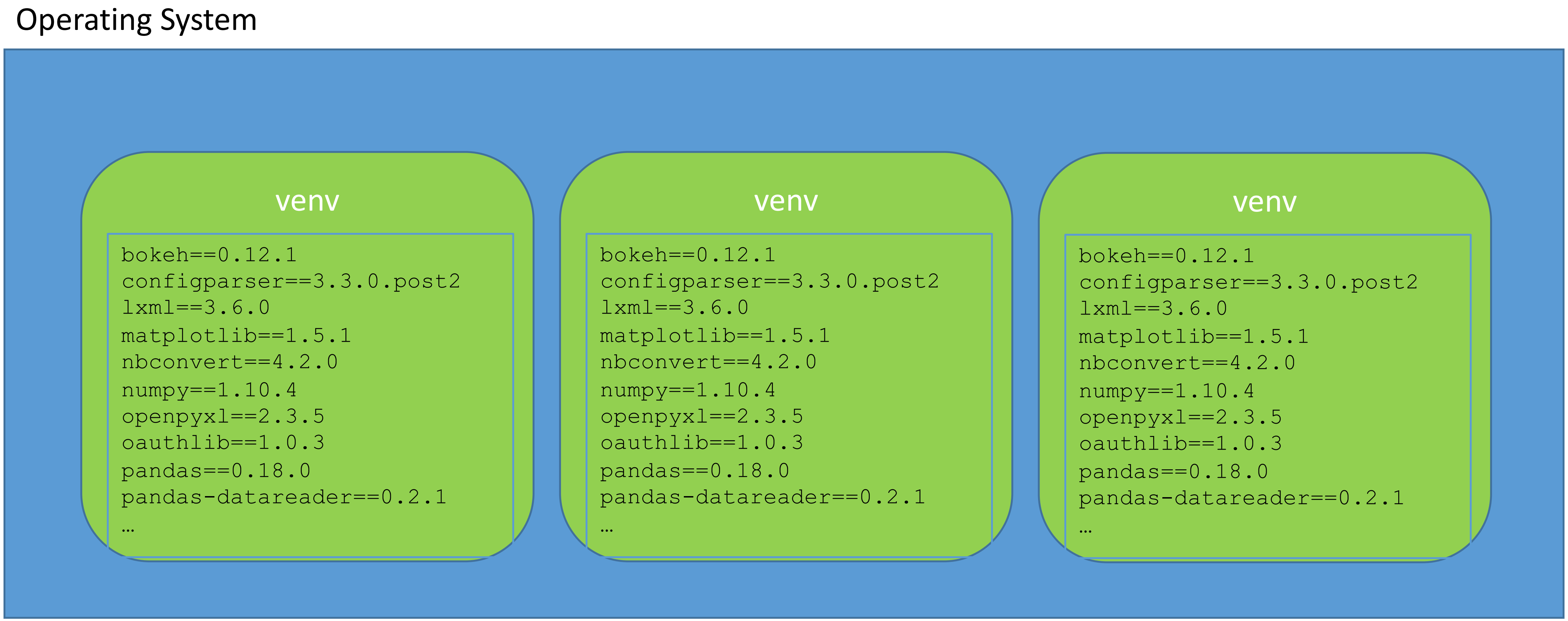
Indenting counts :

Run python as Interpreter

type()
dir()
help()

Virtual Envs
&
Packages

Virtual envs: isolation & portability



Packages we use: `requirements.txt`

```
bokeh==0.12.1
configparser==3.3.0.post2
lxml==3.6.0
matplotlib==1.5.1
nbconvert==4.2.0
numpy==1.10.4
openpyxl==2.3.5
oauthlib==1.0.3
pandas==0.18.0
pandas-datareader==0.2.1
…
```

Using ‘pip’

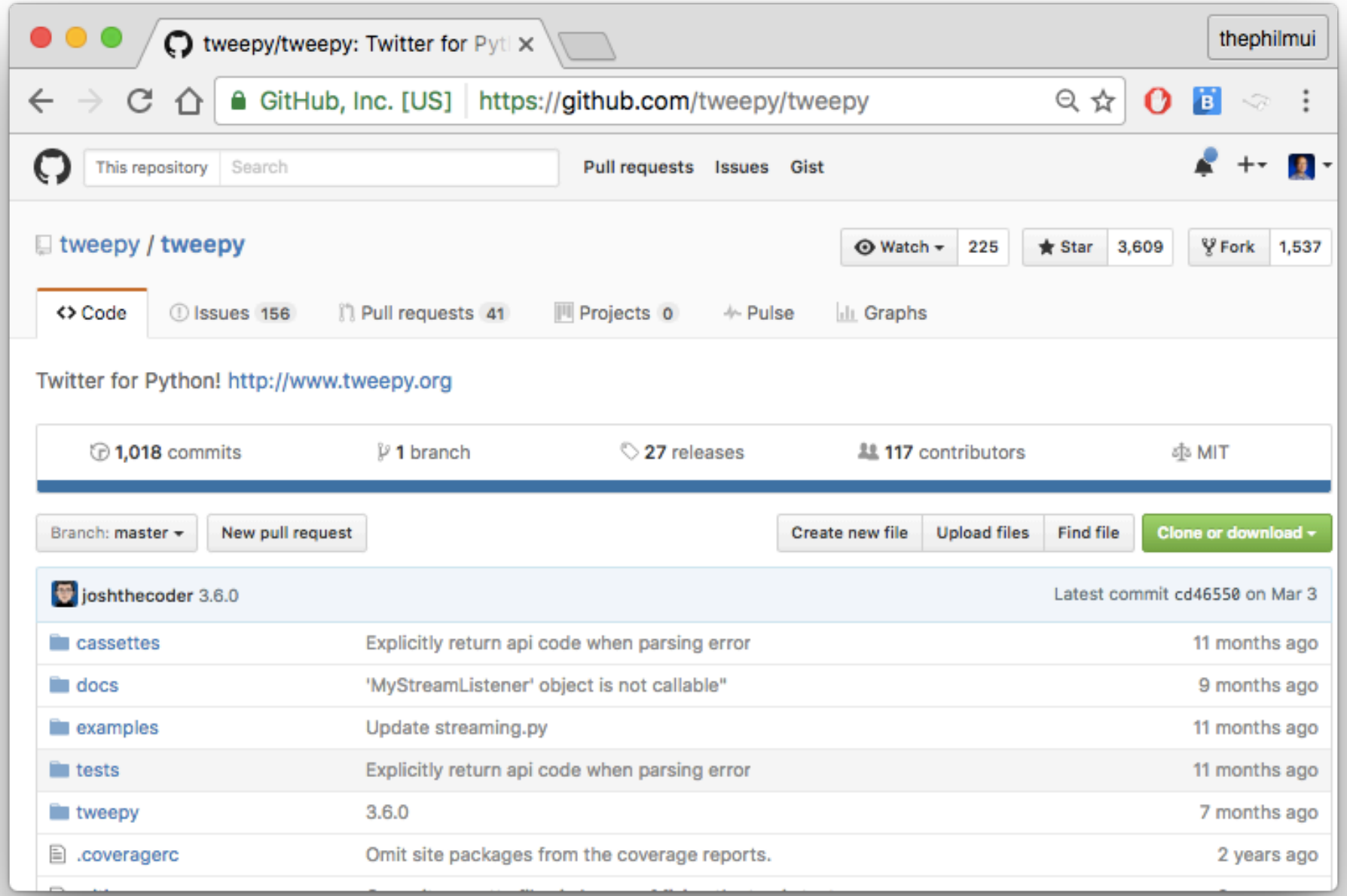
```
$ venv dsenv
$ source dsenv/bin/activate
$ pip install r requirements.txt
```

Starting Jupyter

```
$ source dsenv/bin/activate
(venv) $ jupyter notebook
```

Data Ingestion

lecture02.ingestion.ipynb


tweepy/tweepy: Twitter for Pyt
thephilmui
GitHub, Inc. [US] https://github.com/tweepy/tweepy
This repository Search
Pull requests Issues Gist
tweepy / tweepy
Watch 225 Star 3,609 Fork 1,537
Code Issues 156 Pull requests 41 Projects 0 Pulse Graphs
Twitter for Python! http://www.tweepy.org
1,018 commits 1 branch 27 releases 117 contributors MIT
Branch: master New pull request Create new file Upload files Find file Clone or download
joshthecoder 3.6.0 Latest commit cd46550 on Mar 3
cassettes Explicitly return api code when parsing error 11 months ago
docs 'MyStreamListener' object is not callable" 9 months ago
examples Update streaming.py 11 months ago
tests Explicitly return api code when parsing error 11 months ago
tweepy 3.6.0 7 months ago
.coveragerc Omit site packages from the coverage reports. 2 years ago

Submission for Completing CodeAcademy Assignment

This assignment will represent your completion of any Python online tutorial:

```
https://www.learnpython.org/
https://www.codecademy.com/learn/python
https://developers.google.com/edu/python/
```

Requirements for submission is on next page.

Submission Requirements:

- Use the Twitter API to a good sample of tweets about “trump” and “clinton”. (Suggestion: at least 100 tweets each).
- Count the most frequently associated words for both categories.
- Print the top 10 most associated words which are not “stop words”.
- Assign value 1 for "positive" sentiment, -1 for "negative" sentiment for each of the top 10 words.
- What is the average sentiment value (as represented by the top 10 words) for tweets associated with "trump" and "clinton"?
- Attach your code to your submission

Web Scraping

lecture02.web.scraping.ipynb