Introduction to Python

Ellie Bennett

University of Helsinki

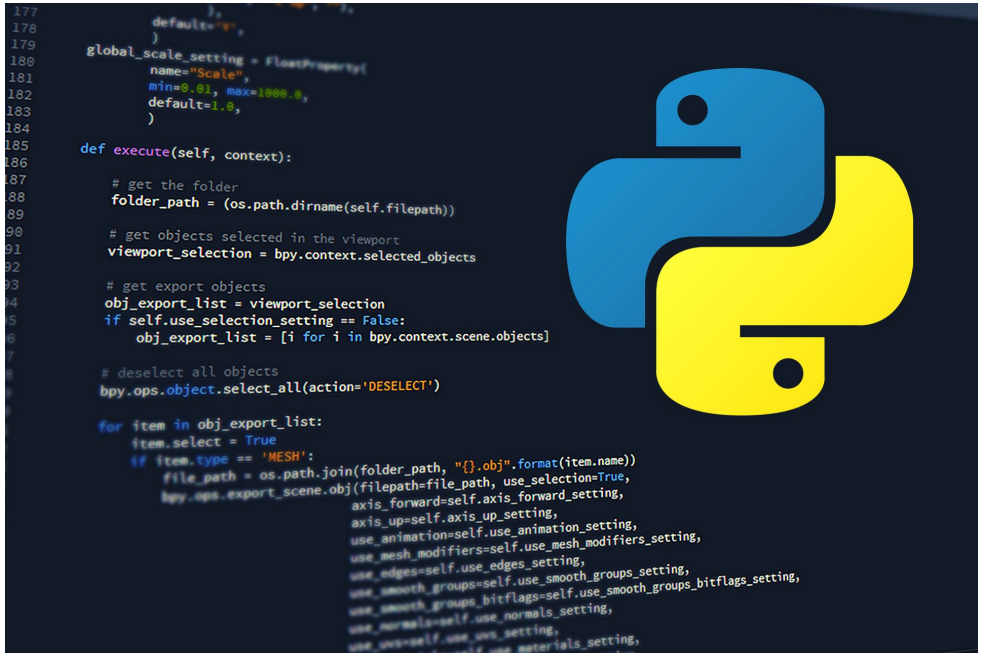
Outline of the module

- Quick introduction to Python and how it is used in digital humanities more widely
- Introduction to user interfaces
- Introduction to:
  - Pandas and dataframes
  - Cleaning data
  - Exploring your data
  - Saving as a .txt file
- Small group exercises

Scan me for the slides!

What is Python?

- Programming language
- Different versions
- Huge versatility and applicability
- Easy to learn and write code



What is Python used for in digital humanities?

Statistical Analysis

Especially powerful with tagged data like XML markup

Manipulating Tables

Can manipulate tables to organise data

Automated Data Cleaning

Python can find patterns in your data and standardise or clean it

Scraping Data from the Internet

Python code can go through internet pages and make a copy of the data it finds

General Purpose Coding Language

Lots of packages available mean it can do a lot of different things.

The large community using Python means there is a lot of support for any issues you might have.

User Interfaces: IDLE

IDLE Shell 3.9.7

File Edit Shell Debug Options Window Help

```
Python 3.9.7 (tags/v3.9.7:1016ef3, Aug 30 2021, 20:19:38) [MSC v.1
929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more inform
ation.
>>>
================ RESTART: D:\DAA 2023\Python\2. cleaning df.py ==
==============
```

Squeezed text (26304 lines).

```
>>>
========== RESTART: D:\DAA 2023\Python\2. cleaning df.py =========
```

Squeezed text (26304 lines).

```
>>>
========== RESTART: D:\DAA 2023\Python\2. cleaning df.py =========
Traceback (most recent call last):
  File "D:\DAA 2023\Python\2. cleaning df.py", line 8, in <module>
    noduplicates = table.drop_duplicates(subset=['id_text'], keep=
'last')
NameError: name 'table' is not defined
>>>
========== RESTART: D:\DAA 2023\Python\2. cleaning df.py =========
     id_text  ... pleiades_id
9    P313416  ...      874621
10   P313417  ...      874621
11   P313425  ...      874621
12   P313427  ...      874621
13   P313435  ...      874621
..       ...  ...         ...
268  P224433  ...      894019
269  P224447  ...      894019
270  P224485  ...      894019
271  P224487  ...      894019
272  P224587  ...      894019

[264 rows x 16 columns]
>>>
```

Ln: 31 Col: 4

2. cleaning df.py - D:\DAA 2023\Python\2. cleaning df.py (3.9.7)

File Edit Format Run Options Window Help

```
import pandas as pd

df = pd.read_csv('SAA1lemmaandmetadupes_03042023.csv', sep=',')#wh

#print(df.to_string()) #this is an enormous table, so it takes for

#remove duplicates
noduplicates = df.drop_duplicates(subset=['id_text'], keep='last')
print(noduplicates)

#re-order so lemma goes last

#remove all columns that aren't P numbers, designation, ancient au

#filtering according to a column
#ilu = df[df['word2']=='ilu[god]N']
#print(ilu)

#display only one column
#words = df[['word1']]
#print(words)

#saving the file
#savefile =  'dataframe.csv'
#with open(savefile, 'w', encoding="utf-8") as w:
#    ilu.to_csv(w, index=True) #beginning of this is the last data

#-----------------------------------------------------
#----EXTRA FANCY DATAFRAME STUFF WE WON'T DO TODAY----
#-----------------------------------------------------

#add two columns together
#we already have one column (words), so we need to make another
#words2 = df[['word2']]
#print(words2)
#rename column to 'word1'
#words2.columns = ['word1']
```

Ln: 8 Col: 17

User Interfaces: Jupyter Notebooks

- Widely used
- browser-based
- Can separate code into different blocks that can be run on their own
  - Useful for learning python and figuring out where your code has gone wrong

jupyter PMI info to Gephi Node Lists Last Checkpoint: 12/04/2021 (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

2426	udīni[up-till-now]AV	2427
2427	birtu[fort]N	2428
2428	libbu[interior]N	2429

2429 rows × 2 columns

In [11]:

```
#reorder columns (Id, then Label)
Nodes = Nodes[['Id', 'Label']]
Nodes
```

Out[11]:

	Id	Label
0	1	Aššur[1]DN
1	2	Ištar[1]DN
2	3	Lamaštu[1]DN
3	4	Ninurta[1]DN
4	5	Pazūzu[(a-demon)]N
...	...	...
2424	2425	šanû[be(come)-changed]V
2425	2426	pānu[front]N
2426	2427	udīni[up-till-now]AV
2427	2428	birtu[fort]N
2428	2429	libbu[interior]N

2429 rows × 2 columns

In [12]:

```
#remove duplicates
#Nodes = Nodes.drop_duplicates(subset=['Label'], keep='last')
#Nodes
```

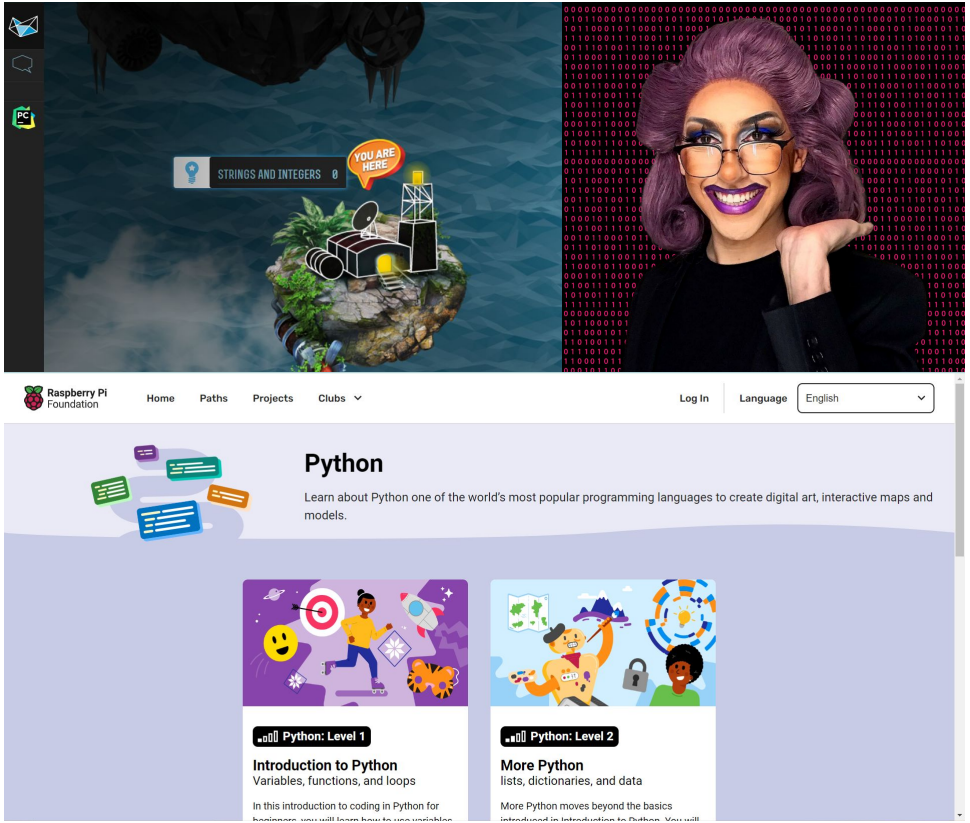
In [13]:

```
#save as a .csv file
savefile =  'NodeList_Top_20_2021_4_29.csv'
with open('output/' + savefile, 'w', encoding="utf-8") as w:
    Nodes.to_csv(w, index=False)
```

Small group questions

1. What percentage of the SAA1 dataset (not just the Nimrud texts) are personal names (lemmas ending in ‘PN’)?
    a. How many unique names are there?
    b. How does this compare to nouns (‘N’), verbs (‘V’), or adjectives (‘AJ’)?
    c. How does this compare to place names (‘GN’, ‘SN’)?
    d. How does this compare to god names (‘DN’) or temple names (‘TN’)?
2. How many times is ‘Ariye[1]PN’ mentioned in the whole dataset?
    a. How does this compare to the letters whole provenience is Nimrud?
    b. How does this compare to the letters whose provenience is Nineveh?
3. What are the unique place names (lemmas ending in ‘GN’, or ‘SN’) in the whole dataset?
    a. How does this compare to the letters whole provenience is Nimrud?
    b. How does this compare to the letters whose provenience is Nineveh?
4. How many unique ethnonyms (‘EN’) are the whole dataset?
5. In the SAA1 dataset, how many recipients are there compared to how many senders?
    a. How does this compare to the letters whole provenience is Nimrud?
    b. How does this compare to the letters whose provenience is Nineveh?

Resources



- Programming historian
- Python's beginner's guide
- Regular Expressions for Python library
- StackOverflow
- Free online course
- Youtube channel Python Tutorials for Digital Humanities (and accompanying website)
- University of Helsinki Applied Language Technology MOOC
- There are LOTS of games to help you learn!

Remember…

Scan me for the slides!

YOU’RE ALL CODERS NOW!