Python in 10 minutes

Part 6

Dr. Mark Williamson, PhD
Biostatistics, Epidemiology, and Research Design Core (BERDC)
Dakota Cancer Collaborative on Translational Activity (DaCCoTA)
University of North Dakota (UND)

Purpose:

- Quick, bite-size guides to basic usage and tasks in Python
- I’m no expert, I’ve just used it for various tasks, and it has made my life easier and allowed me to do things I couldn’t manually
- I’d like to share that working knowledge with you

Lesson 6: Editing data

Last time, we learned different ways to extract specific data and add it to lists, dictionaries, and new files. Today, we'll dive more into editing data. We'll examine how to:

1) manage whitespace

2) add content

3) delete content

4) change content

Lesson 6: The Dataset in Question

Fasta file of 16S rDNA from Archaea

- 116 sequences from the silva database (https://www.arb-silva.de/)
- Halobacteria -> Halobaterales -> Haloadaptaceae -> Haladaptatus
- Available at: https://med.und.edu/daccota/files/docs/berdc_docs/haladaptatus_16s.txt

First Entry

```
>AB663345.1.1471 Archaea;Halobacterota;Halobacteria;Halobacterales;Haloadaptaceae;Haladaptatus;Haladaptatus
cibarius
AUUCUGGUUGAUCCUGCCAGAGGCCAUUGCUAUUGGAGUUCGAUUUAGCCAUGCUAGUUGCACGAACUUAGAUUCGUAGC
GGAAAGCUCAGUAACACGUGAUCAAACUACCCUAUAGACCAGCAUAACCUCGGGAAACUGAGGCUAAUACUGGAUAACGC
UCCUACGUUUGAAUACGGGGAGCCGGAAACGCUCCGGCGCUAUAGGAUGUGAUUGCGGCCGAUUAGGUAGACGGUGGGGU
AACGGCCCACCGUGCCGAUAAUCGGUACGGGUUGUGAGAGCAAGAGCCCGGAGAUGGAUUCUGAGACAAGAAUCCAGGCC
CUACGGGGCGCAGCAGGCGCGAAAACUUUACACUGCACGACAGUGCGAUAAGGGAACUCCAAGUGCGAGGGCAUAUAGUC
CUCGCUUUUGUGUAUCGUAAGGUGGUACACGAAUAAGAGCUGGGCAAGACCGGUGCCAGCCGCCGCGGUAAUACCGGCAG
CUCGAGUGAUGGCCAAUUUUAUUGGGCCUAAAGCGUCCGUAGCCUGCCAGACAGGUCCGUCGGGAAAUCUGCUCGCUCAA
CGAGCAGGCGUCCGGCGGAAACCAGCUGGCUUGGGGCCGGAAGACCCAAGGGGUACGUCUGGGGUAGGAGUGAAAUCCUG
UAAUCCUAGACGGACCACCGAUCGCGAAAGCACCUUGGGAGGACGGACCCGACGGUGAGGGACGAAAGCUAGGGUCACGA
ACCGGAUUAGAUACCCGGGUAGUCCUAGCUGUAAACGAUGCUCGCUAGGUGUGGCACAGGCUACGAGCCUGUGCUGUGCC
ACAGUGAAGACGUAAGCGAGCCGCCUGGGAAGUACGUCUGCAAGGAUGAAACUUAAAGGAAUUGGCGGGGGAGCACUACA
ACCGGAGGAGCCUGCGGUUUAAUUGGACUCAACGCCGGACAUCUCACCAGCACCGACAAUAGCUGUGACGGUCAGUUUGA
UGAGCUUACUAGAGCUUUUGAGAGGAGGUGCAUGGCCGCCGUCAGCUCGUACCGUGAGGCAUCCUGUUAAGUCAGGCAAC
GAGCGAGAUCCGCGUCCGUAAUUGCCAGCAGCACCCUUGUGGUGGCUGGGUACAUUACGGAGACUGCCGCUGCUAAAGCG
GAGGAAGGAACGGGCAACGGUAGGUCAGCAUGCCCCGAAUGUGCUGGGCUACACGCGGGCUACAAUGGCCAAGACAAUGG
GUUCCAACCCCGAGAGGGGACGGUAAUCUCCGAAACUUGGUCGUAGUUCGGAUUGAGGGCUGAAACUCGCCCUCAUGAAG
CUGGAUUCGGUAGUAAUCGCGCUUCAGCAGAGCGCGGUGAAUACGUCCCUGCUCCUUGCACACACCGCCCGUCAAAGCAC
CCGAGUGAGGUCCGGAUGAGGCCAUCAGGCGAUGGUCGAAUCUGGGCUUCGCAAGGGGGCUUAAGUCGUAACAAGGUAGC
CGUAGGGGAAUCUGCGGCUGGAUCACCUCCU
```

Lesson 6: Whitespace management

Goal: Remove and add whitespace

Procedure

- Download the dataset and change it from a .txt to .fasta file (File-> Save As -> File name: haladaptatus.fasta, Save as type: All Files)
- Open Python and start a new file
- Create a **path** and **file** variable, and two outfiles

#add tab whitespace to sequence lines

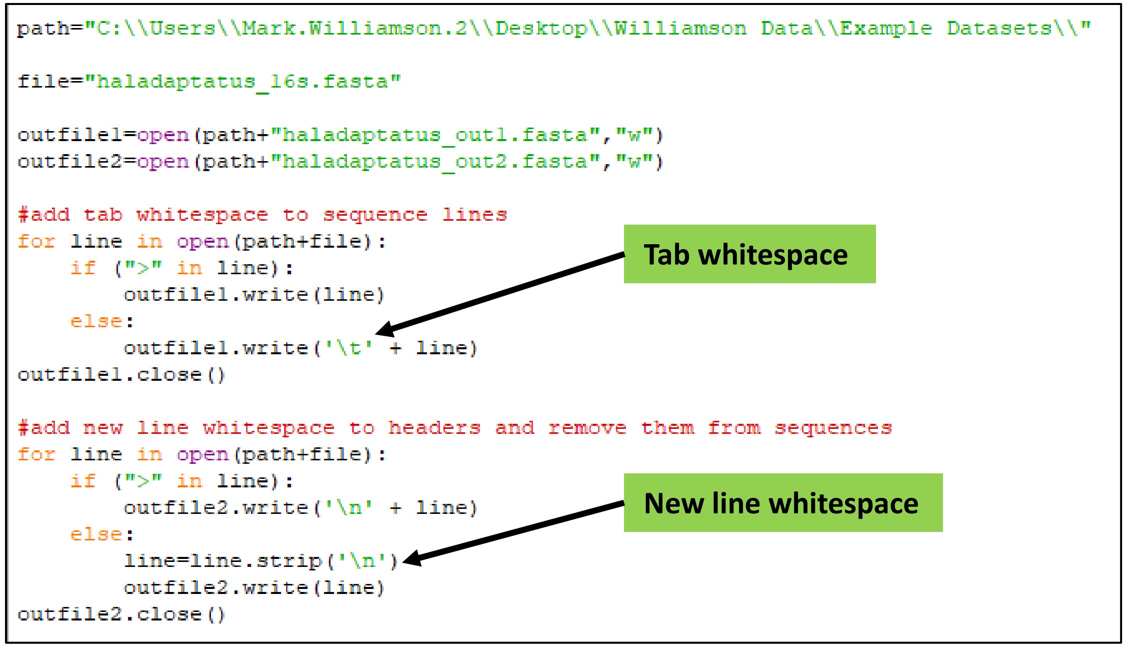
- Create a for-loop for each line
- Create an if-else statement that checks if “>” is in the line (indicates a header) and writes the to the outfile as normal if true
- Else, write to the outfile with a tab (‘\t’) followed by the line
- Close the outfile and view it in Notepad or Excel -> *Notepad will show the sequences tab-indented, while in Excel, the sequences will be indented by one column*

#add new line whitespace to headers and remove them from sequences

- This time, have the if-else statement add a new line (‘\n’) to the header and for the rest of the lines (sequences), strip the new line
- Close the outfile and view it in Excel -> *all the sequences will be one line*

Illumination

Whitespace is the term for characters used for spacing in python. They are invisible in a typical document (.doc, .txt, .xlsx, etc.) but show up in python and are important to deal with in certain situations. The two most relevant are \t (tab) and \n (new line).



```
path="C:\\Users\\Mark.Williamson.2\\Desktop\\Williamson Data\\Example Datasets\\"

file="haladaptatus_16s.fasta"

outfile1=open(path+"haladaptatus_out1.fasta","w")
outfile2=open(path+"haladaptatus_out2.fasta","w")

#add tab whitespace to sequence lines
for line in open(path+file):
    if (">" in line):
        outfile1.write(line)
    else:
        outfile1.write('\t' + line)
outfile1.close()

#add new line whitespace to headers and remove them from sequences
for line in open(path+file):
    if (">" in line):
        outfile2.write('\n' + line)
    else:
        line=line.strip('\n')
        outfile2.write(line)
outfile2.close()
```

Exploration

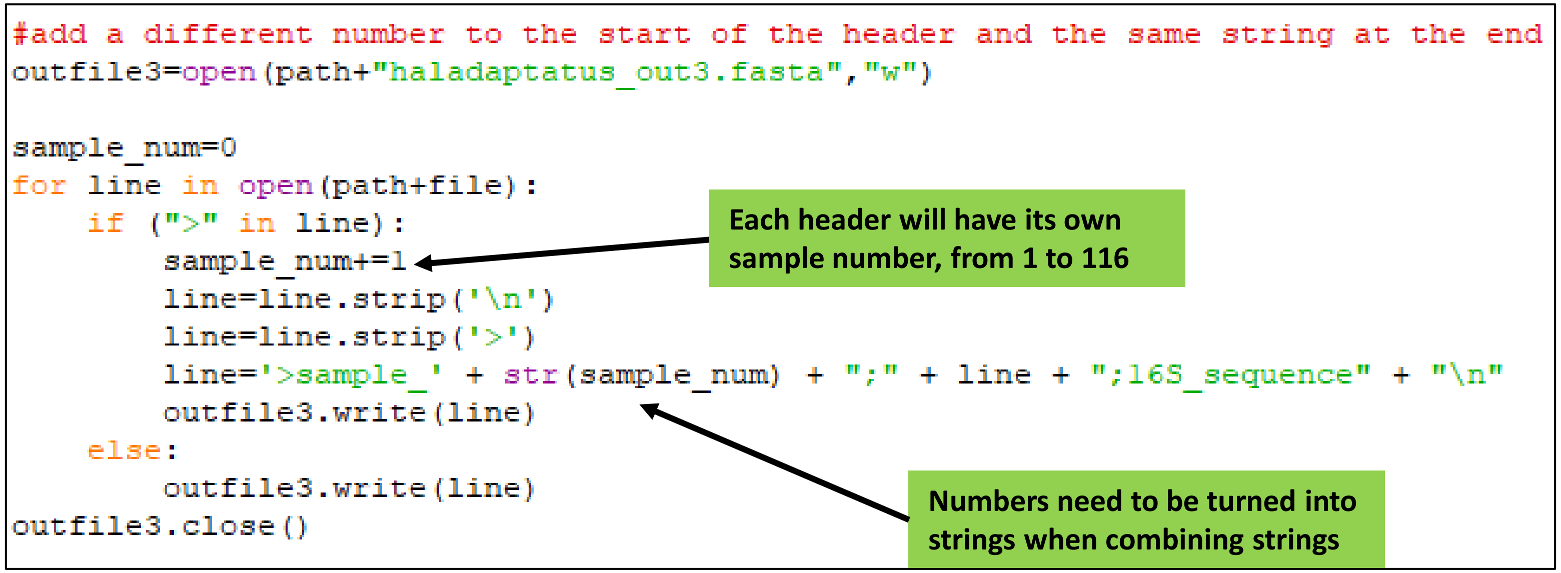
Try stripping all new lines (‘\n’) from all lines and write it to a new file. Then open it up in Excel to see the results

Lesson 6: Adding content

Goal: Add content to the headers

Procedure

- Create a new outfile and a variable called **sample_num** and set it to 0
- Inside a for-loop for each line, create an if-else statement that checks if “>” is in the line
- If true:
    - Add 1 to **sample_num**
    - Then, edit the line by stripping the new line whitespace and the ‘>’ string from it
    - Next, further edit by adding a string (>sample_), a string of the **sample_num**, a string of a semicolon, and ending with another string (;16S_sequence), and a new line whitespace
    - Finally, write all that to the outfile
- If false, simply write the line to the outfile



```
#add a different number to the start of the header and the same string at the end
outfile3=open(path+"haladaptatus_out3.fasta","w")

sample_num=0
for line in open(path+file):
    if (">" in line):
        sample_num+=1
        line=line.strip('\n')
        line=line.strip('>')
        line='>sample_' + str(sample_num) + ";" + line + ";16S_sequence" + "\n"
        outfile3.write(line)
    else:
        outfile3.write(line)
outfile3.close()
```

Each header will have its own sample number, from 1 to 116

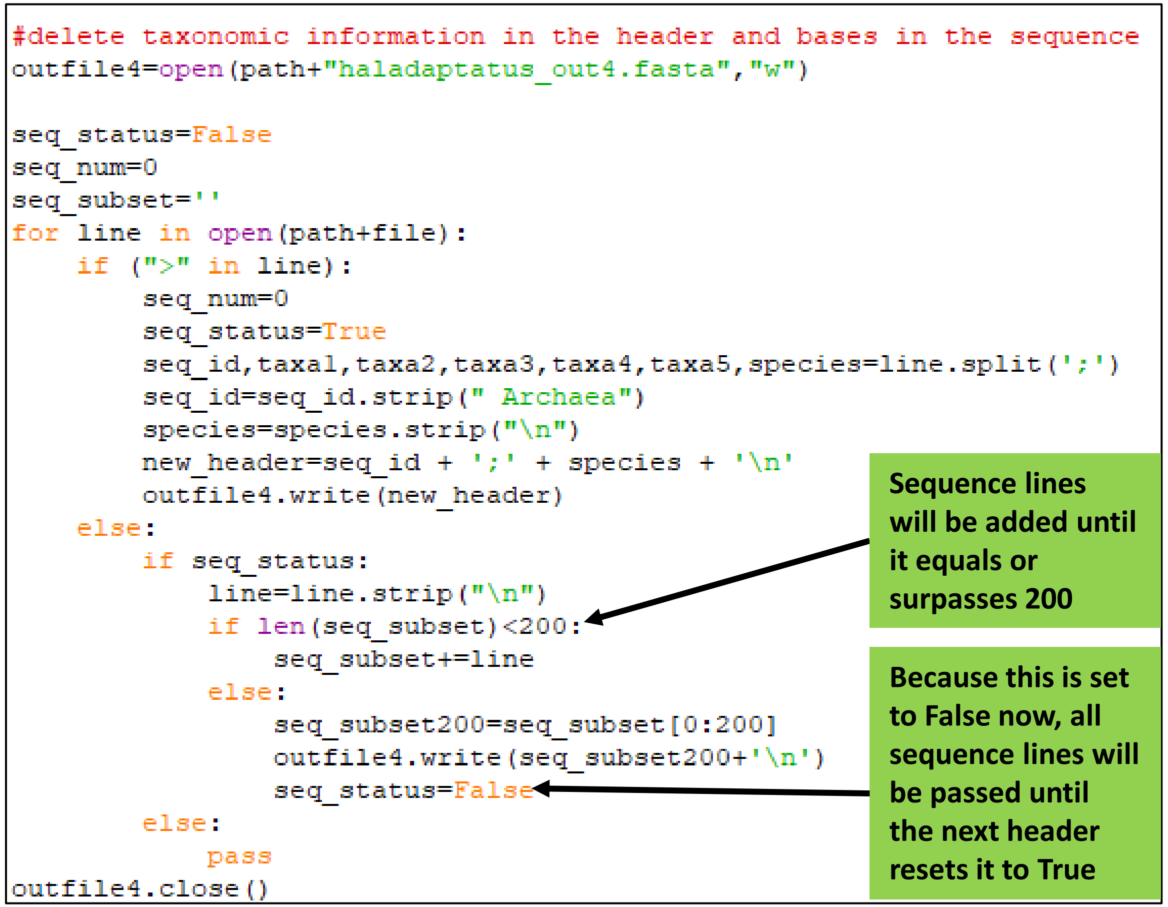
Numbers need to be turned into strings when combining strings

Lesson 6: Subtracting content

Goal: Subtract content from the headers and sequences

Procedure

- Create a new outfile and the variables **seq_status** (=False), **seq_num** (=0), and **seq_subset** (='')
- Inside a for-loop for each line, create an if-else statement that checks if ">" is in the line
- If true:
    - Set **seq_num** to 0 and **seq_status** to True
    - Split the header into seven variables using line.split
    - Strip " Archaea" from **seq_id** and the new line whitespace from **species**
    - Create a new header that only includes the modified **seq_id** and **species**, followed by a new line whitespace and write this to the outfile
- If false:
    - Create an if-else statement that check to see if **seq_status** is true
    - If true:
        - Strip the whitespace from the line and create another if-else statement
            - If the length of **seq_subset** is less than 200, add the line to **seq_subset**
            - Otherwise, create a new variable with only the first 200 bases, write that to the outfile and set **seq_status** to False
    - If false, pass



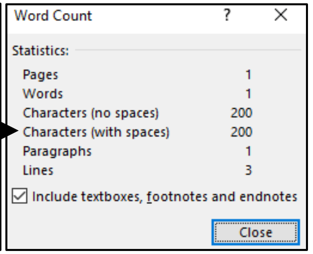
```
#delete taxonomic information in the header and bases in the sequence
outfile4=open(path+"haladaptatus_out4.fasta","w")

seq_status=False
seq_num=0
seq_subset=''
for line in open(path+file):
    if (">" in line):
        seq_num=0
        seq_status=True
        seq_id,taxa1,taxa2,taxa3,taxa4,taxa5,species=line.split(';')
        seq_id=seq_id.strip(" Archaea")
        species=species.strip("\n")
        new_header=seq_id + ';' + species + '\n'
        outfile4.write(new_header)
    else:
        if seq_status:
            line=line.strip("\n")
            if len(seq_subset)<200:
                seq_subset+=line
            else:
                seq_subset200=seq_subset[0:200]
                outfile4.write(seq_subset200+'\n')
                seq_status=False
        else:
            pass
outfile4.close()
```

Sequence lines will be added until it equals or surpasses 200

Because this is set to False now, all sequence lines will be passed until the next header resets it to True

Exploration

Open the outfile, copy and past one of the sequence lines into a word document, and check the number of characters using the review tool. It should equal 200.



Word Count

Statistics:

Pages	1
Words	1
Characters (no spaces)	200
Characters (with spaces)	200
Paragraphs	1
Lines	3

☑ Include textboxes, footnotes and endnotes

Lesson 6: Changing content

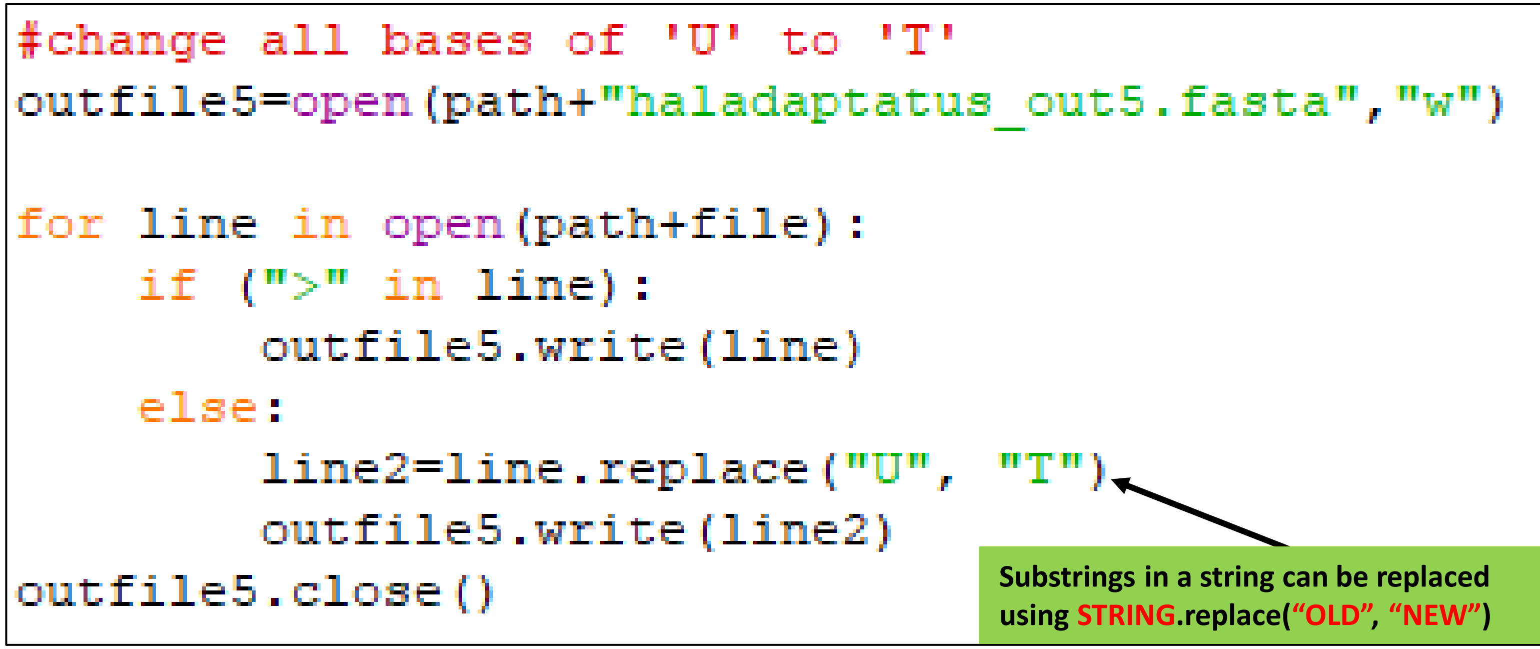
Goal: Change all ‘U’ bases to ‘T’ in the sequences

Procedure

- Create a new outfile
- Inside a for-loop for each line, create an if-else statement that checks if “>” is in the line
- If true:
    - write the line to the outfile as normal
- If false:
    - Create a new variable called **line2** and replace all instances of ‘U’ with ‘T’
    - Write **line2** to the outfile

Illumination

Biopython (https://biopython.org/) is an add-on that is useful for genomic data. It has a variety of functions, including back_transcribe(), which is what this example is doing manually.



```
#change all bases of 'U' to 'T'
outfile5=open(path+"haladaptatus_out5.fasta","w")

for line in open(path+file):
    if (">" in line):
        outfile5.write(line)
    else:
        line2=line.replace("U", "T")
        outfile5.write(line2)
outfile5.close()
```

Substrings in a string can be replaced using STRING.replace(“OLD”, “NEW”)

Lesson 6: Changing content cont.

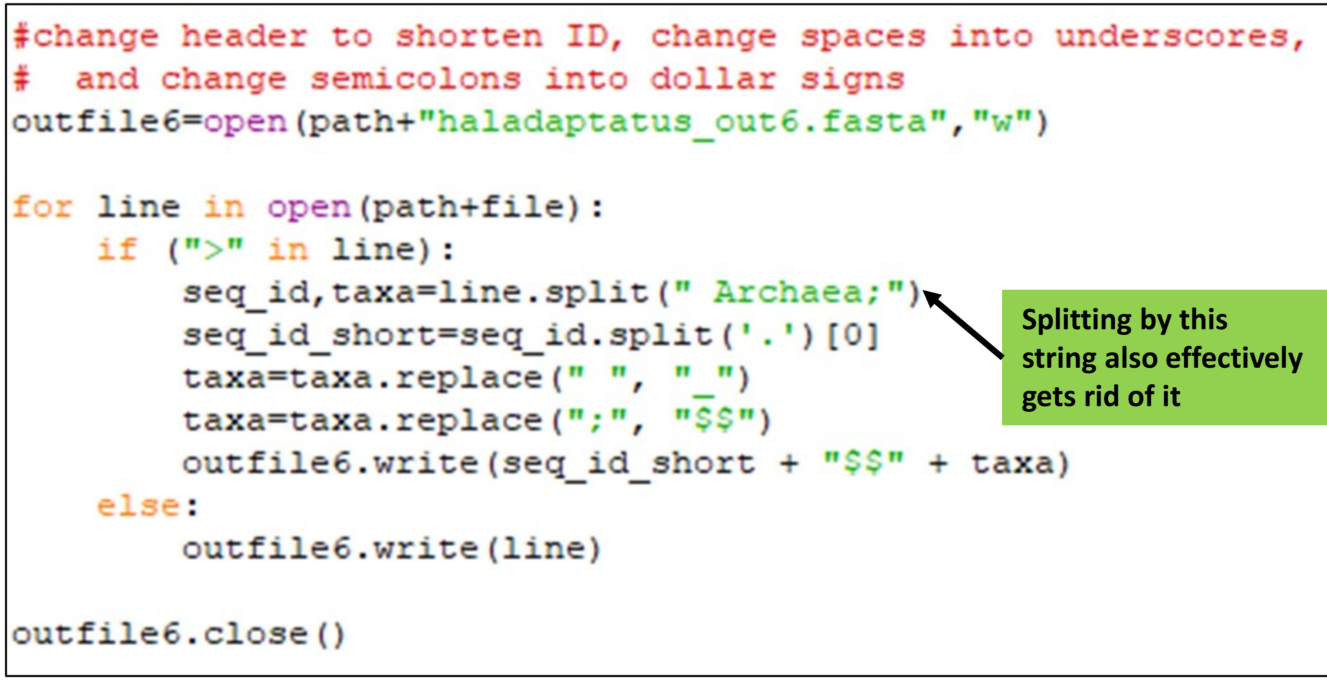
Goal: Change the headers by shortening the sequence ID, change any spaces into underscore, and change semicolons into dollar signs

Procedure

- Create a new outfile
- Inside a for-loop for each line, create an if-else statement that checks if “>” is in the line
- If true:
  - Split the line into **seq_id** and **taxa** by the string “ Archaea”
  - Shorten **seq_id** by splitting by a period and only including the first split [0]
  - In taxa, use .replace() for both spaces and semicolons to turn them into underscores or double dollar signs, respectively
  - Write the resulting **seq_id** and **taxa** to the outfile with ‘$$’ between them
- If false:
  - write the line to the outfile as normal

Illumination

There are times where it is very useful to change one string to another (; -> $$). For example, some procedures only recognize certain syntax, or you may want to avoid splitting with a commonly used symbol because it would split in incorrect places.



```
#change header to shorten ID, change spaces into underscores,
#  and change semicolons into dollar signs
outfile6=open(path+"haladaptatus_out6.fasta","w")

for line in open(path+file):
    if (">" in line):
        seq_id,taxa=line.split(" Archaea;")
        seq_id_short=seq_id.split('.')[0]
        taxa=taxa.replace(" ", "_")
        taxa=taxa.replace(";", "$$")
        outfile6.write(seq_id_short + "$$" + taxa)
    else:
        outfile6.write(line)

outfile6.close()
```

Splitting by this string also effectively gets rid of it

Exploration

Try replacing the “;” in the header by your own replacement (ex. “#####” or “|” or even “¯_(ツ)_/¯”)

Lesson 6: Summary

- Python can edit data for a variety of purposes
- Whitespace can be added or deleted to change how lines of data are formatted in new files
- Content can be added, subtracted, or changed in lines, strings, and other structures
- Please complete a brief, 5-question assessment:

  https://und.qualtrics.com/jfe/form/SV_25bRWt0UWqvGEa9