Programming for Data Science: File I/O lab

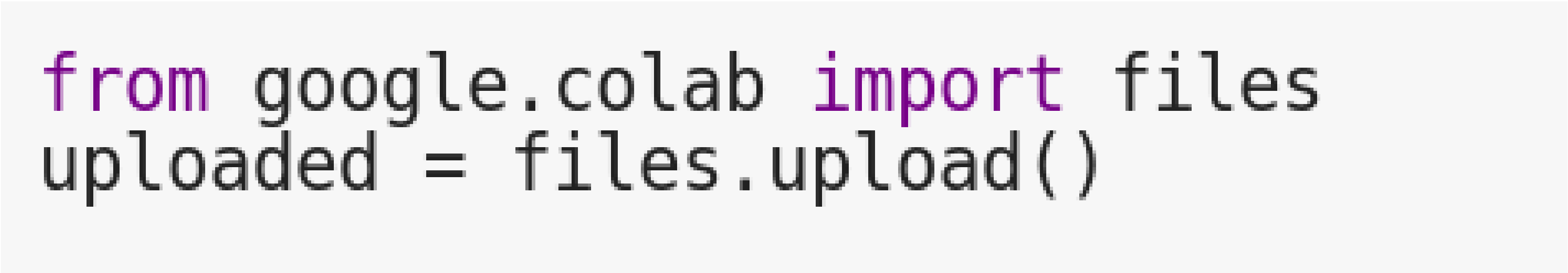
Instructor: Jakramate Bootkrajang

Outlines

- Uploading file to colab
- DNA sequence file
- Reading from file
- Exercises

Colab file uploading

- In order to practice opening file we need a file to begin with
- You can upload files to Colab using the following code snippet



```
from google.colab import files
uploaded = files.upload()
```

- The uploaded file is only available for the current login session

DNA sequence file

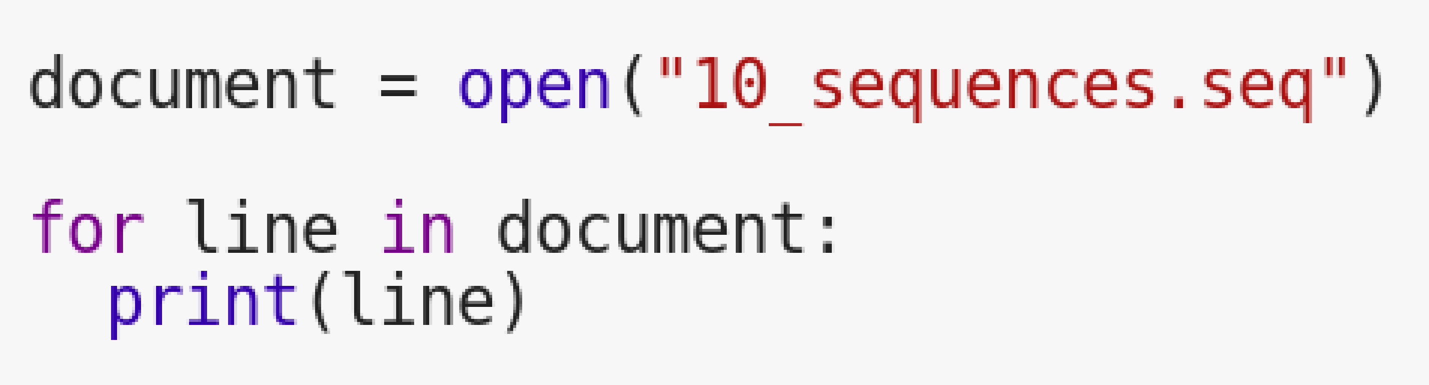
- Download the file from
- http://www.dalkescientific.com/writings/NBN/data/10_sequences.seq
- Save the file in your PC
- Upload the saved file to Colab using the code snippet in the previous slide

Processing sequences stored in a file

- We’ve previously worked with DNA sequence and did some interesting processing
- We will now generalise the process to cover multiple sequences.
- The sequences are stored in 10_sequences.seq file

Reading lines from a file

- Using open() function and loop over each line
- Document is a variable that stores input file handler
- Recall that file handler can be iterated

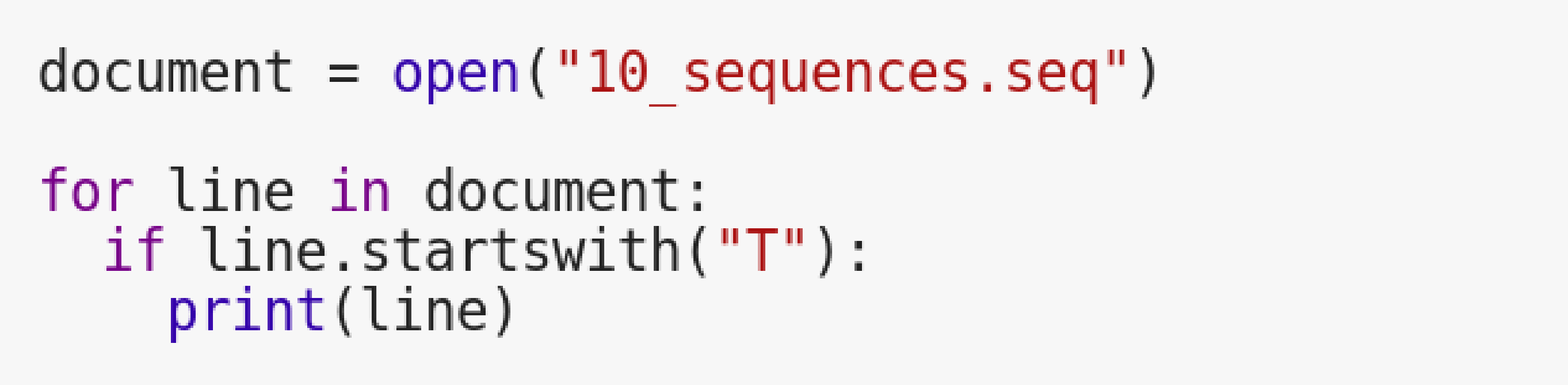


```
document = open("10_sequences.seq")

for line in document:
  print(line)
```

More complex example

- List the sequences starting with a thymine



```
document = open("10_sequences.seq")

for line in document:
  if line.startswith("T"):
    print(line)
```

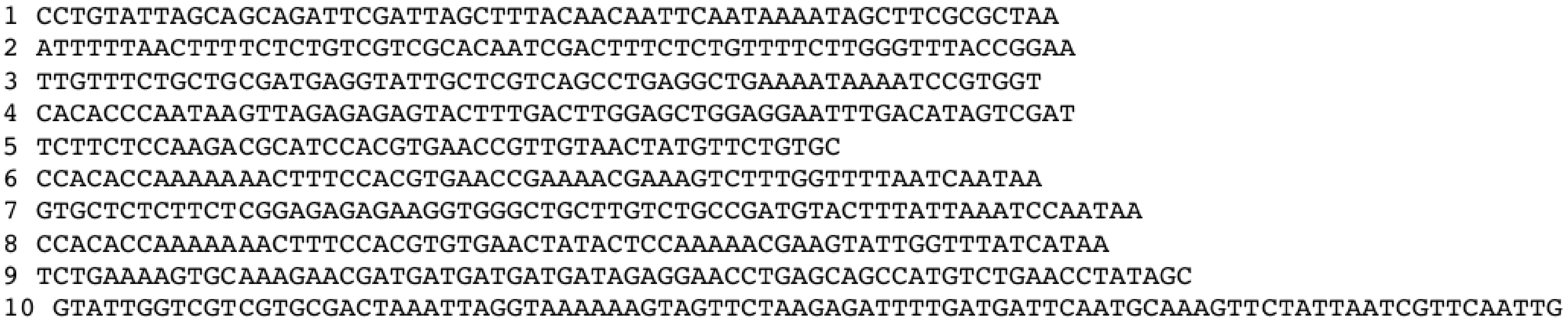
- Each line contains additional ‘\n’, what is its effect ?
- How to remove the extra ‘\n’ ?

rstrip() and lstrip()

- To remove extra white space at the end use rstrip() method
- To remove extra white space at the beginning of a string use lstrip() method

Exercise 1: Number lines in a file

- Read the file 10_sequences.seq. Print out the line number (starting with 1) then the line. Remember to use rstrip() to remove the extra newline.
- The output should look like this



```
1  CCTGTATTAGCAGCAGATTCGATTAGCTTTACAACAATTCAATAAAATAGCTTCGCGCTAA
2  ATTTTTAACTTTTCTCTGTCGTCGCACAATCGACTTTCTCTGTTTTCTTGGGTTTACCGGAA
3  TTGTTTCTGCTGCGATGAGGTATTGCTCGTCAGCCTGAGGCTGAAAATAAAATCCGTGGT
4  CACACCCAATAAGTTAGAGAGAGTACTTTGACTTGGAGCTGGAGGAATTTGACATAGTCGAT
5  TCTTCTCCAAGACGCATCCACGTGAACCGTTGTAACTATGTTCTGTGC
6  CCACACCAAAAAAACTTTCCACGTGAACCGAAAACGAAAGTCTTTGGTTTTAATCAATAA
7  GTGCTCTCTTCTCGGAGAGAGAAGGTGGGCTGCTTGTCTGCCGATGTACTTTATTAAATCCAATAA
8  CCACACCAAAAAAACTTTCCACGTGTGAACTATACTCCAAAAACGAAGTATTGGTTTATCATAA
9  TCTGAAAAGTGCAAAGAACGATGATGATGATGATAGAGGAACCTGAGCAGCCATGTCTGAACCTATAGC
10  GTATTGGTCGTCGTGCGACTAAATTAGGTAAAAAAGTAGTTCTAAGAGATTTTGATGATTCAATGCAAAGTTCTATTAATCGTTCAATTG
```

Exercise 2

- List the sequences in 10_sequences.seq which have the pattern CTATA

Exercise 3

- Modify the previous program to print the index of the first time CTATA pattern is found

Exercise 4

- Based on sequences.seq write a program to answer the following questions

1) How many sequences are in that file ?

2) How many have the pattern CTATA ?

3) How many have more than 500 bases ?

4) How many have over 50% GC composition ?

GC composition is the ratio of the number of G and C over the length of the sequence

References

- http://www.dalkescientific.com/writings/NBN/