text manipulation - parsing.

files and file manipulations

name_handle = open(‘students’, ‘r’)

slicing and splitting

slicing and splitting

Create a function ‘extract_acc’ that extracts the Accession Number from the protein fasta file.

Tip: If you are not familiar with fasta file format - open the file and identify the common patterns. The accession numbers are the first 12 characters after each ‘>’.

Example file: pelagibacter.faa

text manipulation

regular expressions - quick guide

^ matches the beginning of a line
$ matches the end of a line
. matches any character
\s matches whitespace
\S matches any non-whitespace character
* repeats a character zero or more times
*? repeats a character zero or more times (none-greedy)
+ repeats a character one or more times
+? repeats a character one or more times (none-greedy)
[aeiou] matches a single character in the listed set
[^XYZ] matches a single character not in the listed set
[a-z0-9] the set of characters can include a range
(Indicates where string extraction is to start
) Indicates where string extraction is to end

regular expressions - quick guide

Change the function so you can collect multiple occurrences. For example search for ‘hydrolases’ and collect the fasta headers in a list.

modules = python files with .py extension

What are modules?

os

sys module

third party modules
matplotlib
pandas
numpy

FIN