Problem #1

Subject: Data Mining

Filename: hw06_01.py
Allowed modules: re, pathlib

In this problem you are provided with a zip file containing a single top-level folder. This folder has many sub-folders, with each of these contains many files. Inside these files are lines of random strings, but some of these strings contain numbers that are delimited by dollar signs. A given line will have, at most, one such number. The length of these numbers varies, and the only ones we are interested in are those numbers that are exactly five digits long. Note that this is the length of the string representing the number, not the number itself. For instance, 00318 is a five-digit number, while 073914 is not. Your task is to write a script that scans the contents of these files, finds the five-digit numbers, and sums them up. Your script should output the following information, reasonably presented:

WORK

Pseudo code:

```
Setup regex as string
For folders in data
    Subfolders++
    For files in subfolders
        Files++
        For each line
            Find regex
            Add number to total
            Total_numbers++
Print following:
Total subfolders
Total files
Total 5 digit numbers
Sum of all 5 digit numbers
```

OUTPUT

```
Number of subfolders: 50
Number of files: 1000
Number of 5 digit numbers: 4745
Sum of numbers: 235762121
```

CODE

```
'''
PROGRAMMER: Christopher D Colbert
USERNAME: ccolbert
PROGRAM: hw06_01.py

DESCRIPTION: Given a folder containing subfolders and files within those subfolders,
find 5 digit long numbers and sum total. Prints total files, total subfolders, total
number of 5 digit numbers, and sum of numbers.
'''
import re, pathlib
total = 0
subfolders = 0
files = 0
num_total = 0

regex = '\$([0-9]{5})\$'
p = pathlib.Path('data')
```

```
#for subfolders in main file
for folder in p.iterdir():
  subfolders += 1


  #for files in each subfolder
  for file in folder.iterdir():
    files +=1


    #open file
    with file.open() as f:
      #for each line find regex
      for line in f:
        numbers = re.findall(regex, line)


        #extract number from list returned from findall
        for number in numbers:
          num_total += 1
          total += int(number)


print(f"Number of subfolders: {subfolders}")
print(f"Number of files: {files}")
print(f"Number of 5 digit numbers: {num_total}")
print(f"Sum of numbers: {total}")
```