Assignment 1: Raw Data Processing CSCE 5310 – Empirical Analysis Due: September 18, 2022, by 11:59pm

Working with raw data is not easy, data coming from the real world is generally messy. In this assignment you will work with files containing raw data and try to parse and manipulate the data. Download zoo dataset files from Canvas and write code to convert the raw data into more human-readable CSV format.

aardvark,1,0,0,1,0,0,1,1,1,1,0,0,4,0,0,1,1 antelope,1,0,0,1,0,0,0,1,1,1,0,0,4,1,0,1,1 bass,0,0,1,0,0,1,1,1,1,0,0,1,0,1,0,0,4

Lot of times it is important to keep the data into compact machine-readable formats when the data is large, but this data is not large. Goal is to write a code to convert it to readable CSV file (Python preferably, if you are comfortable in some other programming language, talk to me). Note that there are no column headers in the data above, so not clear what these 1s and 0s are telling you need to cross-reference them from another file. Specifically, one would like the output to look like:

AnmialName,Hair,Feathers,Eggs,Milk,Airborne,Aquatic,Pedator...
aardvark,Yes,No,No,Yes,No,NO,Yes,Yes,Yes,Yes,...
antelope,Yes,No,No,Yes,No,No,No,Yes,Yes,Yes,No,...

To do this, you need to check both files “zoo.data” and “zoo.names”. The information in “zoo.names” is relatively unstructured, you can find the Attribute Information in this file. You also need to look at types of each attribute (Boolean, Numeric, etc.). Next, you need to load the data from “zoo.data”, convert Boolean value 1 to “Yes” and 0 to “No”, make sure Numeric remain numeric. You need to write the values to CSV file.

In [54]:

```
import pandas as pd
import csv

#below class is to return type of animal as given in the zoo.names document


class Animal():
    def typeofanimal(self, ani):
        class1 = ["aardvark", "antelope", "bear", "boar", "buffalo", "calf", "cav
                  "porpoise", "puma", "pussycat", "raccoon", "reindeer", "seal",
        class2 = ["chicken", "crow", "dove", "duck", "flamingo", "gull", "hawk",
        class3 = ["pitviper", "seasnake", "slowworm", "tortoise", "tuatara"]
        class4 = ["bass", "carp", "catfish", "chub", "dogfish", "haddock", "herri
        class5 = ["frog", "frog", "newt", "toad"]
        class6 = ["flea", "gnat", "honeybee", "housefly", "ladybird", "moth", "te
        class7 = ["clam", "crab", "crayfish", "lobster", "octopus", "scorpion", "
        if ani in class1:
            return 1
        elif ani in class2:
            return 2
        elif ani in class3:
            return 3
        elif ani in class4:
            return 4
        elif ani in class5:
            return 5
        elif ani in class6:
            return 6
        elif ani in class7:
            return 7
```

In [55]:

```
#input and output files are declared

input_filename = 'zoo-1.data'
output_filename = 'output.csv'

#opened input and output file

with open(input_filename, 'r', newline='') as infile, \
     open(output_filename, 'w', newline='') as outfile:
    reader = csv.reader(infile, skipinitialspace=True)
    writer = csv.writer(outfile)

#column headings for the output file
    heading = ['Animal Name', "Hair", "Feathers", "Eggs", "Milk", "Airborne", "Ac
             "Toothed", "Backbone", "Breathes", "Venomous", "Fins", "Legs", "Tai
    writer.writerow(heading)
    #i stands for row input
    i = []
    leg = ['0','2','4','5','6','8']
    # loop in to read each row in the input file
    for row in reader:
        i = row
        #loop in to get 18 attributes
        for x in range(0,18):
            if x == 0:
                #get animal name into a variable
                ani = i[x]
            elif x == 13:
                #get numberic value of number of legs (set of values: {0,2,4,5,6,
                if i[x] in leg:
                    i[x] = int(i[x])
                else:
                    i[x]== 0
            elif x == 17:
                # call Animals class
                d = Animal()
                #call "typeofanimal" function in Animal class to get the type of
                h = d.typeofanimal(ani)
                i[x] = int(h)
            else:
                #for other boolean attributes change 0 to no and 1 to yes
                if i[x]== "1":
                    i[x] = "yes"
                elif i[x] == "0":
                    i[x] = "no"
                elif i[x] == " ":
                     i[x] = "none"
        #finally write the row into the output file
        writer.writerow(i)
```

In [56]:
```
#now read the output file to display the data
df = pd.read_csv("output.csv")
print(df.shape)
```

```
(101, 18)
```

In [57]:
```
display(df)
```

	Animal Name	Hair	Feathers	Eggs	Milk	Airborne	Aquatic	Predator	Toothed	Backbone	Br
0	aardvark	yes	no	no	yes	no	no	yes	yes	yes	
1	antelope	yes	no	no	yes	no	no	no	yes	yes	
2	bass	no	no	yes	no	no	yes	yes	yes	yes	
3	bear	yes	no	no	yes	no	no	yes	yes	yes	
4	boar	yes	no	no	yes	no	no	yes	yes	yes	
...	...	...	...	...	...	...	...	...	...	...	
96	wallaby	yes	no	no	yes	no	no	no	yes	yes	
97	wasp	yes	no	yes	no	yes	no	no	no	no	
98	wolf	yes	no	no	yes	no	no	yes	yes	yes	
99	worm	no	no	yes	no	no	no	no	no	no	
100	wren	no	yes	yes	no	yes	no	no	no	yes	

In [ ]:

In [ ]:

In [ ]: