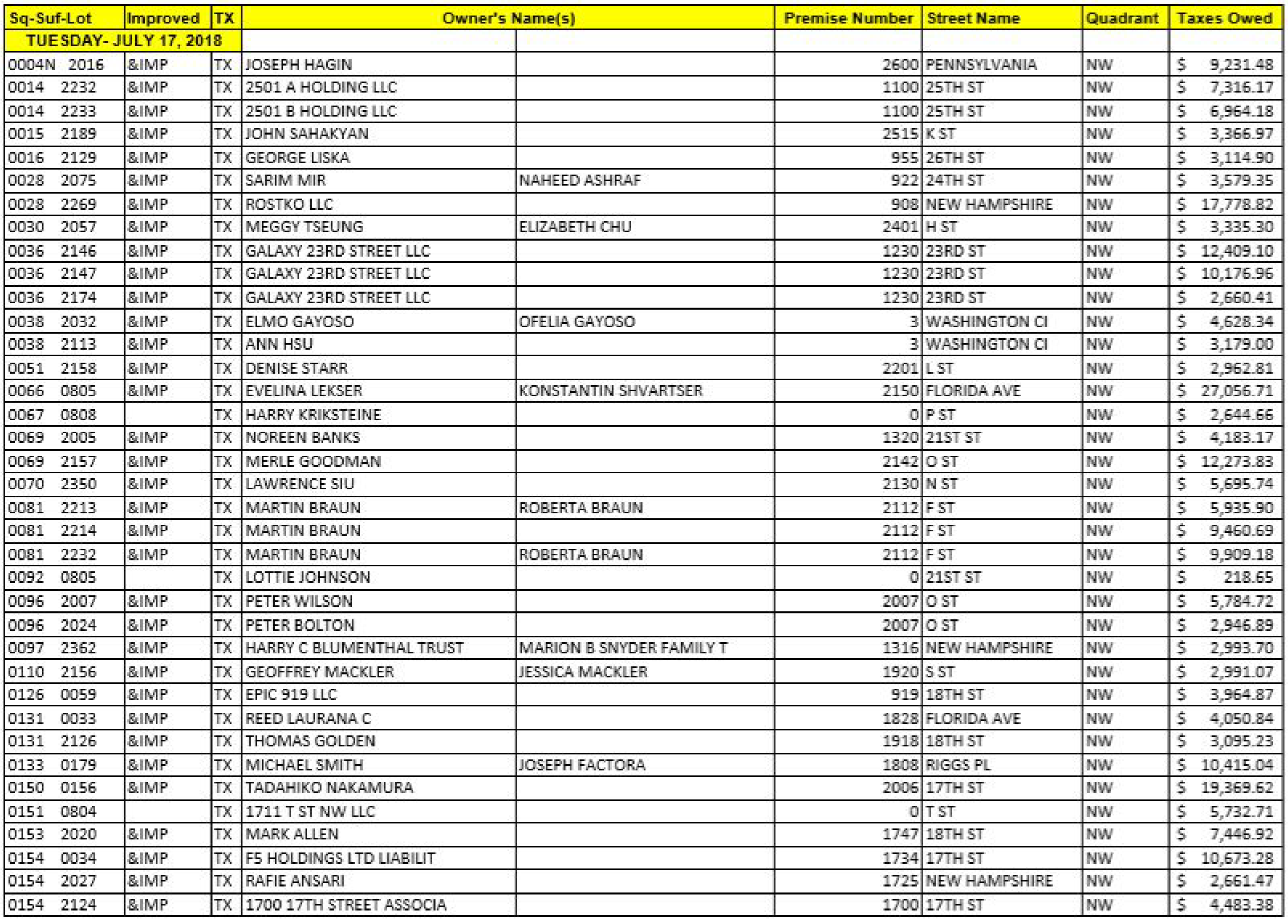
Taxes Owed List

Notes: This python scripts scrapes a folder of PDFs from a Taxes Owed List and creates a csv output file for each PDF. The following column headers are created and the associated data is collected: **Suf-Lot and Taxes Owed**. The PDF filenames are not unique and each CSV file is named based on the current number in a series. Two formats of parcel (Sq-Suf-Lot) where accounted for in the code.

PDF



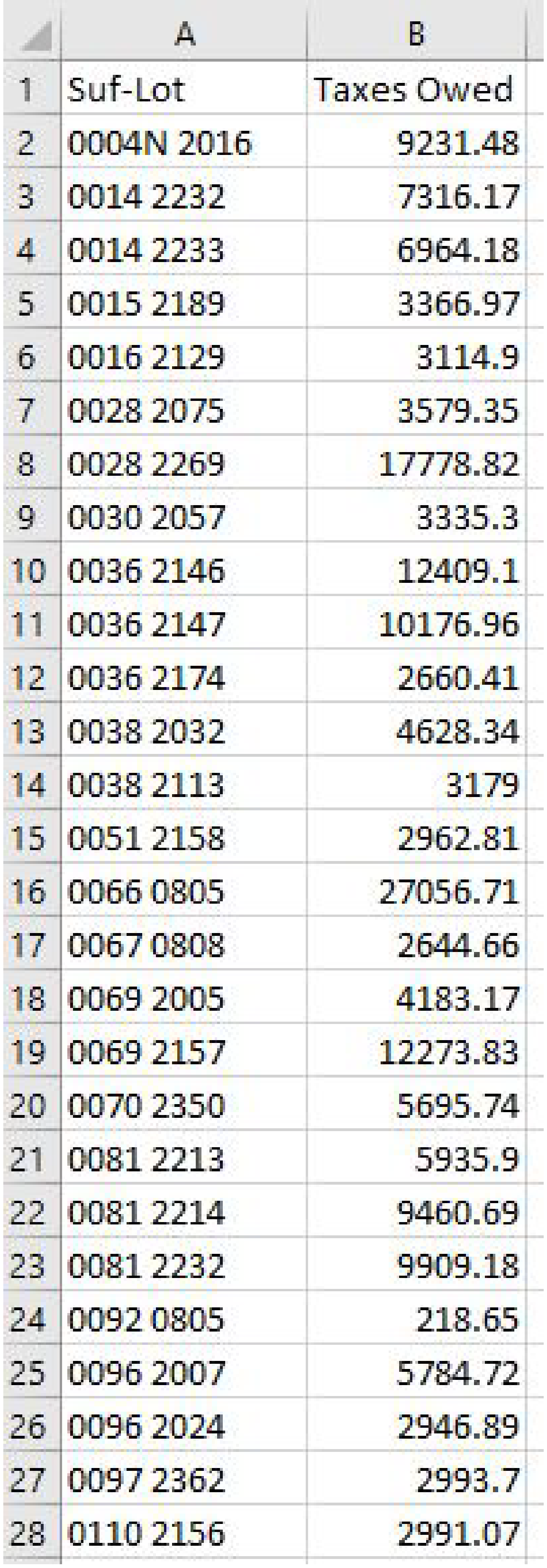
Sq-Suf-Lot	Improved	TX	Owner's Name(s)		Premise Number	Street Name	Quadrant	Taxes Owed
TUESDAY- JULY 17, 2018								
0004N 2016	&IMP	TX	JOSEPH HAGIN		2600	PENNSYLVANIA	NW	$ 9,231.48
0014 2232	&IMP	TX	2501 A HOLDING LLC		1100	25TH ST	NW	$ 7,316.17
0014 2233	&IMP	TX	2501 B HOLDING LLC		1100	25TH ST	NW	$ 6,964.18
0015 2189	&IMP	TX	JOHN SAHAKYAN		2515	K ST	NW	$ 3,366.97
0016 2129	&IMP	TX	GEORGE LISKA		955	26TH ST	NW	$ 3,114.90
0028 2075	&IMP	TX	SARIM MIR	NAHEED ASHRAF	922	24TH ST	NW	$ 3,579.35
0028 2269	&IMP	TX	ROSTKO LLC		908	NEW HAMPSHIRE	NW	$ 17,778.82
0030 2057	&IMP	TX	MEGGY TSEUNG	ELIZABETH CHU	2401	H ST	NW	$ 3,335.30
0036 2146	&IMP	TX	GALAXY 23RD STREET LLC		1230	23RD ST	NW	$ 12,409.10
0036 2147	&IMP	TX	GALAXY 23RD STREET LLC		1230	23RD ST	NW	$ 10,176.96
0036 2174	&IMP	TX	GALAXY 23RD STREET LLC		1230	23RD ST	NW	$ 2,660.41
0038 2032	&IMP	TX	ELMO GAYOSO	OFELIA GAYOSO	3	WASHINGTON CI	NW	$ 4,628.34
0038 2113	&IMP	TX	ANN HSU		3	WASHINGTON CI	NW	$ 3,179.00
0051 2158	&IMP	TX	DENISE STARR		2201	L ST	NW	$ 2,962.81
0066 0805	&IMP	TX	EVELINA LEKSER	KONSTANTIN SHVARTSER	2150	FLORIDA AVE	NW	$ 27,056.71
0067 0808		TX	HARRY KRIKSTEINE		0	P ST	NW	$ 2,644.66
0069 2005	&IMP	TX	NOREEN BANKS		1320	21ST ST	NW	$ 4,183.17
0069 2157	&IMP	TX	MERLE GOODMAN		2142	O ST	NW	$ 12,273.83
0070 2350	&IMP	TX	LAWRENCE SIU		2130	N ST	NW	$ 5,695.74
0081 2213	&IMP	TX	MARTIN BRAUN	ROBERTA BRAUN	2112	F ST	NW	$ 5,935.90
0081 2214	&IMP	TX	MARTIN BRAUN		2112	F ST	NW	$ 9,460.69
0081 2232	&IMP	TX	MARTIN BRAUN	ROBERTA BRAUN	2112	F ST	NW	$ 9,909.18
0092 0805		TX	LOTTIE JOHNSON		0	21ST ST	NW	$ 218.65
0096 2007	&IMP	TX	PETER WILSON		2007	O ST	NW	$ 5,784.72
0096 2024	&IMP	TX	PETER BOLTON		2007	O ST	NW	$ 2,946.89
0097 2362	&IMP	TX	HARRY C BLUMENTHAL TRUST	MARION B SNYDER FAMILY T	1316	NEW HAMPSHIRE	NW	$ 2,993.70
0110 2156	&IMP	TX	GEOFFREY MACKLER	JESSICA MACKLER	1920	S ST	NW	$ 2,991.07
0126 0059	&IMP	TX	EPIC 919 LLC		919	18TH ST	NW	$ 3,964.87
0131 0033	&IMP	TX	REED LAURANA C		1828	FLORIDA AVE	NW	$ 4,050.84
0131 2126	&IMP	TX	THOMAS GOLDEN		1918	18TH ST	NW	$ 3,095.23
0133 0179	&IMP	TX	MICHAEL SMITH	JOSEPH FACTORA	1808	RIGGS PL	NW	$ 10,415.04
0150 0156	&IMP	TX	TADAHIKO NAKAMURA		2006	17TH ST	NW	$ 19,369.62
0151 0804		TX	1711 T ST NW LLC		0	T ST	NW	$ 5,732.71
0153 2020	&IMP	TX	MARK ALLEN		1747	18TH ST	NW	$ 7,446.92
0154 0034	&IMP	TX	F5 HOLDINGS LTD LIABILIT		1734	17TH ST	NW	$ 10,673.28
0154 2027	&IMP	TX	RAFIE ANSARI		1725	NEW HAMPSHIRE	NW	$ 2,661.47
0154 2124	&IMP	TX	1700 17TH STREET ASSOCIA		1700	17TH ST	NW	$ 4,483.38

Python Code (20190819_TaxSaleListing.py)

```
# Read Sac County Tax Sale PDFs and extract Suf-Lot and Taxes Owed
# Accumatch
# by Michael Keller
# Aug 19 - 20, 2019
# requires pdfplumber, re, datetime, os

# Libraries
import pdfplumber
import re
from datetime import datetime
import os
import sys

# Start Timer
start = datetime.now()

# Booleans
SW_Found = False
Lot_Found = False
Parcel_Found = False
TaxesOwed_Found = False

# Variables
lot = ''
taxesOwed = ''
fileCount = 0
space = ' '
nospace = ''
comma = ','
dollarSign = '$'
_return = '\r'
outString = []
outFilename = 'SAC_County_'
outputType = '.csv'
colHeadings = 'Suf-Lot,Taxes Owed'
```

```
# Regular Expressions
regLot = '\d{4}[A-Z]?.{2,5}\d{4}'
regParcel = 'PAR \d{8}'
regOwed = "['$'][ ,0-9]+['.']\d{2}"
regSW = '(SW)\d{10}'
spaces = '\s{2,}'

# Messages
messageSaveBeg = 'Saving \''
messageSaveEnd = '\' file . . .'
messageEOL = '. . . end of line'

# Functions
def getCurrentDirectory():
    os.chdir(os.path.dirname(__file__))
    return os.getcwd()

def getFiles(ext):
    for(dirpath,dirnames,filenames) in os.walk(getCurrentDirectory()):
        return (f for f in filenames if f.endswith(ext))

def printPDFfilename(fn):
    return 'Reading ' + fn + ' file . . .'

def openFile(fn):
    # Create output file
    try:
        file = open(fn,'w+')
        # Write out lines from array
        file.write(colHeadings + _return)
        file.close()
        print('Opening ' + fn + ' file . . .')
    except FileNotFoundError:
        sys.exit('File: ' + fn + ' does not exist')
    except PermissionError:
        sys.exit('Unable able to open output file for writing. Is the file currently open?')

def writeFile(fn,os):
    # Create and append to output file
    try:
        file = open(fn,'a+')
        for taxLine in os:
            file.write(taxLine + _return)
        file.close()
        print('Saving ' + fn + ' file . . .')
    except FileNotFoundError:
        sys.exit('File: ' + fn + ' does not exist')
    except PermissionError:
        sys.exit('Unable able to open output file for writing. Is the file currently open?')

PDFs = getFiles('.pdf')

for x in PDFs:
    ## Message Run
    print(printPDFfilename(x))

    ## Open PDF file for reading
    pdf = pdfplumber.open(x)

    lastPage = len(pdf.pages)

    ## Reset counters
    count = 0
    fileCount = fileCount + 1

    print('Number of Pages: ' + str(lastPage))

    fileName = outFilename + str(fileCount) + outputType
    openFile(fileName)

    for page in pdf.pages:
        p1 = pdf.pages[count]

        count = count + 1

        p1text = p1.extract_text()

        text = p1text.splitlines()


##    Break for Testing

##        if(count == 6):

##            break


        for item in text:

            TaxesOwed_Found = re.search(regOwed,item)

            if not (TaxesOwed_Found is None):

                taxesOwed = item[TaxesOwed_Found.start():TaxesOwed_Found.end()]

                if(item[0:2] == 'SW'):

                    SW_Found = re.search(regSW,item)

                    if not (SW_Found is None):

                        lot = item[SW_Found.start():SW_Found.end()]

                else:

                    Lot_Found = re.search(regLot,item)

                    Parcel_Found = re.search(regParcel,item)

                if not (Lot_Found is None):

                    lot = item[Lot_Found.start():Lot_Found.end()]

                if not (Parcel_Found is None):

                    lot = item[Parcel_Found.start():Parcel_Found.end()]


                carp1 = re.sub(spaces,space,lot)

                carp2 = taxesOwed.replace(dollarSign,nospace)

                carp2 = carp2.strip()

                carp2 = carp2.replace(comma,nospace)

                carp2 = carp2.replace(space,nospace)

                outString.append(carp1 + comma + carp2)

##                print(carp1 + comma + carp2)


    ## Close PDF

    pdf.close()
    ## Write CSV file
    writeFile(fileName,outString)

    ## Clear variables
    ## Make sure outString is clear for next file's data
    outString.clear()

#end timer
print('Runtime: ' + str(datetime.now() - start))

print(messageEOL)

# EOF
```

CSV Output



	A	B
1	Suf-Lot	Taxes Owed
2	0004N 2016	9231.48
3	0014 2232	7316.17
4	0014 2233	6964.18
5	0015 2189	3366.97
6	0016 2129	3114.9
7	0028 2075	3579.35
8	0028 2269	17778.82
9	0030 2057	3335.3
10	0036 2146	12409.1
11	0036 2147	10176.96
12	0036 2174	2660.41
13	0038 2032	4628.34
14	0038 2113	3179
15	0051 2158	2962.81
16	0066 0805	27056.71
17	0067 0808	2644.66
18	0069 2005	4183.17
19	0069 2157	12273.83
20	0070 2350	5695.74
21	0081 2213	5935.9
22	0081 2214	9460.69
23	0081 2232	9909.18
24	0092 0805	218.65
25	0096 2007	5784.72
26	0096 2024	2946.89
27	0097 2362	2993.7
28	0110 2156	2991.07

Michael Keller
Accumatch
20190819_TaxSaleListing
Aug 20, 2019