Wasatch County – Delinquent Tax Billing Report

Notes: This python scripts scrapes a Wasatch County PDF and creates a csv output file. The following column headers are created, and the associated data is collected: **Parcel, Year, Fees, Principal, Penalty, Interest, Total Due**. The PDF filename is not unique. What was challenging about scraping this particular PDF was that the Parcel Number comes after most of the tax data.

PDF

08/12/2019 **WASATCH COUNTY** 04:11:25PM

Delinquent Tax Billing Report

			Year	Fees	Principal	Penalty	Interest	Total Due
1 KL LLC			2018		1,678.70	41.97	89.26	1,809.93
3415 E SUNDOWNER RIDGE DR			2017		1,645.52	41.14	197.35	1,884.01
HEBER			2016		1,621.93	40.55	304.19	1,966.67
LOT TP1, TUHAYE TWIN PEAKS SUBDI			2015		1,079.30	26.98	279.86	1,386.14
00-0020-9955	(Active)	**Parcel Totals:**		**0.00**	**6,025.45**	**150.64**	**870.66**	**7,046.75**
1130 LONGVIEW LLC			2018		1,507.85	37.70	80.17	1,625.72
1130 E LONGVIEW DR			2017		1,460.38	36.51	175.15	1,672.04
00-0020-3998	(Active)	**Parcel Totals:**		**0.00**	**2,968.23**	**74.21**	**255.32**	**3,297.76**
2491 DANIELS ROAD LLC			2018		5,226.18	130.65	277.88	5,634.71
2491 S DANIELS RD			2017		4,948.11	123.70	593.44	5,665.25
00-0020-5970	(Active)	**Parcel Totals:**		**0.00**	**10,174.29**	**254.35**	**871.32**	**11,299.96**
7MW HOLDINGS LLC			2018		10,932.45	273.31	581.28	11,787.04
00-0015-5346	(Active)	**Parcel Totals:**		**0.00**	**10,932.45**	**273.31**	**581.28**	**11,787.04**
7MW HOLDINGS LLC			2018		3,401.54	85.04	180.86	3,667.44
00-0015-5353	(Active)	**Parcel Totals:**		**0.00**	**3,401.54**	**85.04**	**180.86**	**3,667.44**
841 PROPERTIES LLC			2018		5,383.21	0.00	13.90	5,397.11
847 S MAIN ST			2017		6,009.15	150.23	720.69	6,880.07
00-0005-9118	(Active)	**Parcel Totals:**		**0.00**	**11,392.36**	**150.23**	**734.59**	**12,277.18**
ACTIUM HIGH YIELD LOAN FUND III LLC			2017		714.48	17.86	85.69	818.03
00-0021-2027	(Active)	**Parcel Totals:**		**0.00**	**714.48**	**17.86**	**85.69**	**818.03**
ACTIUM HIGH YIELD LOAN FUND III LLC			2017		6,785.34	169.63	813.78	7,768.75
00-0021-2028	(Active)	**Parcel Totals:**		**0.00**	**6,785.34**	**169.63**	**813.78**	**7,768.75**
ADAMSON GREG			2018		24.78	37.18	3.21	65.17
00-0020-7305	(Active)	**Parcel Totals:**		**0.00**	**24.78**	**37.18**	**3.21**	**65.17**
AGARWAL SWAPNIL			2018		10.00	513.90	27.18	551.08
00-0020-3725	(Active)	**Parcel Totals:**		**0.00**	**10.00**	**513.90**	**27.18**	**551.08**
AGUIRRE JAIME			2018		929.33	23.23	49.41	1,001.97

Python Code (20190816_WasatchCounty.py)

```
# Read a Wasatch County Delinquent Tax Billing Report and extracting
# Parcel ID, Year, Principal, Penalty, Interest, Total Due and Totals
# Accumatch
# by Michael Keller
# Aug 15, 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
Parcel_Found = False
Year_Found = False
lineStop = False
lineStart = False
dumpData = False
goodYear = False

# Variables
fn = 'delq-listing.pdf'
messageEOL = '. . . end of line'
dollarSign = '$'
count = 0
nospace = ''
space = ' '
comma = ','
outString = []
outFilename = 'WasatchCounty_'
```

```
outputType = '.csv'
dataLineStart = 'Year Fees Principal Penalty Interest Total Due'
dataLineStop = 'Page'
colHeadings = 'Parcel,Year,Fees,Principal,Penalty,Interest,Total Due'
_return = '\r'
taxdatacount = 0
totalDue = ''
interest = ''
penalty = ''
principal = ''
fees = ''
year = ''
# capture last five years (plus one)
yearsToCollect = 6

# Regular Expressions
regParcel = '\d{2}[-]\d{4}[-]\d{4}'
regYear = '201\d{1}'

# 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 printNumberOfPages(np):
    return 'Number of Pages: ' + str(np)

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 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 to open output file for writing. Is the file currently open?')

def clearArray(a):
    # Clear and recreate array
    try:
        a.clear()
        rows = yearsToCollect
        cols = 6
        a = [['' for j in range(rows)] for i in range(cols)]
    except PermissionError:
        sys.exit('Unable to create two dimensional array. Sorry.')
    return a

# PDF File Open
```

```
print(printPDFfilename(fn))
pdf = pdfplumber.open(fn)

lastPage = len(pdf.pages)
print(printNumberOfPages(lastPage))

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

a = clearArray([])

for page in pdf.pages:
    p1 = pdf.pages[count]
    pltext = p1.extract_text()
    text = pltext.splitlines()

##    Break for Testing
##    if(count == 9):
##        break

    for item in text:
        if(item[0:4] == dataLineStop):
            lineStop = True
            lineStart = False

        if((lineStart == True) and (lineStop == False)):
            blob = item.rsplit(space,4)
            totalDue = blob[4].replace(comma,nospace)
            totalDue = totalDue.replace(dollarSign,nospace)
            interest = blob[3].replace(comma,nospace)
            interest = interest.replace(dollarSign,nospace)
            penalty = blob[2].replace(comma,nospace)
            penalty = penalty.replace(dollarSign,nospace)
            principal = blob[1].replace(comma,nospace)
            principal = principal.replace(dollarSign,nospace)
            restOfLine = blob[0]
```

```
            Parcel_Found = re.search(regParcel,restOfLine)
            if not (Parcel_Found is None):
                Parcel = restOfLine[Parcel_Found.start():Parcel_Found.end()]
                restOfLine = restOfLine[Parcel_Found.end():len(restOfLine)]
                blob2 = restOfLine.rsplit(space,1)
                fees = blob2[1].replace(comma,nospace)
                fees = fees.replace(dollarSign,nospace)
                dumpData = True
                a[taxdatacount][0] = ''
                a[taxdatacount][1] = fees
            else:
                Year_Found = re.search(regYear,restOfLine)
                if not (Year_Found is None):
                    year = item[Year_Found.start():Year_Found.end()]

                if(int(year) >= 2015):
                    a[taxdatacount][0] = year
                    a[taxdatacount][1] = ''
                    goodYear = True

            if((dumpData == True) or (goodYear == True)):
                goodYear = False
                a[taxdatacount][5] = totalDue
                a[taxdatacount][4] = interest
                a[taxdatacount][3] = penalty
                a[taxdatacount][2] = principal
                taxdatacount = taxdatacount + 1

            if(dumpData == True):
                cols = len(a)
                rows = 0
                if cols:
                    rows = len(a[0])
                for j in range(taxdatacount):
                    outString.append(Parcel + comma + a[j][0] + comma + a[j][1] + comma + a[j][2]
+ comma + a[j][3] + comma + a[j][4] + comma + a[j][5])
```

```
                a = clearArray(a)
                dumpData = False
                taxdatacount = 0

        if(item.strip() == dataLineStart):
            lineStop = False
            lineStart = True

    count = count + 1

## Close PDF
pdf.close()

## Write CSV file
writeFile(fileName,outString)

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

print(messageEOL)

# EOF
```

CSV Output

	A	B	C	D	E	F	G
1	Parcel	Year	Fees	Principal	Penalty	Interest	Total Due
2	00-0020-9955	2018		1678.7	41.97	89.26	1809.93
3	00-0020-9955	2017		1645.52	41.14	197.35	1884.01
4	00-0020-9955	2016		1621.93	40.55	304.19	1966.67
5	00-0020-9955	2015		1079.3	26.98	279.86	1386.14
6	00-0020-9955		0	6025.45	150.64	870.66	7046.75
7	00-0020-3998	2018		1507.85	37.7	80.17	1625.72
8	00-0020-3998	2017		1460.38	36.51	175.15	1672.04
9	00-0020-3998		0	2968.23	74.21	255.32	3297.76
10	00-0020-5970	2018		5226.18	130.65	277.88	5634.71
11	00-0020-5970	2017		4948.11	123.7	593.44	5665.25
12	00-0020-5970		0	10174.29	254.35	871.32	11299.96
13	00-0015-5346	2018		10932.45	273.31	581.28	11787.04
14	00-0015-5346		0	10932.45	273.31	581.28	11787.04
15	00-0015-5353	2018		3401.54	85.04	180.86	3667.44
16	00-0015-5353		0	3401.54	85.04	180.86	3667.44
17	00-0005-9118	2018		5383.21	0	13.9	5397.11
18	00-0005-9118	2017		6009.15	150.23	720.69	6880.07
19	00-0005-9118		0	11392.36	150.23	734.59	12277.18
20	00-0021-2027	2017		714.48	17.86	85.69	818.03
21	00-0021-2027		0	714.48	17.86	85.69	818.03
22	00-0021-2028	2017		6785.34	169.63	813.78	7768.75
23	00-0021-2028		0	6785.34	169.63	813.78	7768.75
24	00-0020-7305	2018		24.78	37.18	3.21	65.17
25	00-0020-7305		0	24.78	37.18	3.21	65.17
26	00-0020-3725	2018		10	513.9	27.18	551.08
27	00-0020-3725		0	10	513.9	27.18	551.08
28	00-0020-6989	2018		929.33	23.23	49.41	1001.97
29	00-0020-6989		0	929.33	23.23	49.41	1001.97

Michael Keller
Accumatch
20190815_WasatchCounty
Aug 16, 2019