

Township / Property List

Notes: This python scripts scrapes PDFs from a Township/Property List from this PDF, and other in the directory, and creates a csv output file for each PDF. The following column headers are created and the associated data is collected: **Parcel and Amount**. The PDF filenames are unique and each CSV file is named based on the township located on the second line.

PDF

township|property_list

BELLEVILLE Parcel Count: 1,579

FRANKLIN, JAMES L & LINDA	02-25.0-309-024	\$2,130.16
JOHNSON, IDA ET AL	02-25.0-316-001	\$264.04
JOHNSON, IDA ET AL	02-25.0-316-002	\$1,048.61
GROOMS, DOROTHY	02-25.0-316-005	\$338.00
SHANKLIN, ETHEL	02-25.0-316-009	\$1,626.81
SMITH, ROGER	02-25.0-317-001	\$1,785.46
SPENCER, MILDRED	02-25.0-317-002	\$2,262.35
SMITH, ROGER	02-25.0-317-005	\$1,228.58
CHOLET, CRAIG	02-25.0-317-006	\$1,633.41
HUSTON, LESA	02-25.0-319-004	\$1,792.65
MONTGOMERY, MARY	02-25.0-319-010	\$4,977.25
DONLOE, LEON B & CYNTHIA R	02-26.0-417-015	\$1,214.69
DIGGS YEKITA & , RIGSBY CA	02-26.0-417-016	\$2,428.58
YOUNG, LAVONDA	02-26.0-417-017	\$1,771.75
YATES, CLINTON	02-26.0-417-018	\$246.10
MURRAY, CORNELIUS	02-26.0-417-033	\$264.36
LEMONS, MACK R REV	02-26.0-417-054	\$1,834.06
SPEARMAN, SHIRLEY	02-26.0-417-063	\$134.53
DORSEY, BOBBIE	02-26.0-417-064	\$864.11
BENTON, EUGENE JR & LUCIND	02-26.0-417-066	\$1,305.15
BENTON, CASSANDRA	02-26.0-417-069	\$1,560.27
SKY BLUE DEVELOPMENT INC	02-26.0-417-096	\$7,656.29
MOORE, FRED A & JANIE M	02-26.0-417-104	\$1,616.58
BERNARD, O BASGA & BARBARA	02-26.0-418-004	\$1,219.71
TURMAN, DAVID	02-35.0-203-007	\$5,376.18
TURMAN, DAVID	02-35.0-203-028	\$64.25
DESTINY C OHARA-WALKER	02-35.0-206-010	\$464.53
CAVE CREEK PROPERTIES	02-35.0-207-007	\$84.94
CAVE CREEK PROPERTIES	02-35.0-207-008	\$1,060.76
MIDDENDORF, DIANA	02-35.0-207-026	\$1,203.50
VAUGHN, KRISTOPHER	02-35.0-207-044	\$151.99
SIMONSON, EDWARD & MARGARE	02-35.0-207-056	\$1,415.71
MADLOCK, VIOLA	02-35.0-208-041	\$92.27
TJN LAND TRUST NO 3080 C/O	02-35.0-208-054	\$7,090.92

Python Code (20190806_StClair.py)

```
# Read a list of PDF files in folder, open, scan, and dump data individual text files
# Accumatch
# by Michael Keller
# Aug 7, 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
Township_Found = False
_TownshipFound = False
Parcel_Found = False
Amount_Found = False

# Variables
Township = ''
TownshipFlag = 'Parcel Count'
FileFlag = 'file:'
space = ' '
nospace = ''
comma = ','
_return = '\r'
count = 0
outString = []
outputType = '.csv'
fileName = ''
regParcel = '\\d{2}[-]\\d{2}[.]\\d{1}[-]\\d{3}[-]\\d{3}'
```

```

regAmount = "['\$'] [0-9',']*['.']\d{2}"
parcel = ''
amount = ''
dollarSign = '$'
colHeadings = 'Parcel,Amount'

# 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)
    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

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

    for page in pdf.pages:
        p1 = pdf.pages[count]
        count = count + 1
        pltext = p1.extract_text()
        text = pltext.splitlines()

    ## Break for Testing
    ## if(count == 6):
    ## break
```

```

    for item in text:
        if((TownshipFlag in item) and (_TownshipFound == False)):
            carp = item.split(space,1)
            Township = carp[0].strip()
            print('Township Name: ' + Township)
            _TownshipFound = True
            fileName = Township + outputType
            openFile(fileName)
        else:
            if(item[0:5] != FileFlag):
                Parcel_Found = re.search(regParcel,item)
                if not (Parcel_Found is None):
                    Parcel = item[Parcel_Found.start():Parcel_Found.end()]
                    carp2 = item[Parcel_Found.end():len(item)].strip()
                    Amount_Found = re.search(regAmount,carp2)
                    if not (Amount_Found is None):
                        Amount = carp2[Amount_Found.start():Amount_Found.end()]
                        Amount = Amount.replace(comma,nospace)
                        Amount = Amount.replace(dollarSign,nospace)
                        outString.append(Parcel + comma + Amount)

## Close PDF
pdf.close()

## Write CSV file
writeFile(fileName,outString)

## Clear variables
pl = ''
pltext = ''
text = ''
_TownshipFound = False
## 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	Parcel	Amount
2	02-25.0-309-024	\$2,130.16
3	02-25.0-316-001	\$264.04
4	02-25.0-316-002	\$1,048.61
5	02-25.0-316-005	\$338.00
6	02-25.0-316-009	\$1,626.81
7	02-25.0-317-001	\$1,785.46
8	02-25.0-317-002	\$2,262.35
9	02-25.0-317-005	\$1,228.58
10	02-25.0-317-006	\$1,633.41
11	02-25.0-319-004	\$1,792.65
12	02-25.0-319-010	\$4,977.25
13	02-26.0-417-015	\$1,214.69
14	02-26.0-417-016	\$2,428.58
15	02-26.0-417-017	\$1,771.75
16	02-26.0-417-018	\$246.10
17	02-26.0-417-033	\$264.36
18	02-26.0-417-054	\$1,834.06
19	02-26.0-417-063	\$134.53
20	02-26.0-417-064	\$864.11
21	02-26.0-417-066	\$1,305.15
22	02-26.0-417-069	\$1,560.27
23	02-26.0-417-096	\$7,656.29
24	02-26.0-417-104	\$1,616.58
25	02-26.0-418-004	\$1,219.71
26	02-35.0-203-007	\$5,376.18
27	02-35.0-203-028	\$64.25
28	02-35.0-206-010	\$464.53

Michael Keller
Accumatch
20190806_StClair
Aug 7, 2019