In [122]:

```
# -*- coding: utf-8 -*-
# python 3.6
from __future__ import print_function

import sys
import re
import os
from bs4 import BeautifulSoup
import urllib.request
import pandas as pd
import pprint
import datetime
# importing necessary packages

example_table_link = "https://jobs.mo.gov/content/missouri-warn-notices-py-2017"


def remove_control_chart(s):
    """
    :param s: string that may not be utf-8 encode
    :return:
    """
    s = s.replace('\xa0', '')
    s = re.sub(r'\s\s', ' ', s)
    s = '"' + s + '"'
    return s


def to_string(s):
    """
    makes a string
    input example: 10000
    output example: '1000'
    """
    try:
        return str(s)
    except:
        # Change the encoding type if needed
        return s.encode('utf-8')


def clean_up_cols(s):
    """
    :param s: string for example # AFFECTED
    :return: space removed for example #_affected
    """
    s = to_string(s)
    s = s.lower()
    s = s.replace(' ', '_')
    s = s.replace('#', 'number')
```

```
    s = '"' + s + '"'

    return s


def WEBSCRAPeTABLES(link, outputfile_name):
    """
    :param link: this will be the link where you want to scrape table
    :return: json and csv that returns metadata or table data with header keys
    """
    keys_ls = []
    with urllib.request.urlopen(link) as response:
        fout = open(outputfile_name,'w')
        the_page = response.read()
        bs = BeautifulSoup(the_page,features="lxml")
        table = bs.find(lambda tag: tag.name == 'table')
        tabletr = table.tr
        print("RAW XML HEADERS:")
        pprint.pprint(tabletr)
        list_columns = tabletr.find_all('th')
        print("\nCLEANED HEADERS:")
        # this provides the list of columns using the clean_up_cols fxn defined
above
        colnames = [clean_up_cols(lc.text) for lc in list_columns]
        header = ','.join(colnames)
        print(header)
        fout.write(header)
        rows = table.findAll(lambda tag: tag.name == 'tr')
        data = []
        counter = 0
        print("\nROW CONTENTS:")
        for row_content in rows:
            counter += 1
            values = row_content.find_all('td')
            csv_row = ','.join([(remove_control_chart(to_string(val.text))) for
val in values])
            print(csv_row)
            fout.write(csv_row +
                       '\n')
    print("\n\n*******\n\nwrote out", to_string(counter), "rows from table\n in
csv file called " + outputfile_name)

name = to_string(example_table_link.split('/')[-1]) + ".csv"
WEBSCRAPeTABLES(example_table_link, name)
```

```
RAW XML HEADERS:
<tr><th scope="col"><span>DATE RECEIVED</span></th>
<th scope="col"><span>COMPANY NAME</span></th>
<th scope="col"><span>LOCATION</span></th>
<th scope="col"><span>COUNTY</span></th>
<th scope="col"><span>REGION</span></th>
<th scope="col"><span>TYPE</span></th>
<th scope="col"><span>LAYOFF DATE</span></th>
<th scope="col"><span># AFFECTED</span></th>
```

</tr>

CLEANED HEADERS:
"date_received","company_name","location","county","region","type","
layoff_date","number_affected"

ROW CONTENTS:

"07/06/17","SunEdison, Inc.","St. Louis","St. Louis County","St. Lou
is County","Layoff","07/10/2017","6"
"07/24/17","Neuterra d/b/a Fulton Medical Center ("Hospital")","Fult
on","Callaway","Central","Closing","09/22/2017","158"
"07/31/17","Lozier Corporation","Union","Franklin","Jefferson/Frankl
in Consortium","Closing","10/06/2017 - Fall 2017","92"
"08/03/17","Diodes Fabtech, Inc.","Lee's Summit","Jackson","Kansas C
ity & Vicinity","Closing","10/06/2017","167"
"08/16/17","Frontier Communications Corporation","Weldon Spring","St
. Charles","St. Charles County","Closing","10/16/2017 - 10/20/2017",
"141"
"09/20/17","Positronic Industries Inc.","Mt. Vernon","Lawrence","Sou
thwest","Layoff","12/01/2017","97"
"10/10/17","TD AmeritradeCorp. (updated 08-22-2018)","Multiple Locat
ions","Multiple Locations","Multiple Locations","Layoff","11/21/2017
","1208"
"10/20/17","Zhongding Sealing Parts (USA), Inc. d/b/a Buckhorn Rubbe
r Products","Hannibal","Ralls","Northeast","Closing","12/22/2017","1
19"
"10/27/17","Amcor Rigid Plastics USA LLC(updated 12-27-2017)","Jeffe
rson City","Cole","Central","Layoff","12/29/2017-3/29/2018","72"
"11/02/17","Kmart","Independence","Jackson","Kansas City & Vicinity"
,"Closing","01/21/2018","62"
"11/07/17","Kindred Hospitals East, LLC d/b/a Kindred Hospital","Kan
sas City","Jackson","Kansas City & Vicinity","Closing","01/06/2018-0
1/20/2018","115"
"11/16/17","HM Dunn AeroSystems, Incorporated","St. Louis","St. Loui
s City","St. Louis City","Closing","01/28/18","53"
"11/20/17","Sedgwick LLP","Kansas City","Jackson","Kansas City & Vic
inity","Closing","01/20/2018","75"
"12/05/17","Knappco, Inc.","Riverside","Platte","Kansas City & Vicin
ity","Closing","02/12/2018","58"
"12/14/17","AT&T Communications, Inc.","Kansas City","Jackson","Kans
as City & Vicinity","Layoff","02/17/2018","87"
"12/15/17","Talbot (Leggett & Platt, Inc.)","Neosho","Newton","South
west","Closing","02/13/2018","130"
"12/15/17","Armstrong Energy, Inc.","St. Louis","St. Louis County","
St. Louis County","Closing","02/14/2018","8"
"01/05/18","ConAgra Foods, Inc.(Updated 01-30-2018)","Trenton","Grun
dy","Northwest","Closing","03/09/2018-05/31/2018","282"
"01/11/18","Moon Ridge Foods, LLC","Pleasant Hope","Polk","Southwest
","Layoff","01/11/2018","240"
"01/26/18","Serco, Inc.","Wentzville","St. Charles County","St. Char
les County","Closing","06/30/2018","660"
"02/02/18","ASM Research","St. Louis","St. Louis City","St. Louis Ci

ty","Closing","03/27/2018","14"
"02/02/18","Cognosante, LLC(Updated 03-26-2018)","Wentzville","St. C
harles County","St. Charles County","Closing","04/06/2018","84"
"02/20/18","ConvergysCorporation","Arnold","Jefferson/Franklin Conso
rtium","Jefferson County","Closing","04/30/2018","319"
"03/15/18","Toys "R" Us(Updated 05-15-2018)","Lee's Summit","Jackson
County","Kansas City & Vicinity","Closing","05/14/2018","191"
"03/23/18","Eagle Foods","Seneca","Newton","Southwest","Closing","06
/06/2018","39"
"04/18/18","ABB Inc.(Updated 08-08-2018)","St. Louis","St. Louis Cit
y","St. Louis City","Layoff","May 31, 2018","82"
"04/19/18","Select Medical Corporation","Kansas City","Jackson","Kan
sas City & Vicinity","Closing","May 18, 2018","105"
"04/27/18","Claycomo Releasing, Inc.(Updated 05-10-2018)","Claycomo"
,"Clay","Kansas City & Vicinity","Closing","June 30, 2018","172"
"04/30/18","Twin Rivers Regional Medical Center - Kennett HMA Physic
ians Management, LLC","Kennett","Dunklin","Southeast","Closing","Jun
e 30, 2018","16"
"04/30/18","Twin Rivers Regional Medical Center - KennettHMA, LLC","
Kennett","Dunklin","Southeast","Closing","June 30, 2018","259"
"05/10/18","CassensTransport Company","Claycomo","Clay","Kansas City
& Vicinity","Closing","June 30, 2018","29"
"05/18/18","Car City Motor, Inc.","St. Joseph","Buchanan","Northwest
","Closing","June 18, 2018","74"
"05/30/18","AramarkCampus Services, LLC - SaintLouis University","St
. Louis","St. Louis","St. Louis","Closing","July 31, 2018","188"
"05/29/18","Harley Davidson Motor Company, Inc. - Kansas City","Kans
as City","Jackson","Kansas City & Vicinity","Closing","August 3, 201
8","180"
"06/15/18","Hard Rock Cafe - St. Louis","St. Louis","St. Louis City"
,"St. Louis City","Closing","August 16, 2018","57"
"06/15/18","syncreon U.S.","Kansas City","Kansas City","Kansas City
& Vicinity","Layoff","August 17, 2018","207"
"06/28/18","State Farm Mutual Automobile Insurance Company - Earth C
ity","Earth City","St. Louis County","St. Louis County","Closing","A
ugust 31, 2018","136"
"06/29/18","American Airlines, Inc. - St. LouisPilot Crew Base","St.
Louis","St. Louis","St. Louis","Closing","September 3, 2018","155"
"","","","","","","TOTAL","6137"

****

wrote out 40 rows from table
 in csv file called missouri-warn-notices-py-2017.csv

In [123]:

```
# now checking to see how csv filecomesin

missouri=pd.read_csv("missouri-warn-notices-py-2017.csv")
missouri.head()
```

Out[123]:

	date_received	company_name	location	county	region	type	layoff_date	nu
0	07/06/17	SunEdison, Inc.	St. Louis	St. Louis County	St. Louis County	Layoff	07/10/2017	
1	07/24/17	Neuterra d/b/a Fulton Medical Center (Hospital")"	Fulton	Callaway	Central	Closing	09/22/2017	
2	07/31/17	Lozier Corporation	Union	Franklin	Jefferson/Franklin Consortium	Closing	10/06/2017 - Fall 2017	
3	08/03/17	Diodes Fabtech, Inc.	Lee's Summit	Jackson	Kansas City & Vicinity	Closing	10/06/2017	
4	08/16/17	Frontier Communications Corporation	Weldon Spring	St. Charles	St. Charles County	Closing	10/16/2017 - 10/20/2017	

In [125]:

```
missouri.tail()
# here if you would like you can drop the last column -- wanted to keep to see h
ow these would parse
```

Out[125]:

	date_received	company_name	location	county	region	type	layoff_date	number_affe
34	06/15/18	Hard Rock Cafe - St. Louis	St. Louis	St. Louis City	St. Louis City	Closing	August 16, 2018	
35	06/15/18	syncreon U.S.	Kansas City	Kansas City	Kansas City & Vicinity	Layoff	August 17, 2018	
36	06/28/18	State Farm Mutual Automobile Insurance Company...	Earth City	St. Louis County	St. Louis County	Closing	August 31, 2018	
37	06/29/18	American Airlines, Inc. - St. LouisPilot Crew ...	St. Louis	St. Louis	St. Louis	Closing	September 3, 2018	
38	NaN	NaN	NaN	NaN	NaN	NaN	TOTAL	

In [129]:

```
two_table_link = "http://genelex.com/clinical-guidance/cardiology"

WEBSCRAPeTABLES(two_table_link, outputfile_name="drugCardiologyCyp.csv")
```

```
RAW XML HEADERS:
<tr><th>Drug</th><th>Biomarker</th><th>Drug</th><th>Biomarker</th></
tr>

CLEANED HEADERS:
"drug","biomarker","drug","biomarker"

ROW CONTENTS:

"carvedilol","CYP2D6","propafenone","CYP2D6"
"clopidogrel","CYP2C19","propranolol","CYP2D6"
"isosorbide and hydralazine","NAT1;NAT2","quinidine / dextromethorph
an","CYP2D6"
"metoprolol","CYP2D6","ticagrelor","CYP2C19"
"prasugrel","CYP2C19","warfarin","CYP2C9 and VKORC1"


*******

wrote out 6 rows from table
 in csv file called drugCardiologyCyp.csv
```

In [130]:

```
# issue is
drugCYP = pd.read_csv("drugCardiologyCyp.csv")
drugCYP
```

Out[130]:

	drug	biomarker	drug.1	biomarker.1
0	carvedilol	CYP2D6	propafenone	CYP2D6
1	clopidogrel	CYP2C19	propranolol	CYP2D6
2	isosorbide and hydralazine	NAT1;NAT2	quinidine / dextromethorphan	CYP2D6
3	metoprolol	CYP2D6	ticagrelor	CYP2C19
4	prasugrel	CYP2C19	warfarin	CYP2C9 and VKORC1

In [ ]:

```
# unfortunately this code is too basic and needs a little work to grab more tabl
es from a webpage
```