

# -\*- coding: utf-8 -\*-  
# encoding=utf-8  
import urllib  
import re  
import HTMLParser  
import sys  
from splinter import Browser  
from bs4 import BeautifulSoup  
import time  
import calendar  
import os  
import ConfigParser  
import urllib2  
import requests  
import json  
  
fileText = ""  
name = 1  
stop = False  
recordsFiltered = 0  
  
  
def getHtml(url, c, timeInterval):  
 # 请求url  
 # response = requests.get("https://ygjw.hebga.gov.cn/web/infoqueryxzcf/xzcf/query?draw=2&start=0&length=100&sel\_org\_parent\_id=13----&sel\_city\_org\_id=130533&sel\_qry\_start\_date=2018-01-01+%7C+2019-08-31&keywords=&\_=1566279005")  
 response = requests.get(url)  
 # 获取响应内容  
 your\_dict = json.loads(response.content)  
 list1 = your\_dict.get('data')  
 for i in list1:  
 docId = i.get("docDataId")  
 # 取得文号  
 docDataNumber = i.get("docDataNumber")  
 if docDataNumber is None:  
 docDataNumber = i.get("docDataTitle")  
 docDataTitle = i.get("docDataTitle")  
  
 lastUpdateTime = i.get("lastUpdateTime")[0:10]  
 isUpdatetimeAfter = isTime(lastUpdateTime)  
  
 # 每隔10秒再去执行爬取数据操作  
 time\_remaining = timeInterval - time.time() % timeInterval  
 time.sleep(time\_remaining)  
  
 if isUpdatetimeAfter:  
 urlWs = "https://ygjw.hebga.gov.cn/web/infoqueryxzcf/xzcf/detail/{}/1".format(docId)  
 responseUrlWs = requests.get(urlWs)  
 # urlWs = "https://ygjw.hebga.gov.cn/web/infoqueryxzcf/xzcf/detail/5C592CF3F314435487A8ABF7B2304C4C/1"  
 # browser.visit(urlWs)  
 html = responseUrlWs.content  
 soup = BeautifulSoup(html, "lxml")  
  
 # 如果能够获取p标签为案号的这一类文书，手动换行  
 if soup.find('p', {'class': "titleDW"}) is not None:  
 # 全文内容  
 contents = soup.find\_all(name='div', attrs={'class': 'lbody'})  
 # 正文内容  
 zws = soup.find\_all(name='p', attrs={'class': 'MsoNormal'})  
 wsnr = ""  
 titleWH = ""  
 titleDW = ""  
 titleName = ""  
 buttomDW = ""  
 buttonTime = ""  
 for m in contents:  
 if m.find('p', {'class': "titleDW"}) is not None:  
 titleDW = m.find('p', {'class': "titleDW"}).text  
 wsnr += titleDW + '\n'  
  
 if docDataTitle is not None:  
 wsnr += docDataTitle + '\n'  
 elif m.find('p', {'class': "titleName"}) is not None:  
 titleName = m.find('p', {'class': "titleName"}).text  
 wsnr += titleName + '\n'  
  
 if docDataNumber is not None:  
 wsnr += docDataNumber + '\n'  
 elif m.find('p', {'class': "titleWH"}) is not None:  
 titleWH = m.find('p', {'class': "titleWH"}).text  
 wsnr += titleWH + '\n'  
  
 nr = ""  
 for n in zws:  
 nr += n.text + '\n'  
 if nr is not None:  
 wsnr += nr  
  
 if m.find('p', {'class': "buttomDW"}) is not None:  
 buttomDW = m.find('p', {'class': "buttomDW"}).text  
 wsnr += buttomDW + '\n'  
  
 if m.find('p', {'class': "buttonTime"}) is not None:  
 buttonTime = m.find('p', {'class': "buttonTime"}).text  
 wsnr += buttonTime + '\n'  
  
 # if wsnr is not None:  
 if docDataNumber is not None:  
 docDataNumber = docDataNumber.replace("\u884c\u653f\u5904\u7f5a\u51b3\u5b9a\u4e66", "");  
 saveFile(docDataNumber + lastUpdateTime, wsnr, c)  
 else:  
 saveFile(docId + lastUpdateTime, wsnr, c)  
 # 如果不能够获取p标签为案号的这一类文书，已经换行，无需我们换行  
 else:  
 qwnr = ""  
 if soup.find('div', {'class': "lbody"}) is not None:  
 qwnr = soup.find('div', {'class': "lbody"}).text  
 if qwnr is not None:  
 if docDataNumber is not None:  
 # docDataNumber = docDataNumber.replace("行政处罚决定书", "");  
 saveFile(docDataNumber + lastUpdateTime, qwnr, c)  
 else:  
 saveFile(docId + lastUpdateTime, qwnr, c)  
 else:  
 saveFile(docId + '~' + lastUpdateTime, qwnr, c)  
 return html  
  
  
def isTime(uploadDate):  
 startTime = conf.get("startTime", "startTime")  
 endTime = conf.get("endTime", "endTime")  
 # startTime = '2017-01-01'  
 # endTime = '2018-01-01' # 邢台爬桥西、沙河，2017年全年数据  
 startStructTime = time.strptime(startTime, '%Y-%m-%d')  
 '''当前案件上传时间'''  
 structTime = time.strptime(uploadDate, '%Y-%m-%d')  
 endStructTime = time.strptime(endTime, '%Y-%m-%d')  
 print time.mktime(structTime)  
 print time.mktime(startStructTime)  
 if time.mktime(structTime) >= time.mktime(startStructTime):  
 return True  
 else:  
 return False  
  
  
def saveFile(fileName, contents, c):  
 savePath = conf.get("savePath", "savePath")  
 print savePath  
 folderPath = savePath + c  
 # folderPath = "e:\\crawlerFile\\XT\XTFileByCode20170101To20180101\\" + c  
 filePath = folderPath + "\\" + fileName + ".txt"  
 isExist = os.path.exists(folderPath)  
 if not isExist:  
 os.makedirs(folderPath)  
 print folderPath + '创建成功'  
 else:  
 print folderPath + '目录已存在'  
 fh = open(filePath, 'w')  
 fh.write(contents.encode('utf-8'))  
 fh.close()  
 print "保存成功！"  
  
  
if \_\_name\_\_ == '\_\_main\_\_':  
  
 '''读取配置文件'''  
 # 生成config对象  
 conf = ConfigParser.ConfigParser()  
 # 用config对象读取配置文件  
 conf.read("config.cfg")  
  
 corpId = conf.get("corpId", "corpId")  
 timeInterval = int(conf.get("timeInterval", "timeInterval"))  
 print "corpId:" + corpId  
  
 if "," in corpId:  
 '''英文逗号'''  
 code = corpId.split(",")  
 elif "，" in corpId:  
 '''中文逗号'''  
 code = corpId.split("，")  
 else:  
 code = []  
 code.append(corpId)  
 # code = ['130530','130531','130532','130533','130534','130535','130500','130581','130582','130591','130502','130503','130521','130522','130523','130524','130525','130526','130527','130528','130529']  
 # code = ['130528','130529']  
 # 桥西沙河，2017年全年  
 # code = ['130503','130582']  
 # code = ['130582']  
  
  
 for c in code:  
 page = 0  
 stop = False  
 # urlPre = conf.get("urlPre","urlPre")  
 startTime = conf.get("startTime", "startTime")  
 endTime = conf.get("endTime", "endTime")  
 # 0 start  
 # 1 13  
 # 2 corpid  
 # 3 startDate  
 # 4 endDate  
 # 5 时间戳  
 print "############################地区：" + c + "############################"  
  
 # 获取循环次数  
 timestamp = calendar.timegm(time.gmtime())  
 getCsUrl = "https://ygjw.hebga.gov.cn/web/infoqueryxzcf/xzcf/query?draw=2&start=0&length=100&sel\_org\_parent\_id={0}----&sel\_city\_org\_id={1}&sel\_qry\_start\_date={2}+%7C+{3}&keywords=&\_={4}"  
 getCsUrl = getCsUrl.format(c[0:2], c, startTime, endTime, timestamp)  
 print getCsUrl  
 response = requests.get(getCsUrl)  
 dict\_response = json.loads(response.content)  
 recordsFiltered = dict\_response.get('recordsFiltered')  
 length = recordsFiltered / 100 + 1  
  
 try:  
 for i in range(0, length):  
 urlPre = "https://ygjw.hebga.gov.cn/web/infoqueryxzcf/xzcf/query?draw=2&start={0}&length=100&sel\_org\_parent\_id={1}----&sel\_city\_org\_id={2}&sel\_qry\_start\_date={3}+%7C+{4}&keywords=&\_={5}"  
 url = urlPre.format(i \* 100, c[0:2], c, startTime, endTime, timestamp)  
 print url  
 print "\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*第" + str(i + 1) + "页\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*"  
 getHtml(url, c, timeInterval)  
 print "地区" + c + "文书同步成功，共计" + str(recordsFiltered) + "篇文书，同步完成"  
 except Exception, ex:  
 if (ex == 'Alert Text: None'):  
 continue  
 print Exception, ":", ex

[startTime]

startTime = 2019-07-31

[endTime]

endTime = 2019-08-01

[log]

logFileName = crawlerLog.log

[savePath]

savePath =d:\\crawlerFile\\XTFileByCode20170101To20180101\\

[corpId]

corpId =131024

[timeInterval]

timeInterval =10

document.getElementsByClassName("lbody")

document.getElementsByClassName("lbody")[0].getElementsByTagName('p')

var ppp = document.getElementsByClassName("lbody")[0].getElementsByTagName('p')

for(index in ppp){console.log(ppp[index].innerText)}

var ppp = document.getElementsByClassName("lbody")[0].children

for(index in ppp){console.log(ppp[index].innerText)}

for(index in ppp){console.log(ppp[index].innerText)}