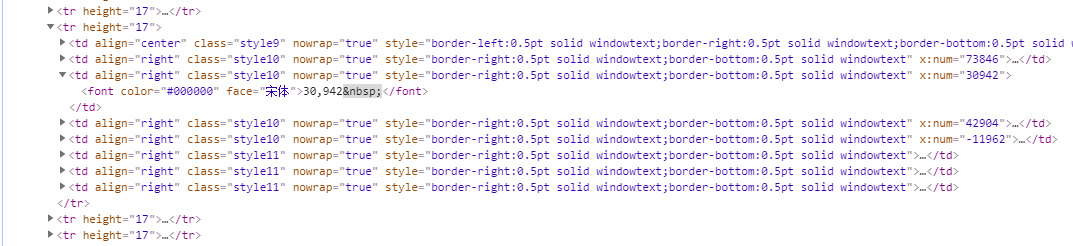
以海关总署的进出口商品国别总值表为例，说明爬虫程序。

**第一部分**，获取网页中的表格数据。此处有两个模块：requests和BeautifulSoup。requests获取由url指定网址的网页html，BeautifulSoup则解析html语法，从中提取所需的文本。soup是网页html，soup.select选择html语法中的tag，提取html语句。利用chrome开发者工具查看html语句，用右键->审查，选左侧element，图中看出，表格数据放在标签tag为‘tr’和‘td’中，tr是表格行的语句，td是表格列的语句。整个网页html只有标签为td的语句包含表格中的数据，因此td作为提取tag。



import requests

url = 'http://www.customs.gov.cn/customs/302249/302274/302277/302276/2709048/index.html'

res = requests.get(url)

from bs4 import BeautifulSoup

res.encoding = 'utf-8'

soup = BeautifulSoup(res.text)

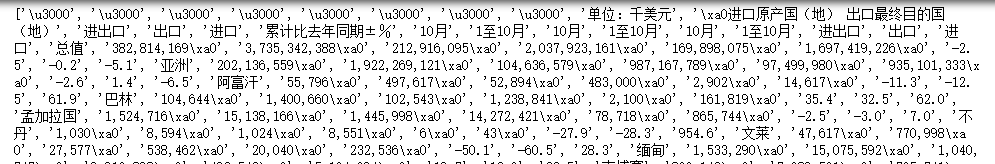
cont1 = soup.select('td')

data1 = [n for n in range(0, len(cont1))]

for i in range(len(cont1)):

    data1[i] = cont1[i].text

cont1得到所有标签为td的html语句，data1从cont1中取出每个语句的文本，即网页表格所提供的数据。



**第二部分**，整理列表。从表格看出，有规律的表格从“总值”开始的那一行，故有用的列表从“总值”所在行开始，这样构成列表data2。从上图看出，表格数据中含“\xa0”，这是空格符。对于linux版本的excel，可以识别该符号。而windows下excel则无法处理这个空格符，也就是不能将文本型数字转换成可计算的数字型。因此需要将列表中的字符串中去除\xa0。



for i in range(0,50):

    if data1[i] == '总值':

        first\_i = i

data2 = [n for n in range(0, len(cont1)-first\_i)]

for k in range(0,len(cont1)-first\_i):

    data2[k] = data1[first\_i+k]

# 去除list中的 \xa0

for i in range(0, len(data2)):

    temp = data2[i]

    s = "".join(temp.split())

    data2[i] = s

**第三部分**，将整理好的列表data2写入xls文件，这里需要xls写入模块 xlwt

file\_path = 'd:\\jupyterNt\\海关数据222.xls'

import xlwt

wbk = xlwt.Workbook()

sheet1 = wbk.add\_sheet(u'sheet1',cell\_overwrite\_ok=True) #创建sheet

#将数据写入第 i 行，第 j 列

for i in range(0,len(data2)//10):

    for j in range(0,10):

        sheet1.write(i,j,data2[i\*10+j])

    i = i + 1

wbk.save(file\_path) # 写入磁盘文件

由此，可以进入excel，建立海关数据表格。需注意：linux下wps的表格虽然和windows excel操作相同，但对于用逗号分隔的文本型数字，如3,285,129,432，若该数字超过32bit的十进制值，则从文本型数字转换成数字型会出错，原因可能是整数位32bit（32位程序而非64位），将因超过范围而出错。在windows下excel则不会出错，也许是该excel是64位程序，没有超过其整数范围。