1. 执行程序名为`公司名称识别并建立上下游关系.py`程序，

```
import xlwt

# global firm names 和 us names 的 Global Company Key 和 Company Name 提取
global_workbk = xlrd.open_workbook(r'C:\Users\jc\Documents\Pydata'+
                                  '\Database Table\global firm names.xlsx')
```

上面最后一行在读取 *global firm names.xlsx* 文件，使用时请改成**对应文件保存的位置**

```
global_worksh = global_workbk.sheet_by_name('0x77igavdumz8vul')
global_cpnames = global_worksh.col_values(colx = 7, start_rowx = 1)
global_cpkey = global_worksh.col_values(colx = 0, start_rowx = 1)
# 组成列表
global_namekey = list(zip(global_cpnames, global_cpkey))



us_workbk = xlrd.open_workbook(r'C:\Users\jc\Documents\Pydata'
                              +'\Database Table\\us names.xlsx')
```

上面最后一行在读取 *us_ names.xlsx* 文件，使用时请**改成对应文件保存的位置**

```
us_worksh = us_workbk .sheet_by_name('76aqys7wh9axjpme')
us_cpnames = us_worksh.col_values(colx = 9, start_rowx = 1)
us_cpkey = us_worksh.col_values(colx = 0, start_rowx = 1)
# 组成列表
us_namekey = list(zip(us_cpnames,us_cpkey))
# 将customer表中的Customer Name列和Global Company Key 提取出
customer_workbk = xlrd.open_workbook(r'C:\Users\jc\Documents\Pydata'+
                                     '\Database Table\customer.xlsx')
```

上面最后一行在读取*customer.xlsx*文件位置，使用时请**改成对应文件保存的位置**

```
customer_worksh = customer_workbk.sheet_by_name('vozkv0ioajsw5wov')
customer_downstream = customer_worksh.col_values(colx = 2, start_rowx = 1)
customer_upkey = customer_worksh.col_values(colx = 0, start_rowx = 1)
# 组成列表
customer_up_down = list(zip(customer_upkey,customer_downstream))
```

```
# global firm names 和 us names 的Global Company Key 和 Company Name列表合并后去重
Allnamekey_lst = list(set( global_namekey+us_namekey))
#对customer_up_down列表也去重
customer_up_down = list(set(customer_up_down))
#按照Global Company Key进行排序
Allnamekey_lst.sort(key=lambda x:x[1])

# 将Global Company Key 和 Company Name 写入Allcompany表中
Allcompany = xlwt.Workbook()
Allcompany_sheet = Allcompany.add_sheet('sheet1')
name_list = ['Global Company Key','Company Name(upstream)','Downstream']
for i in name_list:
    Allcompany_sheet.write(0, name_list.index(i), i)

for namekey in Allnamekey_lst:
    Allcompany_sheet.write(Allnamekey_lst.index(namekey)+1,0,namekey[1])
    Allcompany_sheet.write(Allnamekey_lst.index(namekey)+1,1,namekey[0])

cust_num = [[i[1],0] for i in Allnamekey_lst]
#比较 upstream key(customer_up_down[i][0])和global company key(Allnamekey_lst[j][1])
for upkey_down in customer_up_down:
    for name_key in Allnamekey_lst:
        if upkey_down[0] == name_key[1]:
            Allcompany_sheet.write(Allnamekey_lst.index(name_key)+1
                                   ,2+cust_num[Allnamekey_lst.index(name_key)][1]
                                   ,upkey_down[1])
            cust_num[Allnamekey_lst.index(name_key)][1] += 1

# 保存文件
Allcompany.save('Allcompany.xlsx')
```

最后一行是保存*Allcompany.xlsx*文件的位置，可以不修改，默认保存在同py文件的文件夹下。

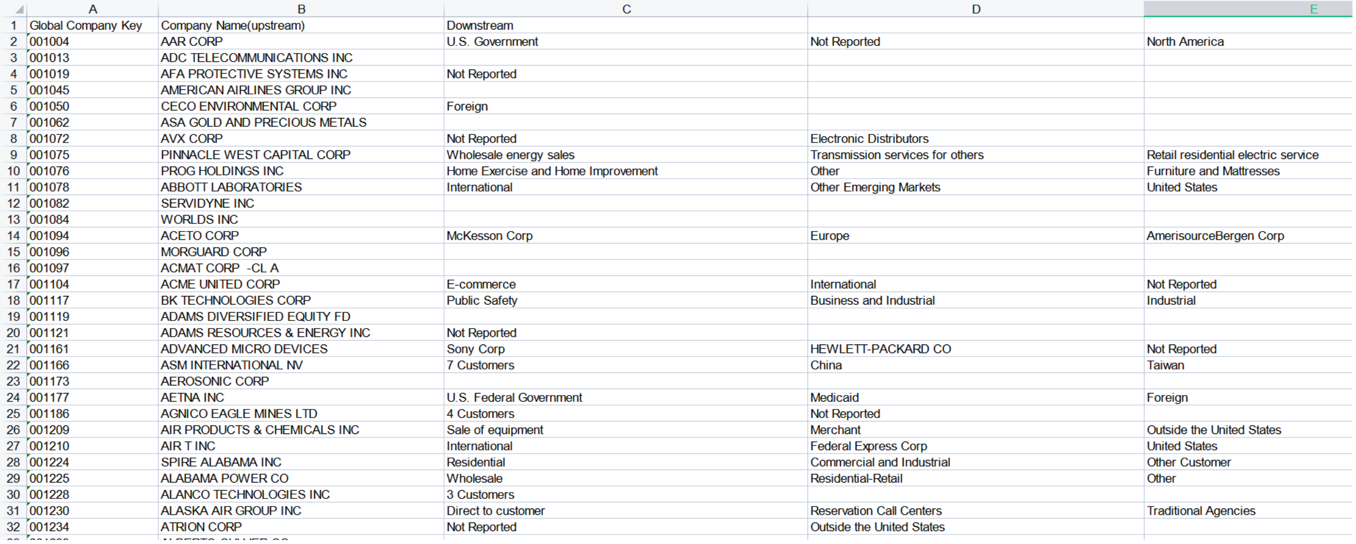
> Allcomany.xlsx文件里保存的有
> 第一列：Global Company Key 每个公司特有的关键字
> 第二列：Company Name(upstream) 上游公司的全称
> 第三列以后：Downstream 与上有公司对应的，下游公司的缩写

> 对数据进行了哪些处理：

> 1. 里面上游公司的名称没有重复，都只出现一次
> 2. 数据都是按照 Global Company Key 公司特有的关键字 从小到大排列

在我的电脑上代码执行大约需要25min

执行后结果如下:



	A	B	C	D	E
1	Global Company Key	Company Name(upstream)	Downstream		
2	001004	AAR CORP	U.S. Government	Not Reported	North America
3	001013	ADC TELECOMMUNICATIONS INC			
4	001019	AFA PROTECTIVE SYSTEMS INC	Not Reported		
5	001045	AMERICAN AIRLINES GROUP INC			
6	001050	CECO ENVIRONMENTAL CORP	Foreign		
7	001062	ASA GOLD AND PRECIOUS METALS			
8	001072	AVX CORP	Not Reported	Electronic Distributors	
9	001075	PINNACLE WEST CAPITAL CORP	Wholesale energy sales	Transmission services for others	Retail residential electric service
10	001076	PROG HOLDINGS INC	Home Exercise and Home Improvement	Other	Furniture and Mattresses
11	001078	ABBOTT LABORATORIES	International	Other Emerging Markets	United States
12	001082	SERVIDYNE INC			
13	001084	WORLDS INC			
14	001094	ACETO CORP	McKesson Corp	Europe	AmerisourceBergen Corp
15	001096	MORGUARD CORP			
16	001097	ACMAT CORP -CL A			
17	001104	ACME UNITED CORP	E-commerce	International	Not Reported
18	001117	BK TECHNOLOGIES CORP	Public Safety	Business and Industrial	Industrial
19	001119	ADAMS DIVERSIFIED EQUITY FD			
20	001121	ADAMS RESOURCES & ENERGY INC	Not Reported		
21	001161	ADVANCED MICRO DEVICES	Sony Corp	HEWLETT-PACKARD CO	Not Reported
22	001166	ASM INTERNATIONAL NV	7 Customers	China	Taiwan
23	001173	AEROSONIC CORP			
24	001177	AETNA INC	U.S. Federal Government	Medicaid	Foreign
25	001186	AGNICO EAGLE MINES LTD	4 Customers	Not Reported	
26	001209	AIR PRODUCTS & CHEMICALS INC	Sale of equipment	Merchant	Outside the United States
27	001210	AIR T INC	International	Federal Express Corp	United States
28	001224	SPIRE ALABAMA INC	Residential	Commercial and Industrial	Other Customer
29	001225	ALABAMA POWER CO	Wholesale	Residential-Retail	Other
30	001228	ALANCO TECHNOLOGIES INC	3 Customers		
31	001230	ALASKA AIR GROUP INC	Direct to customer	Reservation Call Centers	Traditional Agencies
32	001234	ATRION CORP	Not Reported	Outside the United States	

2. **对获取到的全称进行预处理:**
   创建表Allcpdata,全称All Company Data表格，用于存储全称公司的：
   Global Company key, Full name, company type , country ， Cleande Full Name
   其中 Cleande Full Name 表示去除了公司后缀的全称,用于后续与缩写的匹配
   country 用于存储公司可能所在的国家，有些无法识别出就空着
   以下是还没有完成的预处理代码。

```
import xlrd
import xlwt
import xlsxwriter
import re
import jellyfish
from cleanco import cleanco

# def cleanco_string(string):
#     '''
#     消除公司后缀影响
#     ----------
#     string : TYPE
#         full name
#     Returns
#     -------
#     None.
#     '''
#     string = cleanco(string)
#     cotype = string.type()
#     cocountry = string.country()

#     return string.clean_name()



'''
从Allcompany中读取数据
 Allcpkey 用于保存公司缩写
 Allcpname用于保存公司全称
 '''
Allcompany = xlrd.open_workbook(r'C:\Users\jc\Documents\大学'
                                +'\\0大三其他\\_金融数据挖掘科研课题'
                                +'\Allcompany.xlsx')
Allcompany_sheet = Allcompany.sheet_by_name('sheet1')
Allcpkey = Allcompany_sheet.col_values(colx = 0, start_rowx = 1)
Allcpname = Allcompany_sheet.col_values(colx = 1, start_rowx = 1)



'''
创建表Allcpdata,全称All Company Data表格，用于存储全称公司的:
Global Company key, Full name, company type , country ，Cleande Full Name
其中 Cleande Full Name 表示去除了公司后缀的全称,用于后续与缩写的匹配
country 用于存储公司可能所在的国家，有些无法识别出就空着
'''

# 1.创建All Company Data表格，写好表头
Allcpdata_wb = xlwt.Workbook()
Allcpdata = Allcpdata_wb.add_sheet('sheet1')
name_list = ['Global Company key', 'Full name',
             'company type' , 'country' ,'Cleande Full Name']
for i in range(len(name_list)):
    Allcpdata.write(0, i , name_list[i])

# 2.将Global Company key写进表中 第i+1行,第0列
for i in range(len(Allcpkey)):
    Allcpdata.write(i+1, 0 , Allcpkey[i])
```

```
# 3.将Full name写进表中 第i+1行，第1列
for i in range(len(Allcpname)):
    Allcpdata.write(i+1, 1, Allcpname[i])
```