> 今天主要进行的是**第三步**，第一、二步为将最近的代码工作串起来而写上

1. 获取要处理的基本数据

执行程序名为 `公司全称和公司缩写建立上下游关系.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 每个公司特有的一串数字
> 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:

执行 `公司全名的预处理.py` 文件
从第一步中得到的Allcompany.xlsx中获取所有公司的Allcpkey(每个公司特有的一串数字）和Allcpname（公司全名）

注 : 第一行路径对应上面**第一部分代码中生成的Allcompany文件位置**

```
'''
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 match表格的表头

```
# 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'
,'Cleaned 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])
```

4. 将company type写入表中 第i+1行，第2列, 用 cleanco中typesources()和match()判断

```
for i in range(len(Allcpname)):
    Allcpdata.write(i+1, 2, matches(Allcpname[i], typesources()))
```

5. 将country 写入表中 第i+1行，第3列, 用cleanco中match()和countrysources()判断

```
for i in range(len(Allcpname)):
    Allcpdata.write(i+1, 3,matches(Allcpname[i],countrysources()))
```

6. 清理不必要的后缀与符号，并转换为小写字符 写入表中，第i+1行，第4列，用clean_name()判断。

```
for i in range(len(Allcpname)):
    Allcpdata.write(i+1, 4, (clean_name(Allcpname[i]).lower()))
```

关于 clean_name 方法解释如下：

1. 用cleanco中的basename删除后缀
   (A.S./ N.V./ SA/SG S.A.与这一步共同进行单独处理)
   (-CL A,-CL B,-CL C ,-CL I)单独处理
2. 删除未识别出的后缀
   ETF ， A/S ， -ADR ， INC-OLD,group，
3. 再用cleanno中的basename删除一次，确保删除干净了

以下是clean_name方法，**基本能将company name中所有后缀与国家名都删除，并不丢失此外的其他字符。**

```
def clean_name(string):
    '''
    1. 用cleanco中的basename删除后缀
    (A.S./ N.V./  SA/SG S.A.与这一步共同进行单独处理)
    (-CL A,-CL B,-CL C ,-CL I)单独处理
    2.删除未识别出的后缀
    ETF ， A/S  ， -ADR ， INC-OLD,group，
    3. 再用cleanno中的basename删除一次，确保删除干净了
    ---
    return 公司清理后缀后的名称
    '''
    #1.
    # 截取最后5个判断是否有-CL_?类后缀，如果有则删去
    str_suffix0 = string[:len(string)-6:-1][::-1].replace(' ','')
    if '-CL' in str_suffix0:
        string = string[:string.find('-')]
    string = re.sub(r'[\W]',' ',string) # 首先删去字符影响，替换为空格
```

```
    string = cleanco.basename(string) # 用cleanco里自带的basename做第一次后缀清除
    clean_suffix0 = ['A/S','N.V.','SA/AG','S.A.']
    for cs in clean_suffix0:
        if cs in string:
            string = string[:string.rfind(cs[0])]
    #2.
    # 进行第二次后缀清理
    # 截取最后 5 个字符,并删除其中的空格 为什么是 5 个（因为根据观察，自动清除的后缀里，最长的是GROUP，所以取最后5个判断不会有漏）
    str_suffix = string[:len(string)-6:-1][::-1].replace(' ','')
    clean_suffix = ['CO','ETF','AS','ADR','SA','AG','OLD','GROUP'] # 判断是否包含这些后缀
    for cs in clean_suffix:
        if cs in str_suffix:
            string = string[:string.rfind(cs[0])]
    #3.
    #用basename进行第三次后缀处理
    string = cleanco.basename(string)
    return string
```

最后进行文件的保存：

```
Allcpdata_wb.save('All Company Data.xlsx')
```

数据处理结果如下：

	A	B	C	D	E	F
57186	51335	IPD GROUP LIMITED	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	ipd	
57187	51336	SHAPE AUSTRALIA CO LTD	LimitedCorporation	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	shape australia	
57188	51340	LATENT VIEW ANALYTICS LTD	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	latent view analytics	
57189	51341	FINATEXT HOLDINGS LTD	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	finatext holdings	
57190	51343	KOHOKU KOGYO CO. LTD	CorporationLimited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	kohoku kogyo	
57191	51345	POLTREG S.A.	CorporationLimited Liability Company	ArgentinaBrazilChileColumbiaDominican RepublicEcuadorGuatemalaLuxembergMexicoPeruPolandPortugalRomaniaSpain	poltreg	
57192	51346	MARKS ELECTRICAL GRP PLC	Limited Liability Company	NigeriaUnited Kingdom	marks electrical grp	
57193	51350	XPON TECHNOLOGIES GROUP LTD	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	xpon technologies	
57194	51351	HANGZHOU ZHENGDIANG CORP LTD	CorporationLimited	PhilippinesUnited States of AmericaHong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	hangzhou zhengxiang	
57195	51355	HEALTHLEAD PUBLIC COMPANY	Corporation	United States of America	healthlead public	
57196	51358	DUELL OY J	Limited Liability Company	Finland	duell	
57197	51359	NICO RESOURCES LIMITED	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	nico resources	
57198	51362	RENEWABLE JAPAN CO.LTD.			renewable japan	
57199	51366	GUANGDONG LIFESTRONG PHARMAC			guangdong lifestrong pharmac	
57200	51367	MH ROBOT & AUTOMATION CO.LTD	LimitedCorporation	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	mh robot automation	
57201	51368	CAM RANH PORT JOINT STOCK CO	Corporation	United States of America	cam ranh port joint stock	
57202	51370	SICHUAN DISCOVERY DREAM			sichuan discovery dream	
57203	51372	OMNIPOTENT INDUSTRIES LTD	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	omnipotent industries	
57204	51373	GO FASHION LIMITED	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	go fashion	
57205	51374	ARISTON HOLDING N V			ariston holding n v	
57206	51384	ALFONSINO S.P.A.	Corporation	Chileltaly	alfonsino s p a	
57207	51385	INNER MONGOLIA XINHUA			inner mongolia xinhua	
57208	51387	ISBIR SENTETIK DOKUMA SANAYI			isbir sentetik dokuma sanayi	
57209	51389	T.R.V. RUBBER PRODUCTS			t r v rubber products	
57210	51391	PC ILETISIM VE MEDYA			pc iletisim ve medya	
57211	51392	NIVIKA FASTIGHETER AB PUBL	Limited Liability Company	LithuaniaSwedenSwitzerland	nivika fastigheter ab publ	
57212	51393	SANWAYUKA INDUSTRY CORPORATI			sanwayuka industry corporati	
57213	51398	EXAWIZARDS INC	Corporation	PhilippinesUnited States of AmericaUnited States of America	exawizards	
57214	51399	CROOOOBER CO LTD	LimitedCorporation	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	crooooober	
57215	51401	HYBRID TECHNOLOGIES CO LTD	LimitedCorporation	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	hybrid technologies	
57216	51404	BEIJING FJR OPTOELECTRONIC			beijing fjr optoelectronic	
57217	51406	COPLUS INC	Corporation	PhilippinesUnited States of AmericaUnited States of America	coplus	
57218	51407	TURN CLOUD TECHNOLOGY			turn cloud technology	
57219	51408	ENERAQUA TECHNOLOGIES PLC	Limited Liability Company	NigeriaUnited Kingdom	eneraqua technologies	
57220	51416	GROUPE BERKEM SA	Limited Liability Company	ChileDominican RepublicEcuadorGuatemalaHaitiPeruSwitzerland	groupe berkem	
57221	51417	CLOUD MUSIC INC	Corporation	PhilippinesUnited States of AmericaUnited States of America	cloud music	
57222	51420	NIFTY LIFESTYLE CO LTD	LimitedCorporation	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	nifty lifestyle	
57223	51421	ASIAQUEST CO LTD	LimitedCorporation	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	asiaquest	
57224	51426	UNIVERSAL NETWORK SYSTEMS			universal network systems	
57225	51427	ADVANCED GRAPHENE PRODUCTS			advanced graphene products	
57226	51434	ATTURRA LIMITED	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	attura	
57227	51435	DOSILICON CO LTD	LimitedCorporation	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of AmericaUnited States of America	dosilicon	
57228	51436	HSENID BUSINESS SOLUTIONS PL			hsenid business solutions pl	
57229	51437	SUNMOW HOLDING BERHAD			sunmow holding berhad	
57230	51439	ANHUI PROVINCIAL ARCHITECTUR			anhui provincial architectur	
57231	51443	TARSONS PRODUCTS LTD	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	tarsons products	
57232	51445	KLARABO SVERIGE AB (PUBL)	Limited Liability Company	LithuaniaSwedenSwitzerland	klarabo sverige ab publ	
57233	51447	SOLID FORSAKRINGSAKTIEBOLAG			solid forsakringsaktiebol	
57234	51449	DAIWA HOUSE LOGISTICS TRUST			daiwa house logistics trust	
57235	51454	SHANGHAI YIZHONG PHARMA CO	Corporation	United States of America	shanghai yizhong pharma	
57236	51455	ZHANG XIAOQUAN INC	Corporation	PhilippinesUnited States of AmericaUnited States of America	zhang xiaoquan	
57237	51456	BEFOREPAY GROUP LIMITED	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	beforepay	
57238	51457	SECURE INC	Corporation	PhilippinesUnited States of AmericaUnited States of America	secure	
57239	51463	HOEGH AUTOLINERS ASA	Limited Liability Company	Norway	hoegh autoliners	
57240	51465	LAMOR CORPORATION OYJ	CorporationLimited Liability Company	United States of AmericaFinland	lamor	
57241	51478	ASHTEAD TECHNOLOGY HOLDINGS			ashtead technology holdings	
57242	51489	JIANGSU TONGLIN			jiangsu tonglin	
57243	51491	IVECO GROUP N V			iveco group n v	
57244	51492	BIRDDOG TECHNOLOGY LIMITED	Limited	Hong KongIsraelNew ZealandPakistanUnited KingdomUnited States of America	birddog technology	
57245	51497	HEALTHBEACON PLC	Limited Liability Company	NigeriaUnited Kingdom	healthbea	
57246	51498	HANGZHOU BIO SINCERITY PHARM			hangzhou bio sincerity pharm	
57247	51500	ZJMI ENVIRONMENTAL ENERGY CO	Corporation	United States of America	zjmi environmental energy	

3.对获取到的缩写进行预处理

执行`公司缩写的预处理.py`文件，基本步骤与第二步差不多

第一步， 从 `customer.xlsx` 中读取数据， company_abbr 用于保存公司缩写, company_upkey用于保存用于识别缩写对应的上游公司的特殊数字串

```
customer_wb = xlrd.open_workbook(r'C:\Users\jc\Documents\Pydata'+
                                    '\Database Table\customer.xlsx')
customer = customer_wb.sheet_by_name('vozkv0ioajsw5wov')
company_abbr = customer.col_values(colx = 2, start_rowx = 1)
company_upkey = customer.col_values(colx = 0, start_rowx = 1)
# 组成列表
abbr_upkey_zip = list(set((zip(company_abbr,company_upkey))))
# 第i组： abbr_upkey_zip[i][0]公司缩写 abbr_upkey_zip[i][1]上游公司key
```

第二步，创建表Abbrdata,全称Abbreviation Data表格，用于存储缩写公司的：
Global Company key(upstream), Abbreviation, company type , country ， Cleaned Abbreviation
其中 Cleaned Abbreviation 表示去除了公司后缀的缩写,用于与公司全称的匹配
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Abbreviation Data表格，写好表头

```
Abbrdata_wb = xlwt.Workbook()
Abbrdata = Abbrdata_wb.add_sheet('sheet1')
name_list = ['Global Company key(upstream)', 'Abbreviation',
              'company type' , 'country' ,'Cleaned Abbreviation']

for i in range(len(name_list)):
    Abbrdata.write(0, i , name_list[i])
```

2.将Global Company key(upstream)写进表中 第i+1行,第0列

```
for i in range(len(abbr_upkey_zip)):
    Abbrdata.write(i+1, 0 , abbr_upkey_zip[i][1])
```

3.将Abbreviation写进表中 第i+1行，第1列

```
for i in range(len(abbr_upkey_zip)):
    Abbrdata.write(i+1, 1, abbr_upkey_zip[i][0])
```

4.将company type写入表中 第i+1行，第2列，用 cleanco中typesources()和match()判断

```
for i in range(len(abbr_upkey_zip)):
    Abbrdata.write(i+1, 2, matches(abbr_upkey_zip[i][0], typesources()))
```

5.将country 写入表中 第i+1行，第3列, 用cleanco中match()和countrysources()判断

```
for i in range(len(abbr_upkey_zip)):
    Abbrdata.write(i+1, 3,matches(abbr_upkey_zip[i][0],countrysources()))
```

6.清理不必要的后缀与符号，并转换为小写字符 写入表中，第i+1行，第4列，用 cleanco.clean_name()判断

```
for i in range(len(abbr_upkey_zip)):
    Abbrdata.write(i+1, 4, (clean_name(abbr_upkey_zip[i][0]).lower()))

Abbrdata_wb.save('Abbreviation Data.xlsx')
```

最后保存数据在`Abbreviation`

第三步中对缩写的数据清洗有点瓶颈
面临问题：

1. 多个缩写配对同一个公司全称
2. 多个公司全称匹配一个缩写
   以上进一步判断是否该通过人工进行，讨论之后再做决定

其他工作：

> 1. 相关的 Allcompany.xlsx， All Company Data.xlsx，Abbreviation Data.xlsx 文件已上传至OneDrive--Liuchenxin文件夹
> 2. 阅读《Using Natural Language Processing for Supply Chain Mapping》论文，阅读部分论文的笔记`(NOTE) Schöpper, H., Kersten, W., Using Natural Language Processing for Supply Chain Mapping`已上传至OneDrive---Liuchenxin文件夹