今天主要是和方翔讨论的数据匹配的算法。
为了提高**数据匹配的可用率**，讨论结果如下：

1. 在匹配的时候应该消除`./&*][\|*`等类似的特殊符号，之前都没有做到消除
2. 通过方翔分享的参考链接关于在Python中利用字符串 fuzzy matching （模糊匹配）进行数据库merge/join的一些经验和tips
   了解到python中的jellyfish可以很好的进行模糊匹配，返回结果是模糊匹配的匹配比例，尝试用这个写出更有效的代码

> 小反思：在提升代码的时候除了自己想，参考别人的名称模糊匹配代码也是很有效的，这一点方翔做的比我好，明天在注意修改代码的时候应该注意这一点，参考更好的代码在做数据预处理的时候也是更高效的。

以下代码不具有参考价值，仅代表我还在修改中的代码存档

```
import xlrd
import xlwt
import xlsxwriter
import re
import jellyfish

def spl_string(string):
    '''
    以空格为分隔符划分customer name
    :param string:
    :return:
    '''
    string = re.sub(r'[-,/&()]|\'\sBD',r'', string) #预处理  去除customer name 中的字符
    outcome = string.split()
    return outcome

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')
# 公司全称
Allcpname = Allcompany_sheet.col_values(colx = 1, start_rowx = 1)
```

```
customer_workbk = xlrd.open_workbook(r'C:\Users\jc\Documents\Pydata'+
                                      '\Database Table\customer.xlsx')
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)
# 上游公司的key 和 下游公司缩写组成的元组（不可修改）
customer_up_down = list(zip(customer_upkey,customer_downstream))
# 去重
customer_up_down = list(set(customer_up_down))

# 缩写匹配的表格
matchAllname = xlsxwriter.Workbook()
matchAllname_sheet = matchAllname.add_worksheet('sheet1')

name_list = ['Abbreviation','Full name, count',
             'Full name, count','Full name, count']# 0 1
for i in name_list:
    matchAllname_sheet.write_row(0, name_list.index(i), i)

for abb in customer_up_down:
    matchAllname_sheet.write_row(customer_up_down.index(abb)+1,0,abb[1])


# 匹配 下游公司缩写 和 上游公司的全名
cust_num = [[i[1],0] for i in customer_up_down]
for uk_down in customer_up_down:
    # 将down按照空格进行拆分
    down_sp = ''
    for sp in spl_string(uk_down[1]):
        down_sp += sp
    # 创建一个数组，临时存放这些匹配的公司名，并写入数组中
    matchsp = []
    for sp in down_sp:
        for alnamedown in Allcpname:
            if(jellyfish.levenshtein_distance(u'',u''))>0.5:
                # 将匹配的公司名写入数组中
                matchsp.append(alnamedown)
```