江西财经大学
JIANGXI UNIVERSITY OF FINANCE AND ECONOMICS

实 验 报 告

课程名称 金融数据获取与处理

班　　级 金融科技202

学　　号 0204985

姓　　名 陈琳

任课教师 吴燕丰

上课学期 2022 至 2023 学年 第 2 学期

教学班级 001

一、 实验要求：

- 选择行业中的 10 家公司，下载每家公司最近 10 年的年报。
- 将营业收入、归属于上市公司股东的净利润当年数据解析出来。
- 将公司资料中的公司网址、电子信箱、办公地址，以及董事会秘书的姓名、电话、电子信箱解析出来，保存为 csv 文件。
- 将 10 家公司 10 年的营业收入绘制成一张折线图。
- 选择一家公司，将 10 年的营业收入和归属于上市公司股东的净利润绘制成一张折线图。

二、 实验报告内容

1 行业及上市公司选择

1.1 上市公司所属行业

选择的上市公司所属行业的基本信息如表 1-1 所示。

表 1-1 上市公司所属行业信息

门类名称及代码	制造业(C)
行业大类代码	26
行业大类名称	化学原料和化学制品制造

1.2 选择的 10 上市家公司

在化学原料和化学制品制造行业中选择的 10 家上市公司基本信息如表 1-2 所示。

表 1-2 10 家上市公司基本信息

序号	上市公司代码	上市公司简称
1	600075	新疆天业
2	600078	*ST 澄星
3	600096	云天化
4	600135	乐凯胶片

5	600141	兴发集团
6	600160	巨化股份
7	600165	新日恒力
8	600230	沧州大化
9	600249	两面针
10	600273	嘉化能源

2 年报下载

2.1 获取年报链接的 HTML 文件

通过 Web 应用程序自动化测试工具 Selenium 从上海证券交易所网站上获取上市公司年报的下载地址链接，并保存为 HTML 文件，如图 2-1 所示。其中，每个文件中都包含了该公司近年来的年度报告和年度报告摘要，以及修订版或者更正版的年报和公司其他相关报告。在下载年报前，需要先去除不相关文件的链接。

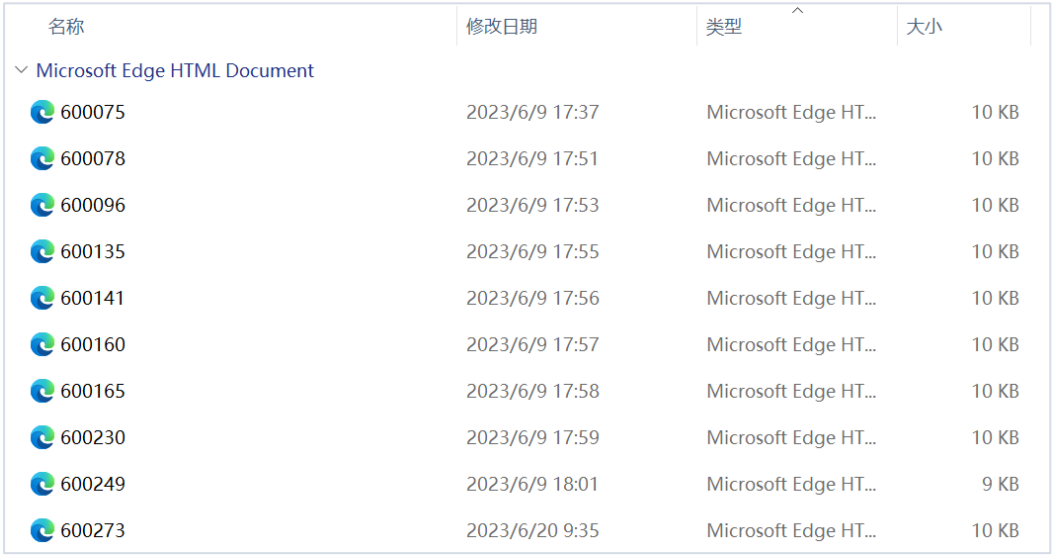

图 2-1 10 家公司包含年报链接的 HTML 文件

2.2 根据链接下载年报

在得到以上 HTML 文件后，首先要通过设定合理的规则对其中的链接地址进行过滤，具体规则如下：

(1) 若标题中包含“更正版”，则修改其对应的日期为所报告年度的下一年的 12

月 31 日。该条规则能确保在按照日期降序排列后，未更正版年报位于更正版年报的下一条记录，方便后续删除未更正版年报的下载链接，同时也能确保下载 pdf 年报时命名的正确性。

(2) 若标题中包含“修订版”或“更正版”，则删除该条记录的下一条记录。该条规则能保证下载的年报是以及修订或者更正后的。

(3) 删除所有标题中包含“摘要”的记录所对应的链接，并且只保留标题中包含“年报”或者“年度报告”的记录所对应的链接。根据该条规则可以去除所有年报摘要和非年报文件的下载链接。

(4) 只保留日期在近 10 年内的记录。该条规则能确保下载的年报是最近 10 年的而不包括其他年份。

根据过滤后的年报链接下载的 10 家上市公司近 10 年(2013 年-2022 年)的年报如图 2-2 所示。

图 2-2　下载的 10 家公司 2013-2022 年全部年报

3　年报解析

3.1　提取公司基本信息

通过 PyMuPDF 包在 Python 中读入下载好的年报 pdf 文件，通过代码将公司简介中的公司办公地址、公司网址、电子信箱以及董事会秘书的姓名、电话电子

信箱解析出来。并且以上公司基本信息以最新公布的为准，即通过各公司 2022 年的年度报告内容提取出来。得到的结果分别如图 3-1 和图 3-2 所示，图 3-1 是 10 家公司的办公地址、公司网址以及电子信箱的相关信息，而图 3-2 则是董事会秘书姓名、电话和电子信箱的相关信息。

公司代码	公司简称	公司办公地址	公司网址	电子信箱
600075	新疆天业	新疆石河子市经济技术开发区北三东路36号	http://www.xj-tianye.com	master@xj-tianye.com
600078	*ST澄星	江苏省江阴市梅园大街618号	www.cxpcchina.com	cx@cxpcchina.com
600096	云天化	云南省昆明市滇池路1417号	www.yyth.com.cn	zqb@yth.cn
600135	乐凯胶片	河北省保定市乐凯南大街6号	http://lkjp.luckyfilm.com/	stock@luckyfilm.com
600141	兴发集团	湖北省宜昌市高新区发展大道62号悦和大厦	www.xingfagroup.com	dmb@xingfagroup.com
600160	巨化股份	浙江省衢州市柯城区	http://www.jhgf.com.cn	jhgf@juhua.com
600165	新日恒力	宁夏银川市金凤区宁安东巷108号创新园A栋	www.xinrihengli.com	official@xinrihengli.com
600230	沧州大化	沧州市运河区永济东路20号	www.czdh.chemchina.com	caiwu@czdh.com.cn
600249	两面针	广西柳州市东环大道282号	www.lmz.com.cn	lmzstock@lmz.com.cn
600273	嘉化能源	浙江省嘉兴市乍浦滨海大道2288号	www.jhec.com.cn	jhnydsh@163.com

图 3-1　公司办公地址、公司网址和电子信箱信息

公司代码	公司简称	董事会秘书姓名	董事会秘书电话	董事会秘书电子信箱
600075	新疆天业	李升龙	0993-2623118	Lishenglong11223@163.com
600078	*ST澄星	汪洋	0510-80622329	cx@cxpcchina.com
600096	云天化	钟德红	（0871）64327127	zhongdehong@yth.cn
600135	乐凯胶片	张永光	0312-7922692	stock@luckyfilm.com
600141	兴发集团	鲍伯颖	0717-6760939	dmb@xingfagroup.com
600160	巨化股份	刘云华	0570-3091758	gfzqb@juhua.com
600165	新日恒力	张宝林	0951-6898015	official@xinrihengli.com
600230	沧州大化	刘晓婧	0317-3556143	caiwu@czdh.com.cn
600249	两面针	韦元贤	0772-2506159	lmzstock@lmz.com.cn
600273	嘉化能源	王庆营	0573-85583256	wangqingying@jiahuagufen.com

图 3-2　董事会秘书姓名、电话和电子信箱信息

3.2　提取营业收入和归属于上市公司股东的净利润

3.2.1　营业收入

用 Python 代码将每家公司每一年年报中当年的营业收入解析出来，得到的

营业收入数据如图 3-3 和图 3-4 所示，而表 3-1 是用 Python 计算的各公司近 10 年的平均营业收入。

从数据中可以看出，不同公司的营业收入额差距较大，大多从 10 亿元到 500 亿元不等，其中云天化营业收入大幅高于其他公司，其每年营业收入超过 500 亿元，在 2022 年更是超过了 700 亿元；而另外 9 家公司营业收入则大多在 200 亿元以下。

表 3-1　近 10 年的平均营业收入数据

序号	公司简称	平均营业收入/元
1	新疆天业	6.304570e+09
2	*ST 澄星	3.116169e+09
3	云天化	5.668888e+10
4	乐凯胶片	1.669298e+09
5	兴发集团	1.731466e+10
6	巨化股份	1.396651e+10
7	新日恒力	1.002719e+09
8	沧州大化	3.125419e+09
9	两面针	1.126124e+09
10	嘉化能源	5.520515e+09

索引	新疆天业	*ST澄星	云天化	乐凯胶片	兴发集团
2013	3.94111e+09	2.46019e+09	5.58962e+10	9.34927e+08	1.09344e+10
2014	4.27004e+09	2.59556e+09	5.44923e+10	9.46219e+08	1.1392e+10
2015	2.27504e+09	2.38773e+09	5.02669e+10	1.18317e+09	1.23923e+10
2016	5.59739e+09	3.26825e+09	5.26337e+10	1.42146e+09	1.45412e+10
2017	4.97716e+09	2.98602e+09	5.59714e+10	1.85132e+09	1.57578e+10
2018	4.82776e+09	3.14647e+09	5.2979e+10	1.86279e+09	1.78555e+10
2019	4.50374e+09	3.30996e+09	5.39759e+10	2.13646e+09	1.80387e+10
2020	8.99258e+09	3.13655e+09	5.21108e+10	2.05383e+09	1.83174e+10
2021	1.20146e+10	3.33341e+09	6.32492e+10	2.23503e+09	2.36067e+10
2022	1.16463e+10	4.53755e+09	7.53133e+10	2.06778e+09	3.03107e+10

图 3-3　2013-2022 年营业收入数据 (1)

索引	巨化股份	新日恒力	沧州大化	两面针	嘉化能源
2013	9.73654e+09	1.7639e+09	3.49109e+09	1.18355e+09	1.34943e+09
2014	9.76355e+09	1.34317e+09	3.11603e+09	1.18699e+09	3.38539e+09
2015	9.51616e+09	1.07665e+09	1.79445e+09	1.35319e+09	3.39133e+09
2016	1.01009e+10	2.78505e+09	2.9439e+09	1.56184e+09	4.50334e+09
2017	1.3768e+10	1.24794e+09	4.4172e+09	1.47212e+09	5.57601e+09
2018	1.56563e+10	5.52088e+08	4.43173e+09	1.24473e+09	5.60376e+09
2019	1.55952e+10	2.65757e+08	2.10557e+09	1.18667e+09	5.36903e+09
2020	1.60537e+10	1.19115e+08	1.65326e+09	6.85431e+08	5.56762e+09
2021	1.79856e+10	1.92387e+08	2.38738e+09	7.17526e+08	8.95657e+09
2022	2.14891e+10	6.81134e+08	4.91358e+09	6.69192e+08	1.15027e+10

图 3-4　2013-2022 年营业收入数据 (2)

3.2.2　归属于上市公司股东的净利润

用 Python 代码将每家公司每一年年报中当年的归属于上市公司股东的净利润解析出来，得到的数据如图 3-5 和图 3-6 所示

从数据中可以看出，10 家公司近 10 年的归属于上市公司股东的净利润大多低于 20 亿元，但在 2021 年和 2022 年，云天化和兴发集团归属于上市公司股东的净利润却达到了 60 亿元左右。此外，有些公司在某几年的归属于上市公司股东的净利润为负数，例如新日恒力在 2014、2016、2019、2021 和 2022 年以及两面针在 2015、2017、2019 和 2022 年的归属于上市公司股东的净利润为都为负数。

索引	新疆天业	*ST澄星	云天化	乐凯胶片	兴发集团
2013	-2.15672e+08	2.36238e+07	5.92711e+08	2.4136e+07	6.07608e+07
2014	3.89712e+07	2.22485e+07	-2.58347e+09	2.75548e+07	4.94311e+08
2015	4.11189e+07	1.79306e+07	1.01192e+08	3.66507e+07	7.72737e+07
2016	4.893e+08	6.00704e+07	-3.35949e+09	4.1143e+07	1.02018e+08
2017	5.39018e+08	5.88302e+07	2.01859e+08	5.9238e+07	3.20998e+08
2018	4.93594e+08	1.93336e+07	1.22765e+08	1.47395e+07	4.02261e+08
2019	2.90401e+07	6.03555e+07	1.51898e+08	8.50772e+07	3.02455e+08
2020	8.86522e+08	-2.21586e+09	2.72036e+08	3.72543e+07	6.23942e+08
2021	1.63831e+09	2.0152e+09	3.64194e+09	5.43371e+07	4.24659e+09
2022	8.53216e+08	5.21429e+08	6.02132e+09	3.86408e+07	5.85178e+09

图 3-5　2013-2022 年归属于上市公司股东的净利润数据 (1)

索引	巨化股份	新日恒力	沧州大化	两面针	嘉化能源
2013	2.53578e+08	1.10743e+07	1.22755e+08	1.01048e+07	-3.25137e+07
2014	1.62516e+08	-1.07964e+08	-1.92381e+08	2.19091e+07	5.79081e+08
2015	1.61778e+08	4.60887e+07	-6.09979e+08	-1.73258e+08	6.72367e+08
2016	1.5123e+08	-1.92021e+08	3.71064e+08	2.6903e+07	7.40477e+08
2017	9.35461e+08	3.82302e+07	1.28304e+09	-1.44e+08	9.68529e+08
2018	2.15256e+09	9.40061e+06	9.90949e+08	2.17197e+07	1.10018e+09
2019	8.9536e+08	-4.49521e+07	4.58425e+07	-5.38054e+07	1.22697e+09
2020	9.53752e+07	2.0906e+07	3.69358e+07	5.80063e+07	1.30373e+09
2021	1.10909e+09	-2.29649e+07	2.20281e+08	8.65826e+06	1.80819e+09
2022	2.38073e+09	-1.41484e+08	4.20098e+08	-3.85864e+07	1.59841e+09

图 3-6　2013-2022 年归属于上市公司股东的净利润数据 (2)

3.3　绘制相关折线图

3.3.1　十家公司的营业收入折线图

根据解析出来的 10 家上市公司近 10 年的营业收入数据，用 Python 的 Matplotlib 包将其绘制成折线图，得到的结果如图 3-7 所示。

从图中可以看出，云天化公司营业收入远远高于其他几家公司，并且在 2020

年及以前，其营业收入保持在 500 到 600 亿之间波动，而到 2021 年和 2022 年则出现了较大幅度的增长。

为了能更为清晰地观察其它公司的营业收入变动情况，将其余 9 家公司的营业收入单独绘制一张折线图，如图 3-8 所示。从中可以看出，兴发集团、巨化股份、新疆天业以及嘉化能源的营业收入都大体呈现出逐步上升的趋势；而*ST 澄星、两面针、乐凯胶片以及沧州大化则是在 50 亿元以下波动。

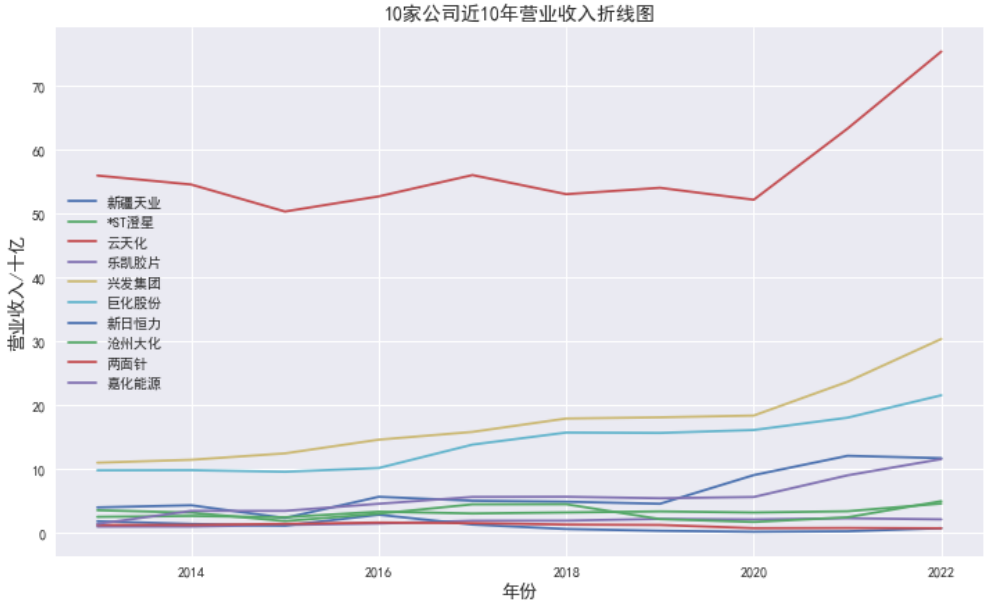

图 3-7　营业收入折线图

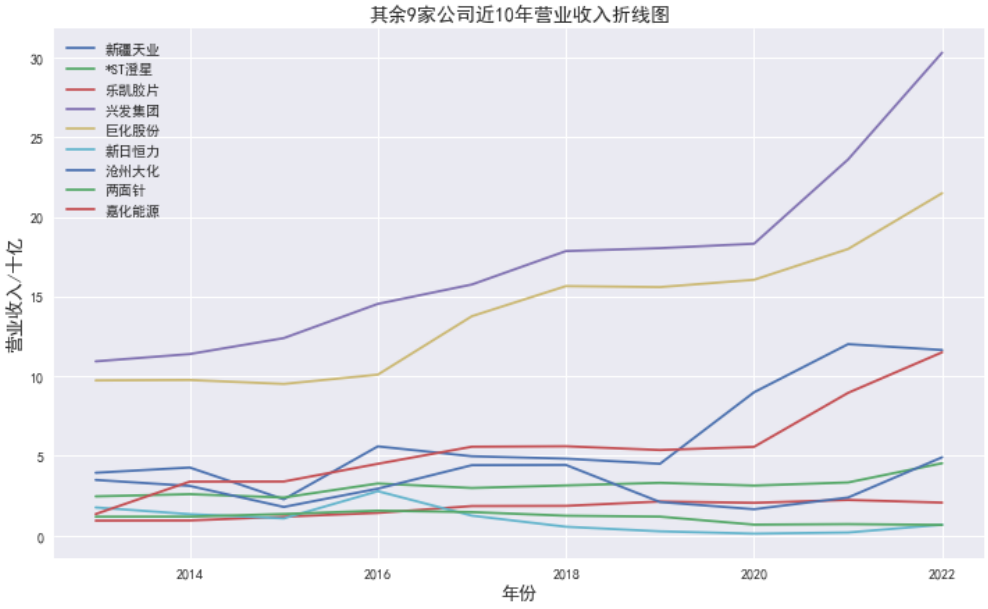

图 3-8　除云天化其余 9 家公司营业收入折线图

3.3.2　一家公司的营业收入和归属于上市公司股东的净利润折线图

选择新疆天业公司绘制其营业收入和归属于上市公司股东的净利润折线图，得到的结果如图 3-9 所示。

根据该图可知，新疆天业的营业收入在在 2019 年及之前都是低于 60 亿的，尤其是 2013 年，只有 20 多亿；而在 2020 年和 2021 年，公司营业收入实现了较大幅的增长，达到了 120 亿元，而在 2022 年又略微有所下降。相比于营业收入，公司近 10 年的归属于上市公司股东的净利润变动则相对平稳，且总体呈现出缓慢上涨的趋势。

图 3-9　新疆天业营业收入和归属于上市公司股东的净利润折线图

4 实验代码

4.1 代码——获取年报下载链接

```
import pytest
import time
import json
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.action_chains import ActionChains
from selenium.webdriver.support import expected_conditions
from selenium.webdriver.support.wait import WebDriverWait
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.desired_capabilities import
DesiredCapabilities
import time
import re
import pandas as pd


#获取年报链接，保存为 html 文件
def get_table_sse(code):
    browser = webdriver.Edge()
    browser.get("http://www.sse.com.cn/disclosure/listedinfo/regular/")
    time.sleep(3)
    browser.set_window_size(1456, 928)
    browser.find_element(By.ID, "inputCode").click()
    browser.find_element(By.ID, "inputCode").send_keys(code)
    browser.find_element(By.CSS_SELECTOR, ".sse_outerItem:nth-
child(4) .filter-option-inner-inner").click()
    browser.find_element(By.LINK_TEXT, "年报").click()
    dropdown = browser.find_element(By.CSS_SELECTOR,
".dropup > .selectpicker")
    dropdown.find_element(By.XPATH, "//option[. = '年报']").click()
    time.sleep(3)

    selector = "body > div.container.sse_content > div > div.col-lg-9.col-
xxl-10 > div > div.sse_colContent.js_regular > div.table-responsive >
table"
    element=browser.find_element(By.CSS_SELECTOR,selector)
    table_html=element.get_attribute('innerHTML')

    fname=f'{code}.html'
    f=open(fname,'w',encoding='utf-8')
```

```
    f.write(table_html)
    f.close()

    browser.quit()


#提取年报的链接地址并保存为 DataFrame
def get_data(tr):
    p_td=re.compile('<td.*?>(.*?)</td>',re.DOTALL)  # 每一行的内容
tds=p_td.findall(tr)
    #股票代码
    s=tds[0].find('>')+1  #起始索引
    e=tds[0].rfind('<')  #结束索引
    code=tds[0][s:e]
    #股票名称
    s=tds[1].find('>')+1
    e=tds[1].rfind('<')
    name=tds[1][s:e]
    #链接
    s=tds[2].find('href="')+6
    e=tds[2].find('.pdf"')+4
    href='http://www.sse.com.cn'+tds[2][s:e]  #补上上交所的顶级域名
    #标题
    s=tds[2].find('$(this))">')+10
    e=tds[2].find('</span>')
    title=tds[2][s:e]
    #日期
    date=tds[3].strip() #strip()如果两端有空格则把空格删除

    data=[code,name,href,title,date]
    return(data)

def parse_data(code):
    fname=f'{code}.html'
    f=open(fname,encoding='utf-8')
    table_html=f.read()
    f.close()

    p=re.compile('<tr>(.+?)</tr>',re.DOTALL)
    trs=p.findall(table_html)
    #
    trs_new=[]  #删除了空行
    for tr in trs:
        if tr.strip() != '':
```

```
            trs_new.append(tr)
    #
    data_all=[get_data(tr) for tr in trs_new[1:]] #第一行是标题行，不需要
    df=pd.DataFrame({
        'code':[d[0] for d in data_all],
        'name':[d[1] for d in data_all],
        'href':[d[2] for d in data_all],
        'title':[d[3] for d in data_all],
        'date':[d[4] for d in data_all]
        })
    return(df)
```

4.2 代码——过滤年报下载链接

```
import re

#去除摘要等其他非年报内容
def filter_words(words,df,include=True):
    ls=[]
    for word in words:
        if include:
            ls.append([word in f for f in df['title']])
        else:
            ls.append([word not in f for f in df['title']])
    index=[]
    for r in range(len(df)):
        flag=not include
        for c in range(len(words)):
            if include:
                flag=flag or ls[c][r]
            else:
                flag=flag and ls[c][r]
        index.append(flag)
    df2=df[index]
    return df2

#只保留 10 年的年报
def filter_date(start,end,df):
    date=df['date']
    v=[d >= start and d<= end for d in date]
    df_new=df[v]
    return df_new

#10 年的起止时间
```

```python
import datetime
def start_end_10y():
    dt_now=datetime.datetime.now()
    current_year=dt_now.year
    start=f'{current_year-9}-01-01'
    end=f'{current_year}-12-31'
    return (start,end)

#整合以上代码
def filter_nb_10y(df,keep_words,exclude_words,start=''):
    if start == '':
        start,end=start_end_10y()
    else:
        start_y=int(start[0:4])
        end=f'{start_y+9}-12-31'
    #
    df=filter_words(keep_words,df,include=True)
    df=filter_words(exclude_words,df,include=False)
    df=filter_date(start,end,df)
    return(df)



#处理更正版的日期
def amend_date(df):
    target_index=[]
    for i, row in df.iterrows():
        if '更正' in row['title']:
            target_index.append(i)
            target=''.join(df.loc[target_index,'title'].values)
            matchobj=re.search('\d{4}年',target)
            amend_year=int(matchobj.group()[:4])
            df.loc[target_index,'date']=f'{amend_year+1}-12-31'
            df.sort_values(by="date" , inplace=True, ascending=False)
    return df


#保留修订版和更正版，删除未修订和未更正版
def retain_rev_edt(df):
    df=df.reset_index(drop=True)  #重置行索引
    index_list = df[df['title'].str.contains('修订|更正')].index.tolist()
    if index_list != []:
        index_list = [i+1 for i in index_list]
        if index_list[-1] > len(df)-1:
```

```
            index_list.pop()
        df=df.drop(index_list,axis=0)
    return df
```

4.3 代码——下载年报

```
import requests
import time

#提出链接、年份数据
def prepare_hrefs_years(df):
    hrefs=df['href'].to_list()
    years=[int(d[0:4])-1 for d in df['date']]
    return((hrefs,years))


# 下载 1 家公司 1 年年报
def download_pdf(href,code,year):
    r = requests.get(href, allow_redirects=True)
    fname=f'{code}_{year}.pdf'
    f = open(fname, 'wb')
    f.write(r.content)
    f.close()
    r.close()

# 下载 1 家公司多年年报
def download_pdfs(hrefs,code,years):
    for i in range(len(hrefs)):
        href=hrefs[i]
        year=years[i]
        download_pdf(href, code, year)
        print(f'Successfully downloaded: {code}_{year}')
        time.sleep(5)
    return()

# 下载多家公司多年年报
def download_pdfs_codes(list_hrefs,codes,list_years):
    for i in range(len(list_hrefs)):
        hrefs=list_hrefs[i]
        years=list_years[i]
        code=codes[i]
        download_pdfs(hrefs, code, years)
        return()
```

4.4 代码——解析年报

4.4.1 代码——解析会计数据

```
import fitz
import re


# 获取指定文本
def get_subtxt(doc,bounds=('主要会计数据和财务指标','总资产')):
    #默认设置为首尾页码
    start_pageno=0
    end_pageno=len(doc)-1
    #
    lb,ub=bounds
    #获取左界页码
    for n in range(len(doc)):
        page=doc[n]
        txt=page.get_text()
        if lb in txt:
            start_pageno=n
            break
    #获取右界页码
    for n in range(start_pageno,len(doc)):
        if ub in doc[n].get_text():
            end_pageno=n
            break
    #获取小范围内字符串
    txt=''
    for n in range(start_pageno,end_pageno+1):
        page=doc[n]
        txt += page.get_text()
    return(txt)


# 获取指定的会计数据值
def get_account_data(account,txt):
    p_txt='%s\s*(-*\d{1,3}(?:,\d{3})*(?:\.\d+)?)' % account   #%s 是占位符，
用‘account’替换，\D 是非数字，\d{1,3}是数字 1 或 2 或 3 个，*可重复，？非贪婪，（）
内是所要的数字，小数点后\d+表示小数点后至少一位数字
    p=re.compile(p_txt)
    matchobj=p.search(txt)
    amt=matchobj.group(1)
    if '.' not in amt:
        p_txt='%s\s*(-*\d{1,3}(?:,\d{3})*(.*))\n' % account
```

```
        p=re.compile(p_txt)
        matchobj=p.search(txt)
        amt=matchobj.group(1)
        #
        s=matchobj.end()
        p_txt1='(.+)\n'
        p1=re.compile(p_txt1)
        matchobj1=p1.search(txt[s:])
        amt += matchobj1.group(1)
    return(amt)



##获取整张会计数据表格

# 获取表头
def get_th_span(txt):
    nianfen='(20\d\d|199\d)\s*年末?'
    s=f'{nianfen}\s*{nianfen}.*?{nianfen}'
    p=re.compile(s,re.DOTALL)
    matchobj=p.search(txt)
    #
    end=matchobj.end()
    year1=matchobj.group(1)
    year2=matchobj.group(2)
    year3=matchobj.group(3)
    #
    flag=(int(year1)-int(year2) == 1) and (int(year2)-int(year3) == 1)
    #
    while (not flag):
        matchobj=p.search(txt[end:])
        end=matchobj.end()
        year1=matchobj.group(1)
        year2=matchobj.group(2)
        year3=matchobj.group(3)
        flag=(int(year1)-int(year2) == 1)
        flag=flag and (int(year2)-int(year3) ==1)
    return(matchobj.span())

#获取表格边界
def get_bounds(txt):
    th_span_1st=get_th_span(txt)
    end=th_span_1st[1]
    th_span_2nd=get_th_span(txt[end:])
```

```
    th_span_2nd=(end+th_span_2nd[0],end+th_span_2nd[1])
    #
    s=th_span_1st[1]
    e=th_span_2nd[0]-1
    #
    while (txt[e] not in '0123456789'):  #如果最后一个不是数字
        e=e-1
    return(s,e+1)

#获取表格内的会计数据关键字
def get_keywords(txt):
    p=re.compile(r'\d+\s*?\n\s*?([\u2E80-\u9FFF]+)')
    keywords=p.findall(txt)
    keywords.insert(0,'营业收入')
    return(keywords)

# 获取整张表格内容
def parse_key_fin_data(subtxt,keywords):
    ss=[]
    s=0
    for kw in keywords:
        n=subtxt.find(kw,s)  #参数 s:从第 s 个位置开始找
        ss.append(n)  #所有 keywords 的起始位置
        s=n+len(kw)
    ss.append(len(subtxt))
    data=[]
    p=re.compile('[^0123456789-]+(?:\s+\D*)?(?:（.*）|\(.*\))?')
    p2=re.compile('\s')
    for n in range(len(ss)-1):
        s=ss[n]
        e=ss[n+1]
        line=subtxt[s:e]
        #获取可能换行的账户名称
        matchobj=p.search(line)
        account_name=p2.sub('',matchobj.group())
        #获取三年数据
        amnts=line[matchobj.end():].split()
        #加上账户名称
        amnts.insert(0,account_name)
        #追加到总数据
        data.append(amnts)
    return data
```

4.4.2 代码——解析公司基本信息

```python
import re

# 提取公司基本情况信息
def get_com_ifm(txt,keywords=['公司办公地址','公司网址','电子信箱']):
    s=txt.find('基本情况简介')
    e=txt.find('信息披露及备置地点',s)
    subtxt=txt[s:e]
    data=[]
    for kw in keywords:
        p=re.compile('%s\s*\n\s*(.+)' % kw)
        matchobj=p.search(subtxt)
        if matchobj:
            ifm=matchobj.group(1)
            if ifm[-1] == ' ':
                ifm=ifm[:-1]
        else:
            ifm='无'
        data.append([kw,ifm])
    return data

# 提取董事会秘书基本信息
def get_cts_ifm(txt,keywords=['姓名','电话','电子信箱']):
    s=txt.find('联系人和联系方式')
    e=txt.find('基本情况简介',s)
    subtxt=txt[s:e]
    data=[]
    for kw in keywords:
        p=re.compile('%s\s*\n\s*(.+)' % kw)
        matchobj=p.search(subtxt)
        if matchobj:
            ifm=matchobj.group(1)
            if ifm[-1] == ' ':
                ifm=ifm[:-1]
        else:
            p=re.compile('%s\s*(.+)' % kw)
            matchobj=p.search(subtxt)
            if matchobj:
                ifm=matchobj.group(1)
                if ifm[-1] == ' ':
                    ifm=ifm[:-1]
            else:
                ifm='无'
        data.append([kw,ifm])
    return data
```

4.5 代码主体

```
import fitz
import pandas as pd
from pylab import plt, mpl
plt.style.use('seaborn')
mpl.rcParams['font.family'] = 'serif'
plt.rcParams['font.family'] = ['sans-serif']
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus']=False

from sse import get_table_sse,parse_data
from filter_url import filter_nb_10y,amend_date,retain_rev_edt
from download import prepare_hrefs_years,download_pdfs
from parse_ar import get_subtxt,get_account_data
from parse_ifm import get_com_ifm,get_cts_ifm


# 下载年报
codes=[600075,600078,600096,600135,600141,600160,600165,600230,600249,60027
3]

for code in codes:
    #获取年报的下载链接
    get_table_sse(code)
    df=parse_data(code)
    df=amend_date(df)   #处理更正版年报的日期
    df=filter_nb_10y(df, keep_words=['年报','年度报告'], exclude_words=['摘要
'])  #去除年报摘要
    df=retain_rev_edt(df)  #保留修订版和更正版，删除未修订和未更正版
    #下载年报
    hrefs=prepare_hrefs_years(df)[0]
    years=prepare_hrefs_years(df)[1]
    download_pdfs(hrefs=hrefs,code=code,years=years)



# 提取营业收入和归属于上市公司股东的净利润
years=[2013,2014,2015,2016,2017,2018,2019,2020,2021,2022]

revenues=pd.DataFrame(index=years,columns=codes)
profits_shlder=pd.DataFrame(index=years,columns=codes)

for code in codes:
```

```
    for year in years:
        filename=f'{code}_{year}.pdf'
        doc=fitz.open(filename)
        if filename != '600078_2021.pdf':
            txt=get_subtxt(doc,bounds=('主要会计数据和财务指标','总资产'))
        else:
            txt=get_subtxt(doc,bounds=('主要 会计数据和财务指标','总资产'))
        if filename != '600075_2013.pdf':
            revenue=get_account_data('\s*'.join('营业收入'), txt)
        else:
            revenue=get_account_data('营业总收入', txt)
        revenues.loc[year,code]=revenue
        if filename != '600273_2019.pdf':
            profit_shlder=get_account_data('\s*'.join('归属于上市公司股东的净
利润'), txt)
        else:
            profit_shlder=get_account_data('\s*'.join('归属于上市公司股东的'),
txt)
        profits_shlder.loc[year,code]=profit_shlder


# 更改列名
col_name=['新疆天业','*ST 澄星','云天化','乐凯胶片','兴发集团',
          '巨化股份','新日恒力','沧州大化','两面针','嘉化能源']

revenues.columns=col_name
profits_shlder.columns=col_name


# 将数据转化为浮点型
for col in col_name:
    revenues[col] = revenues[col].str.replace(',', '').astype(float)
    profits_shlder[col] = profits_shlder[col].str.replace(',',
'').astype(float)



# 10 家公司的营业收入折线图
(revenues/1e9).plot(figsize=(12,7))
plt.xlabel('年份', fontsize=13)
plt.ylabel('营业收入/十亿', fontsize=13)
plt.title('10 家公司近 10 年营业收入折线图',fontsize=14)
# plt.gca().get_yaxis().get_major_formatter().set_scientific(False)
plt.show()
```

```
# 600075 新疆天业近十年营业收入和归属于上市公司股东的净利润折线图
plt.figure(figsize=(12,7))
plt.plot(revenues['新疆天业']/1e9)
plt.plot(profits_shlder['新疆天业']/1e9)
plt.xlabel('年份', fontsize=13)
plt.ylabel('金额/十亿', fontsize=13)
plt.title('新疆天业近 10 年营业收入和归属于上市公司股东的净利润折线图
',fontsize=14)
plt.legend(['营业收入','归属于上市公司股东的净利润'],fontsize=13)
plt.show()



# 提取公司和董事会秘书基本信息
bsc=pd.DataFrame()
for code in codes:
    filename=f'{code}_2022.pdf'
    doc=fitz.open(filename)
    txt=get_subtxt(doc,bounds=('联系人和联系方式','信息披露及备置地点'))
    #
    data1=get_com_ifm(txt)
    bsc.loc[code,'公司办公地址']=data1[0][1]
    bsc.loc[code,'公司网址']=data1[1][1]
    bsc.loc[code,'电子信箱']=data1[2][1]
    #
    data2=get_cts_ifm(txt)
    bsc.loc[code,'董事会秘书姓名']=data2[0][1]
    bsc.loc[code,'董事会秘书电话']=data2[1][1]
    bsc.loc[code,'董事会秘书电子信箱']=data2[2][1]

bsc = bsc.rename_axis("公司代码")
bsc.insert(0, '公司简称', col_name)

bsc.to_csv('公司及董事会秘书基本信息.csv')
```