Python 程序设计实验报告

实验题目：基于 scrapy 框架实现北邮、西电、成电招聘网站的信息爬取与数据处理

姓　　名：刘帅

学　　号：2020212267

日　　期：2022 年 1 月 1 日

1、需求分析

本次作业是基于 scrapy 框架实现北邮、西电、成电招聘网站招聘信息的爬取和信息处理的任务，因此，首先可将任务分为爬虫和信息处理两个子任务。针对爬虫部分，即 spider 文件的建立，则首先需要确定网页类型，html 结构等。通过分析，本实验对于北邮、西电、成电均依赖 selenium API 实现动态爬取，进而确定爬取过程中最重要的两个流程：1、请求信息的爬取并由 scrapy 向网页发起 request 请求 2、锁定目标数据并添加至本地（items）中。针对信息处理部分，我们依赖 csv 和 openpyxl 库，将数据存储至 csv，并经过数据处理将其存入至 excel 文件，进而实现数据的筛选。

2、程序框架

2.1 爬虫文件部分

以北邮爬虫为例（西电、成电框架类似）

```
class mySpider(scrapy.spiders.Spider):
    name = "BUPT"
    allowed_domains = ["job.bupt.edu.cn"]
    start_urls=['https://job.bupt.edu.cn/frontpage/bupt/html/recruitmentinfoList.html?type=1']
    def start_requests(self):
        driver=webdriver.Chrome()
        driver.get(self.start_urls[0])
        for j in range(85):
            page=driver.page_source
            parse=etree.HTML(page)
            urllist=[]
            for i in range(1,16):
                urllist.append(parse.xpath(f'//*[@id="listPlace"]/div[{i}]/div[2]/a/@href'))
            if(j<=1):
                button = driver.find_element(By.XPATH,'/html/body/div[1]/div[4]/div[2]/ul/li[8]/a')
            elif(j==2):
                button = driver.find_element(By.XPATH,'/html/body/div[1]/div[4]/div[2]/ul/li[9]/a')
            elif(j>2):
                button =driver.find_element(By.XPATH,'/html/body/div[1]/div[4]/div[2]/ul/li[10]/a/em')
            ActionChains(driver).move_to_element(button).click(button).perform()

            time.sleep(1)
            print(urllist)
            for per in urllist:
                yield scrapy.Request(per[0])
```

通过 lxml 的 etree 类，将网页的 html 以文本形式显示，提取每个项目的超链接，并利用 Actionchain 实现“下一页的跳转。”

```
def parse(self, response):
    # name = response.xpath('/html/body/div[1]/div[2]/div[2]/div[1]/div[1]/text()').extract()
    # info = response.xpath('/html/body/div[1]/div[3]/div/text()').extract()
    # print(len(info))
    # date = info[0].split('\xa0', -1)[1][4:14]
    # hot = info[0].split('\xa0', -1)[2][7:-4]
    theme=response.xpath('/html/body/div[1]/div[2]/div[2]/div[1]/div[1]/text()').extract()[0]
    date=response.xpath('/html/body/div[1]/div[3]/div/text()').extract()[0].split('浏')[0].rstrip()[-10:]
    views=response.xpath('/html/body/div[1]/div[3]/div/text()').extract()[0].split(': ')[3].rstrip()
    quantity = response.xpath('/html/body/div[1]/div[4]/div/div[1]/div[2]/table/tbody/tr[2]/td[3]/text()').extract()
    item = BUPTItem()
    try:
        if len(quantity)==0:
            item['quantity']=1
        else:
            item['quantity']=quantity[0]
    except IndexError:
        item['quantity'] = 1
    if datetime.datetime.strptime(date, '%Y-%m-%d') > datetime.datetime.strptime('2021-9-1', '%Y-%m-%d'):
        item['title'] = theme
        item["date"] = date
        item['views'] = views
        yield item
```

Parse 函数则是对于每个网页进行特征提取，将招聘主题、招聘日期、浏览次数（北邮多一个职位需求）提取并存入 item 中。

2.2pipeline 部分

```
import csv
class ScrapyJob10Pipeline:#定义本项目的pipeline处理类
    def process_item(self, item, spider):#对于爬取到的数据进行处理
        try:
            dict_item = dict(item)  #把抓取数据生成列表对象
            if spider.name=='XIDIAN_JIUYE_ALL':
                self.XIDIAN_writer.writerow(dict_item)  #将数据写入到文件中
            elif spider.name=='CHENGDIAN':
                self.BEIDA_writer.writerow(dict_item)  # 将数据写入到文件中
            elif spider.name == 'BUPT':
                self.BUPT_writer.writerow(dict_item)  # 将数据写入到文件中
            return item
        except Exception as err:
            print(err)
    def open_spider(self, spider):
        #爬虫开启时候的动作，写模式打开json文件，并设置为爬虫类的
        self.XIDIAN_file =open('XIDIAN_1.csv', 'w+', newline='', encoding='utf-8')
        self.XIDIAN_writer = csv.DictWriter(self.XIDIAN_file,fieldnames=['title','date','views'])
        self.XIDIAN_writer.writeheader()#写入表头
        self.BEIDA_file = open('CHENGDIAN_1.csv', 'w+', newline='', encoding='utf-8')
        self.BEIDA_writer = csv.DictWriter(self.BEIDA_file, fieldnames=['title', 'date', 'views'])
        self.BEIDA_writer.writeheader()  # 写入表头
        self.BUPT_file = open('BEIYOU_1.csv', 'w+', newline='', encoding='utf-8')
        self.BUPT_writer = csv.DictWriter(self.BUPT_file, fieldnames=['title', 'date', 'views','quantity'])
        self.BUPT_writer.writeheader()  # 写入表头
ScrapyJob10Pipeline › process_item() › try › elif spider.name == 'BUPT'
```

其主要功能是将内存数据读入到 csv 并存储，为后续数据操作进行准备。

2.3middlewares 部分

```
def process_request(self, request, spider):
    if spider.name == "XIDIAN_JIUYE_ALL":
        self.XIDIAN_driver.get(request.url)  ###用driver打开该页面
        time.sleep(1)  ###等候几秒钟比较保险
        return scrapy.http.HtmlResponse(url=request.url, body=self.XIDIAN_driver.page_source.encode('utf-8'),
                                        encoding='utf-8', request=request, status=200)
    elif spider.name == "CHENGDIAN":
        self.BEIDA_driver.get(request.url)  ###用driver打开该页面
        time.sleep(1)  ###等候几秒钟比较保险
        return scrapy.http.HtmlResponse(url=request.url, body=self.BEIDA_driver.page_source.encode('utf-8'),
                                        encoding='utf-8', request=request, status=200)
    elif spider.name == "BUPT":
        self.BUPT_driver.get(request.url)  ###用driver打开该页面
        time.sleep(1)  ###等候几秒钟比较保险
        return scrapy.http.HtmlResponse(url=request.url, body=self.BEIDA_driver.page_source.encode('utf-8'),
                                        encoding='utf-8', request=request, status=200)
# def process_response(self, request, response, spider):
#     # Called with the response returned from the downloader.
#
```

其中，request 函数较为重要，spider 文件将请求信息提交至 process_request，利用 scrapy 框架对每个请求进行打开，进而在 spider 文件中实现信息的读取。

2.4begin 部分

```
from scrapy import cmdline
cmdline.execute("scrapy muti_crawl XIDIAN_JIUYE_ALL".split())
cmdline.execute("scrapy muti_crawl CHENGDIAN".split())
cmdline.execute("scrapy muti_crawl BUPT".split())
```

利用补充模块 muti_crawl 进行多线程爬取，上图为程序启动指令。

3、模型分析

当 scrapy 工程成功的爬取到网页信息并将其存储至 csv 后，下一步则需要对于爬取到的数据进行数据处理，也分为如下两个子环节：1、csv 文件的格式化 2、将 csv 文件数据进行分析并存入 excel。

3.1 csv 文件格式化

```python
import csv
import pandas as pd
data=pd.read_csv("/SCRAPY_JOB_10/BEIYOU_1.csv")
newtitle=data['title'].str.strip()
newdate=data['date'].str.strip()
data['title']=newtitle
data['date']=newdate

data=pd.read_csv("/SCRAPY_JOB_10/CHENGDIAN_1.csv")
newtitle=data['title'].str.strip()
newdate=data['date'].str.strip()
data['title']=newtitle
data['date']=newdate

data=pd.read_csv("/SCRAPY_JOB_10/XIDIAN_1.csv")
newtitle=data['title'].str.strip()
newdate=data['date'].str.strip()
data['title']=newtitle
data['date']=newdate
```

利用 pandas 去除 csv 文件的空格

3.2 excel 数据处理

3.2.1 雇主类型定义（图片请放大查看）

```python
def classify(name):
    if '大学' in name or '学院' in name or '校' in name or '实验室' in name or '博士' in name or '研究院' in name:
        return '教育企业'
    elif '科技' in name or '研发' in name or '信息' in name or '网络' in name or '计算机' in name or '软件' in name or '数据' in name or '智能' in name:
        return '互联网企业'
    elif '律师' in name or '法' in name or '政' in name:
        return '法律企业'
    elif '通信' in name or '通讯' in name:
        return '通信企业'
    elif '医' in name or '药' in name:
        return '医疗企业'
    elif '芯' in name or '电子' in name or '半导体' in name:
        return '电子相关企业'
    elif '移动' in name or '联通' in name:
        return '电信运营商'
    elif '技术' in name:
        return '互联网企业'
    elif '环' in name:
        return '环保相关企业'
    elif '工业' in name or '械' in name or '工程' in name or '机' in name or '天' in name or '电' in name or '厂' in name or '测' in name or '自动' in name or '控制' in name:
        return '工业、航天企业'
    elif '车' in name or '汽' in name:
        return '汽车企业'
    elif '会计' in name or '工商' in name or '管理' in name:
        return '工商管理企业'
    elif '银行' in name or '证券' in name or '金' in name or '财' in name or '信' in name or '期货' in name or '股' in name or '投资' in name or '经济' in name:
        return '金融企业'
    elif '建' in name:
        return '土木、建设类企业'
    else:
        return '其他企业'
```

说明：通过观察招聘主题及信息，我们首先人为标注为教育企业、互联网企业、法律企业、通信企业、医疗企业、电子相关企业、电信运营商、互联网企业、环保相关企业、工业、航天企业、汽车企业、工商管理企业、金融企业、土木、建设类企业、其他企业。定义 classify 函数，对于关键字进行提取，例如出现“学院”“教”“实验室”等信息则判定为“教育企业”

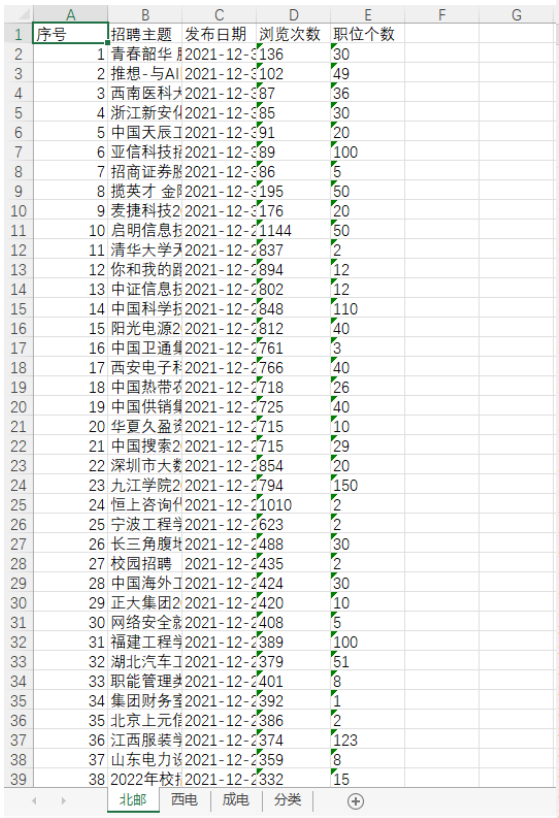
3.2.2 数据排序处理

```
set(sheet)
xidianfile=open(path+"XIDIAN_1.csv",'r',encoding='utf-8')
xidianreader=list(csv.reader(xidianfile))
titlelist=[]
xidianlist=[]
xidiandict={}
finaldict={}
for i in range(1,len(xidianreader)):
    sheet.cell(i+1,1).value=i
    sheet.cell(i+1,2).value=xidianreader[i][0]
    if sheet.cell(i+1,2).value not in titlelist:
        titlelist.append(sheet.cell(i+1,2).value)
        xidianlist.append(sheet.cell(i+1,2).value)#去重并存储西电的title
    sheet.cell(i+1,3).value=xidianreader[i][1]
    sheet.cell(i+1,4).value=xidianreader[i][2]
    xidiandict[xidianreader[i][0]]=int(xidianreader[i][2])
    finaldict[xidianreader[i][0]]=int(xidianreader[i][2])
processed_xidiandict=sorted(xidiandict.items(),key= lambda x:x[1],reverse=True)
for i in xidianlist:
    xidianemployer[classify(i)]+=1
sheet=excel['成电']
set(sheet)
```

将 csv 数据以字典和列表等方式存储并排序，将文件写入 excel。

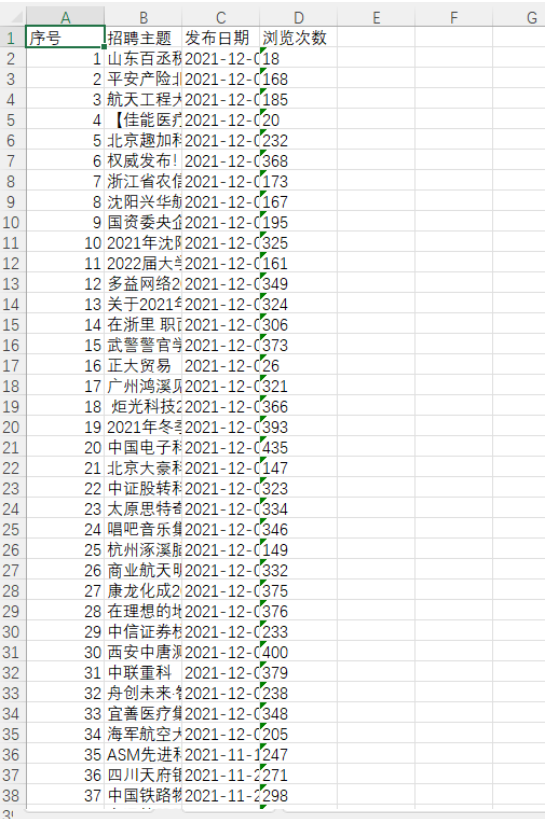
4、数据分析

4.1 预处理完成后的数据

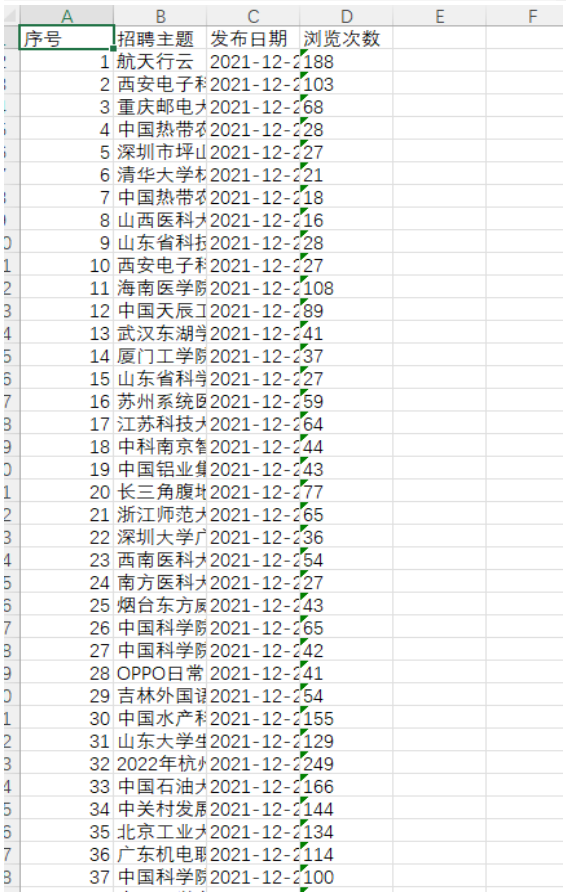


序号	招聘主题	发布日期	浏览次数	职位个数
1	青春韶华	2021-12-3	136	30
2	推想-与AI	2021-12-3	102	49
3	西南医科大	2021-12-3	87	36
4	浙江新安化	2021-12-3	85	30
5	中国天辰工	2021-12-3	91	20
6	亚信科技招	2021-12-3	89	100
7	招商证券股	2021-12-3	86	5
9	揽英才 金	2021-12-3	195	50
9	麦捷科技2	2021-12-3	176	20
10	启明信息招	2021-12-2	1144	50
11	清华大学大	2021-12-2	837	2
12	你和我的青	2021-12-2	894	12
13	中证信息招	2021-12-2	802	12
14	中国科学招	2021-12-2	848	110
15	阳光电源2	2021-12-2	812	40
16	中国卫通集	2021-12-2	761	3
17	西安电子科	2021-12-2	766	40
18	中国热带农	2021-12-2	718	26
19	中国供销集	2021-12-2	725	40
20	华夏久盈资	2021-12-2	715	10
21	中国搜索2	2021-12-2	715	29
22	深圳市大疆	2021-12-2	854	20
23	九江学院2	2021-12-2	794	150
24	恒上咨询代	2021-12-2	1010	2
25	宁波工程学	2021-12-2	623	2
26	长三角腹地	2021-12-2	488	30
27	校园招聘	2021-12-2	435	2
28	中国海外工	2021-12-2	424	30
29	正大集团2	2021-12-2	420	10
30	网络安全系	2021-12-2	408	5
31	福建工程学	2021-12-2	389	100
32	湖北汽车工	2021-12-2	379	51
33	职能管理集	2021-12-2	401	8
34	集团财务有	2021-12-2	392	1
35	北京上元信	2021-12-2	386	2
36	江西服装学	2021-12-2	374	123
37	山东电力设	2021-12-2	359	8
38	2022年校招	2021-12-2	332	15

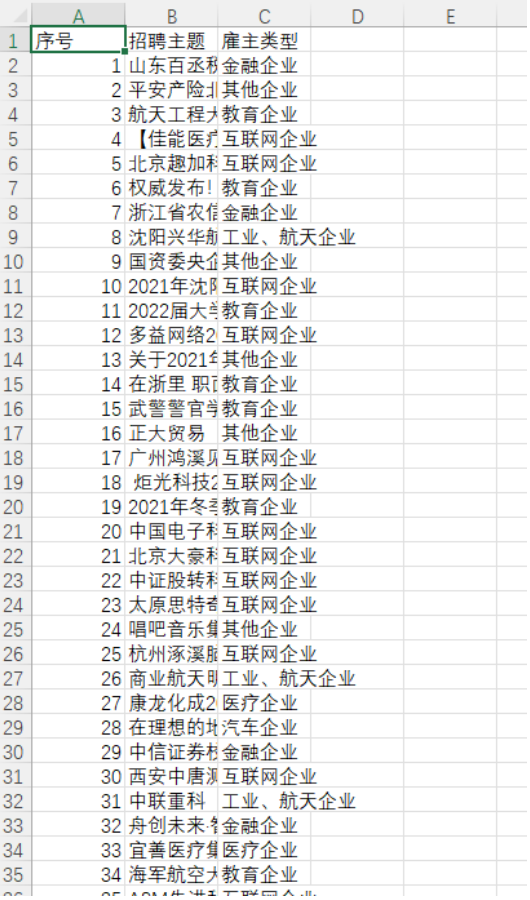
北部 | 西电 | 成电 | 分类



序号	招聘主题	发布日期	浏览次数
1	山东百丞科	2021-12-0	18
2	平安产险北	2021-12-0	168
3	航天工程大	2021-12-0	185
4	【佳能医疗	2021-12-0	20
5	北京趣加科	2021-12-0	232
6	权威发布！	2021-12-0	368
7	浙江省农信	2021-12-0	173
8	沈阳兴华航	2021-12-0	167
9	国资委央企	2021-12-0	195
10	2021年沈阳	2021-12-0	325
11	2022届大学	2021-12-0	161
12	多益网络2	2021-12-0	349
13	关于2021年	2021-12-0	324
14	在浙里 职[	2021-12-0	306
15	武警警官学	2021-12-0	373
16	正大贸易	2021-12-0	26
17	广州鸿溪贝	2021-12-0	321
18	炬光科技2	2021-12-0	366
19	2021年冬季	2021-12-0	393
20	中国电子科	2021-12-0	435
21	北京大豪科	2021-12-0	147
22	中证股转科	2021-12-0	323
23	太原思特奇	2021-12-0	334
24	唱吧音乐集	2021-12-0	346
25	杭州涿溪脑	2021-12-0	149
26	商业航天助	2021-12-0	332
27	康龙化成2	2021-12-0	375
28	在理想的地	2021-12-0	376
29	中信证券招	2021-12-0	233
30	西安中唐源	2021-12-0	400
31	中联重科	2021-12-0	379
32	舟创未来智	2021-12-0	238
33	宜善医疗集	2021-12-0	348
34	海军航空大	2021-12-0	205
35	ASM先进科	2021-11-1	247
36	四川天府银	2021-11-2	271
37	中国铁路物	2021-11-2	298



序号	招聘主题	发布日期	浏览次数
1	航天行云	2021-12-2	188
2	西安电子科	2021-12-2	103
3	重庆邮电大	2021-12-2	68
4	中国热带农	2021-12-2	28
5	深圳市坪山	2021-12-2	27
6	清华大学校	2021-12-2	21
7	中国热带农	2021-12-2	18
8	山西医科大	2021-12-2	16
9	山东省科技	2021-12-2	28
10	西安电子科	2021-12-2	27
11	海南医学院	2021-12-2	108
12	中国天辰工	2021-12-2	89
13	武汉东湖学	2021-12-2	41
14	厦门工学院	2021-12-2	37
15	山东省科学	2021-12-2	27
16	苏州系统医	2021-12-2	59
17	江苏科技大	2021-12-2	64
18	中科南京智	2021-12-2	44
19	中国铝业集	2021-12-2	43
20	长三角腹地	2021-12-2	77
21	浙江师范大	2021-12-2	65
22	深圳大学广	2021-12-2	36
23	西南医科大	2021-12-2	54
24	南方医科大	2021-12-2	27
25	烟台东方质	2021-12-2	43
26	中国科学院	2021-12-2	65
27	中国科学院	2021-12-2	42
28	OPPO日常	2021-12-2	41
29	吉林外国语	2021-12-2	54
30	中国水产科	2021-12-2	155
31	山东大学生	2021-12-2	129
32	2022年杭州	2021-12-2	249
33	中国石油大	2021-12-2	166
34	中关村发展	2021-12-2	144
35	北京工业大	2021-12-2	134
36	广东机电职	2021-12-2	114
37	中国科学院	2021-12-2	100

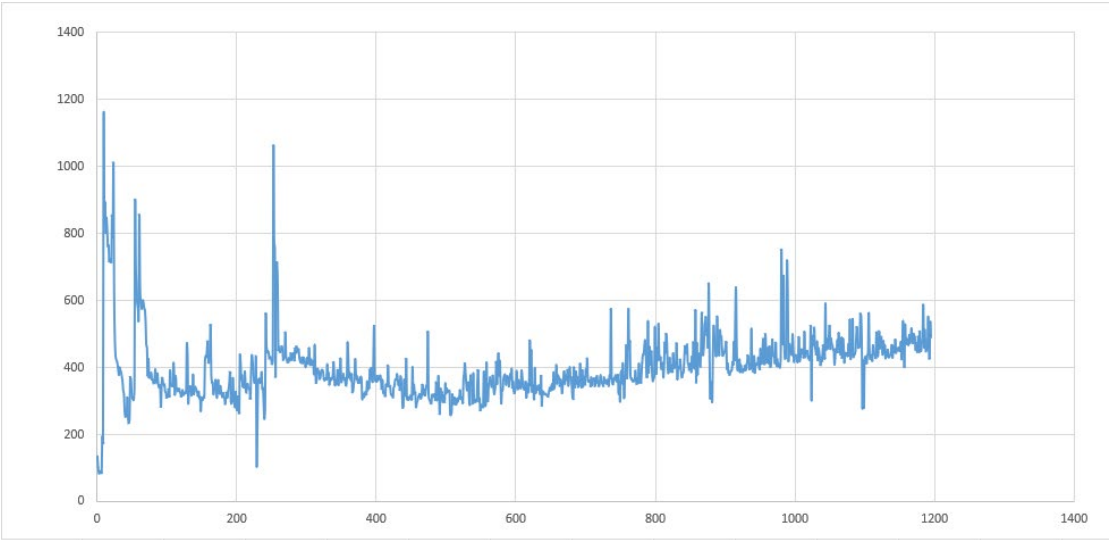


序号	招聘主题	雇主类型
1	山东百丞科	金融企业
2	平安产险北	其他企业
3	航天工程大	教育企业
4	【佳能医疗	互联网企业
5	北京趣加科	互联网企业
6	权威发布！	教育企业
7	浙江省农信	金融企业
8	沈阳兴华航	工业、航天企业
9	国资委央企	其他企业
10	2021年沈阳	互联网企业
11	2022届大学	教育企业
12	多益网络2	互联网企业
13	关于2021年	其他企业
14	在浙里 职[	教育企业
15	武警警官学	教育企业
16	正大贸易	其他企业
17	广州鸿溪贝	互联网企业
18	炬光科技2	互联网企业
19	2021年冬季	教育企业
20	中国电子科	互联网企业
21	北京大豪科	互联网企业
22	中证股转科	互联网企业
23	太原思特奇	互联网企业
24	唱吧音乐集	其他企业
25	杭州涿溪脑	互联网企业
26	商业航天助	工业、航天企业
27	康龙化成2	医疗企业
28	在理想的地	汽车企业
29	中信证券招	金融企业
30	西安中唐源	互联网企业
31	中联重科	工业、航天企业
32	舟创未来智	金融企业
33	宜善医疗集	医疗企业
34	海军航空大	教育企业
35	ASM先进科	互联网企业

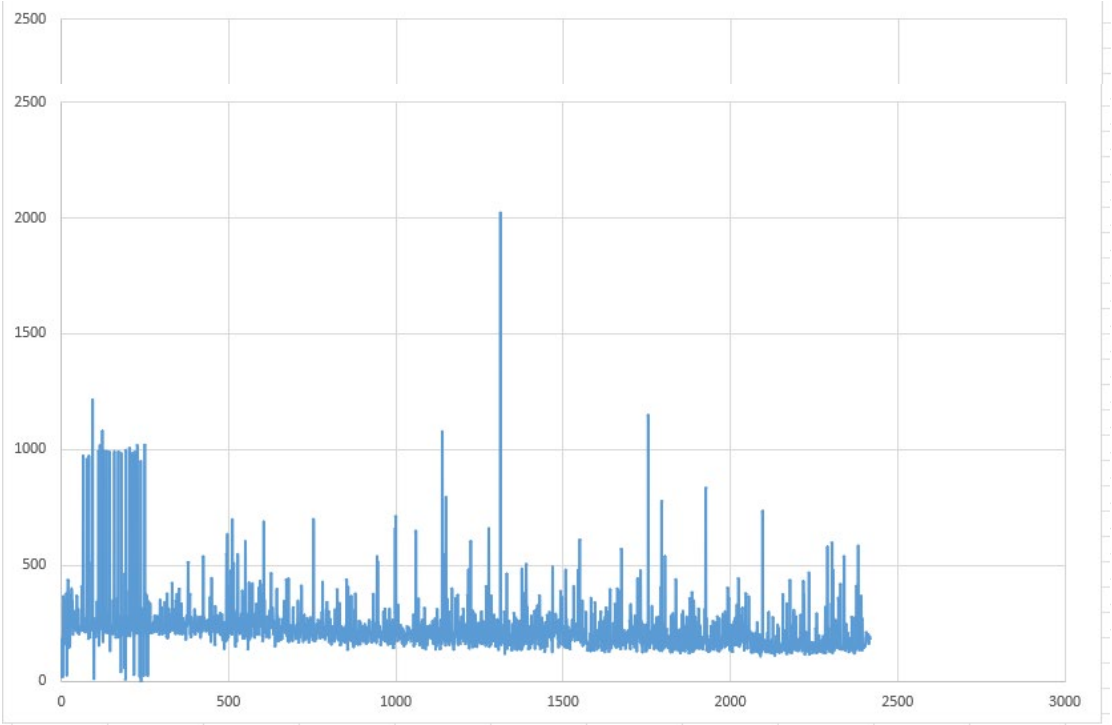
4.2 经数据分析后的数据信息

	最受北邮学生关注的招聘TOP20	最受西电学生关注的招聘TOP20	最受成电学生关注的招聘TOP20
1	中联重科（工业互联网板块）2022届校园招聘简章	相约广电 逐梦前行 陕西广电网络传媒(集团)股份有限公	电科9所2021-2022年校园招聘
2	恒上咨询代宽带资本旗下云天使基金投资经理	西安市事业单位进校园公开招聘2022届高校毕业生公告	中国科学院光电技术研究所2022年校园招聘
3	你和我的距离只有一封简历的厚度	中国航空工业集团公司西安飞机设计研究所2022招聘简	四川传媒学院招聘公告
4	中国电子科技集团公司第五十研究所【2022届校招】	中国电子科技集团公司第二十七研究所2022年度招聘简	中电十所 2022届秋季提前批校园招聘简章
5	北京工业大学先进电池材料与器件研究所（尉海军教授	西安交通工程学院招聘公告	中国航天科技集团第五研究院总体设计部2022年校园招聘
6	深圳市大数据研究院科研人员招聘简章	2021年冬季全国博士、硕士研究生巡回 签约洽谈会——	中国兵器工业集团2022校园招聘全面启动
7	中国科学技术大学先进技术研究院2021-2022年度招聘	潍坊科技学院	中国电子科技集团公司第三十四研究所（军工单位、事业编制）
8	清华大学天津高端装备研究院人才引进	上汽通用东岳汽车有限公司	中国电子科技集团公司第五十一研究所2022届校园招聘简章
9	阳光电源2022届全球校园招聘秋招补录	5G速度 智汇广电 江苏有线2022校园招聘	中电十所 2022届秋季校园招聘简章
10	中证信息技术服务有限责任公司 2021年度招聘公告	天音通信有限公司	中国航空工业集团成都凯天电子股份有限公司 2022届校园招聘
11	九江学院2022年人才招聘公告	全国中小企业股份转让系统有限责任公司 2021年下半年	中国航空制造技术研究院2022届招聘简章
12	西安电子科技大学空间科学与技术学院2021年教师岗位	重庆水利电力职业技术学院2021年公开招聘事业单位工	中铁第一勘察设计院集团有限公司2022年电子科技大学招聘信息
13	中国卫通集团股份有限公司招聘	商业航天明星企业【零壹空间】校招答疑会来啦！！	中国飞行试验研究院2022届校园招聘
14	中信证券北京分公司管培生-校招	承德石油高等专科学校人才引进	中国航天科工集团第四研究院十七所2022届毕业生招聘
15	北京北航天宇长鹰无人机科技有限公司	中国联合工程有限公司2022届校园招聘 大型央企筑梦	中国船舶集团第七〇九研究所2022年校园招聘
16	中国供销集团2022年度应届高校毕业生招聘公告	山东电力设备有限公司2022届高校毕业生招聘简章	郑州大学2021年面向电子科技大学校园专场招聘辅导员（硕士）公告
17	中国热带农业科学院热带作物品种资源研究所2022年普	关于2022年度湖南省国资委"英培计划"人才选拔公告	中国电子科技集团公司第二十二研究所 2022年度校园招聘
18	电子科技大学成都学院教学资源部数据工程师招聘公告	上海电子信息职业技术学院2021年度招聘公告（第一批	中国电子科技集团公司第十三研究所 2022届校园招聘
19	华夏久盈资产管理有限责任公司2022校园招聘	青岛鼎信通讯股份有限公司22届秋季重点岗位补录	民航机场成都电子工程设计有限责任公司2022年校园招聘
20	中国搜索2022年校园招聘简章	赓续百年初心 担当青人使命-大庆师范学院2021年下半	中国电子科技集团航空电子公司 2022届"航U星"招募
	最受北邮学生关注的雇主类型TOP10	最受西电学生关注的雇主类型TOP10	最受成电学生关注的雇主类型TOP10
1	教育企业	互联网企业	教育企业
2	互联网企业	其他企业	互联网企业
3	其他企业	工业、航天企业	其他企业
4	金融企业	金融企业	工业、航天企业
5	工业、航天企业	教育企业	金融企业
6	电子相关企业	电子相关企业	电子相关企业
7	工商管理企业	汽车企业	通信企业
8	法律企业	通信企业	医疗企业
9	医疗企业	电信运营商	工商管理企业
10	通信企业	医疗企业	法律企业
	北邮单个公司招聘职位数TOP10	北邮招聘职位总数及对应雇主类型TOP10	职位总数
1	理想汽车2022届校园招聘秋招补录进行中	汽车企业	1600
2	未来已来 只等你来 ——海南自由贸易港招才引智活动3	教育企业	1000
3	大连人才服务中心 "学子来连、业成滨城"2021-2022 空	其他企业	1000
4	景贤礼士 WE梦相聚 ——北京市石景山区"走进高校引才	其他企业	1000
5	小鹏汽车2022校园招聘"探索者计划"	汽车企业	1000
6	第三届"科技绵州 人才飞阳"四川省绵阳市招才引智 "云	互联网企业	1000
7	中国农业银行研发中心2022年秋季校园招聘	互联网企业	650
8	三环集团2022届秋季校园招聘火热进行中	环保相关企业	500
9	一汽解放汽车有限公司2022年秋季校园招聘启事	汽车企业	500
10	2022年春季全国博士、博士后高级人才专场巡回签约洽	教育企业	500
	最关注ICT行业的招聘主题TOP10	浏览次数	
1	相约广电 逐梦前行 陕西广电网络传媒(集团)股份有限公	2026	
2	西安市事业单位进校园公开招聘2022届高校毕业生公告	1219	
3	中国航空工业集团公司西安飞机设计研究所2022招聘简	1152	
4	电科9所2021-2022年校园招聘	1107	
5	中国电子科技集团公司第二十七研究所2022年度招聘简	1081	
6	中联重科（工业互联网板块）2022届校园招聘简章	1056	
7	西安交通工程学院招聘公告	1023	
8	2021年冬季全国博士、硕士研究生巡回 签约洽谈会——	1021	
9	潍坊科技学院	1021	
10	恒上咨询代宽带资本旗下云天使基金投资经理	1010	

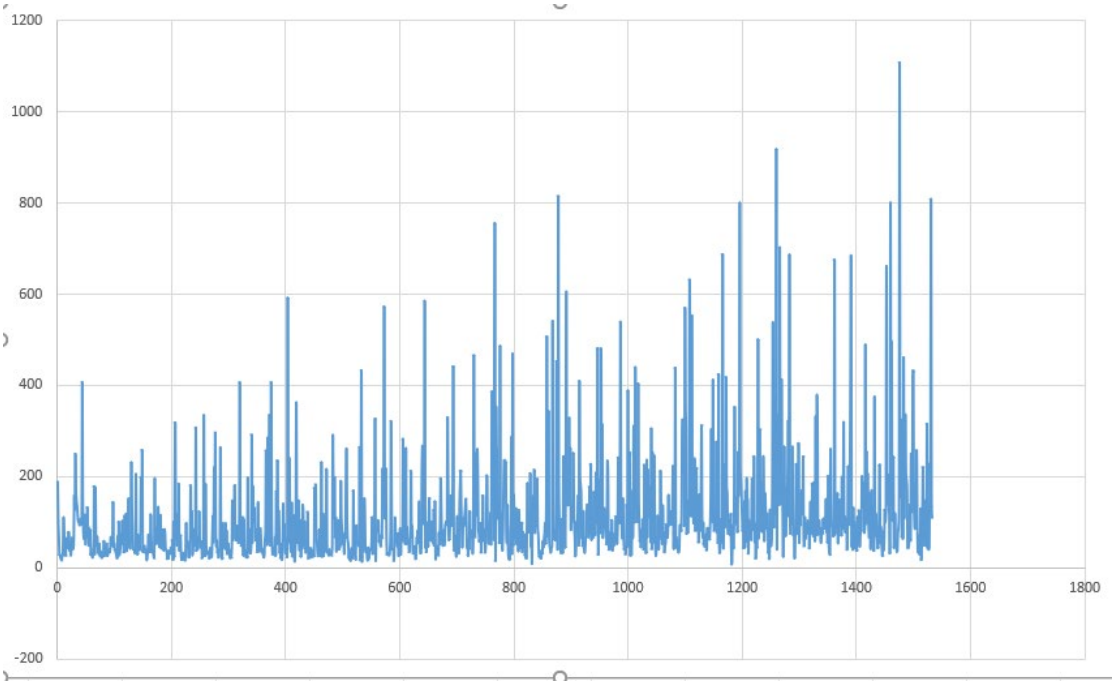
北邮数据可视化（横坐标为标号，纵坐标为浏览次数）



西电数据可视化（横坐标为标号，纵坐标为浏览次数）



成电数据可视化（横坐标为标号，纵坐标为浏览次数）



4.3 结论

可以看出，互联网行业更加青睐西电学生，而教育行业更加青睐北邮和成电的学生，对于“两电一邮”学校，教育行业和互联网行业对于三所学校学生的需求最大，医疗、法律方面对于学校的需求较少。作为人工智能学院的学生，我认为我

们应该更加关心智能医疗领域的教育建设和发展，将医学与科技相结合，同时，建议学校对于人文学院学生加强培养，致力于培养互联网法律相关人才。针对数据分析部分，我们可以看到北邮的最高浏览次数达到 1100 多次，而浏览次数的平均数在 400 次浏览左右；西电的最高浏览次数高达 2000 次，方差较小，数据的平均浏览次数在 250 次左右；成电的最高浏览次数为 1500 多次，浏览次数参差不齐，方差较大，平均浏览次数在 200 次左右。经推断，我认为成电的招生网站因为是研究生网站，同学们对于专业的认知更加清晰，因此对于职位需求的点击量更为有的放矢，个性化更强；而北邮和西电的招聘网站均为本科生招聘网，首先有大量的用户，其次用户的水平残次不齐，招聘信息更加泛化，所以浏览量更为平均。

4、作业小结

本次作业让我了解了 scrapy 爬虫框架，对前端的相关语言的了解程度加深，对于 xpath、css、html 均有了一定程度的认识，有助于后续搭建网页的学习。同时，本次大作业更加考核了我的综合能力，使我对 python 这门语言的掌握应用程度、debug 过程和能力均有了提升，更为后续人工智能有关专业课程的学习打下了夯实基础。感谢老师本学期的辛苦付出，本学期 python 程序设计的学习使我收获颇丰。