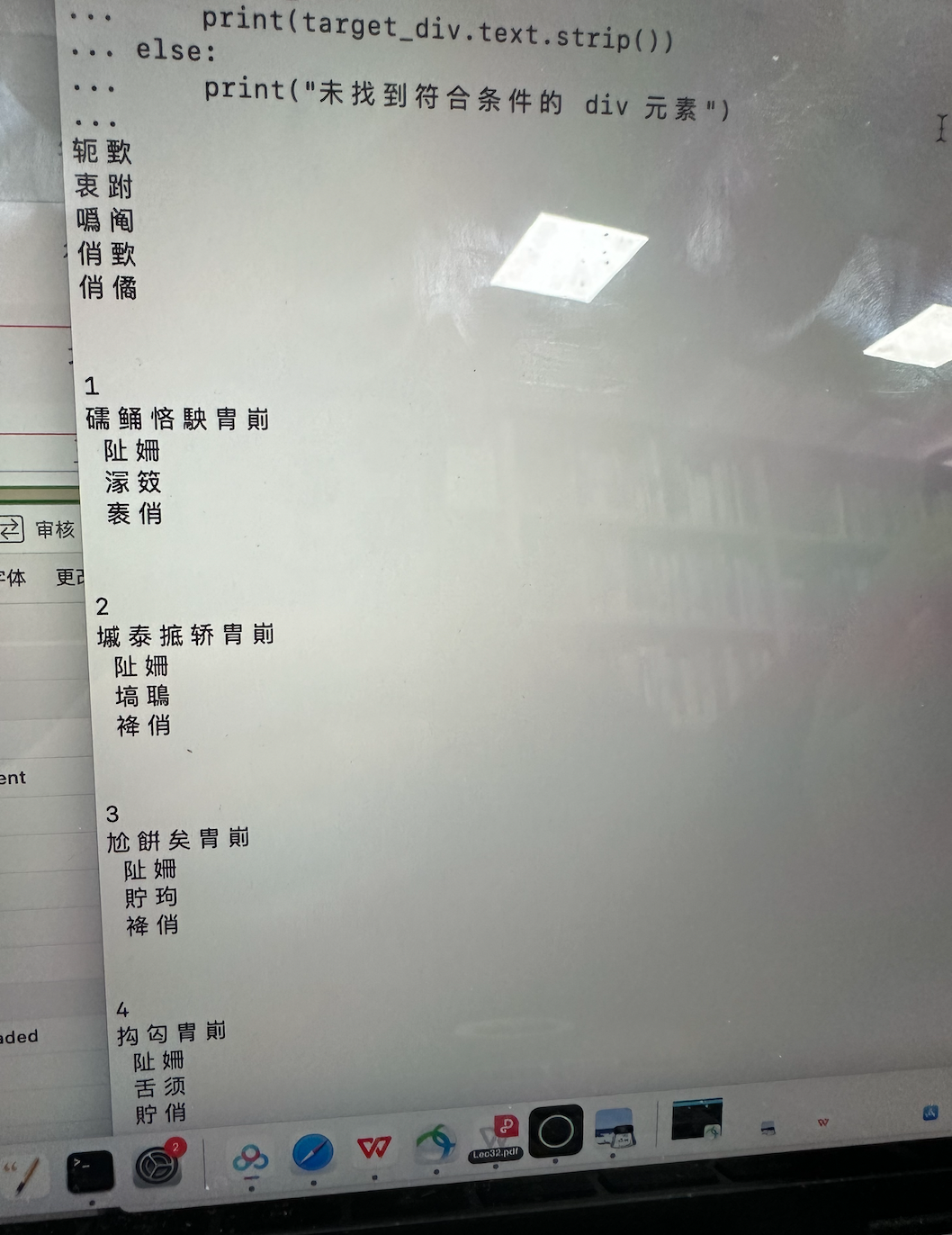
实验心得

在选定“酷狗音乐TOP500榜单”这一题目到做好全部数据为止，时间十分匆忙，完成程度也不甚满意。因为之前的选题是做“汉魏六代石刻”的数据爬取和整理，但一个月都没有完成爬虫代码的问题，想在这里记录一下第一个选题里遇到的困难与吸取的教训，以期后进。

由于代码都没有保存，只存留了当时问同学的聊天记录里拍的照片。

在爬取数据时，遇到了输出乱码的问题。



在解决这个问题时，想了三个方向

第一个是定位问题，定位问题确实出现了问题，没有展开到网页表格最后的定位。

解决后还是乱码。

第二个是编码问题，怀疑是爬虫输出的编码和网页编码不一致。于是又增添了手动指定编码的语句。

content\_type = response.headers.get('Content-Type')

encoding = content\_type.split('charset=')[-1] if 'charset=' in content\_type else 'utf-8'

第三个是登陆问题，由于网站不注册登陆账号申请查看数据库时，只能看到“\*\*\*”看不到具体内容。所以认为增加了如下的语句来解决。

from selenium import webdriver

from selenium.webdriver.chrome.service import Service as ChromeService

from selenium.webdriver.common.by import By

from selenium.webdriver.common.keys import Keys

login\_url = 'https://www.ancientbooks.cn/signin?blackUrl=/c/portal/logout?appKey=RBMQnPl9eZgsVTO8uqFVaBLxA8Exuf0O&backUrl=%2FdocShike%2FshikeSublib.jspx%3FlibId%3D4%26nowDynasty%3D%E8%A5%BF%E6%BC%A2&\_start\_time=1705473253464&base=%2FdocShike&resSys=%2FdocShike%2Fr%2Fcms&res=docShike%2Fr%2Fcms%2Fwww%2Fshike&location=http%3A%2F%2F192.168.10.125%3A8087%2FdocShike%2FssoLogout.jspx%3FreturnUrl%3D%2FdocShike%2FshikeSublib.jspx%3FlibId%3D4%26nowDynasty%3D%E8%A5%BF%E6%BC%A2&rel\_location=%2FssoLogout.jspx%3FreturnUrl%3D%2FshikeSublib.jspx%3FlibId%3D4%26nowDynasty%3D%E8%A5%BF%E6%BC%A2&ssoRedirectUrl=%252FdocShike%252FssoLogout.jspx%253FreturnUrl%253D%252FdocShike%252FshikeSublib.jspx%253FlibId%253D4%2526nowDynasty%253D%25E8%A5%25BF%25E6%25BC%25A2&allBase=http%3A%2F%2F192.168.10.125%3A8087%2FdocShike&path=%2FssoLogout.jspx%3FreturnUrl%3D%2FdocShike%2FshikeSublib.jspx%3FlibId%3D4%26nowDynasty%3D%E8%A5%BF%E6%BC%A2&hasPerm=false&permission=false&outView=true&permit=false&goodsKey=zkcpb-sk&innerURL=http%3A%2F%2F192.168.10.125%2FdocShike%2F&catalogDomain=catalog.ancientbooks.cn&productDomain=inscription.ancientbooks.cn&jhDomain=www.ancientbooks.cn%2F&cidianUrl=https%3A%2F%2Fdictionary.ancientbooks.cn&locale=zh\_CN'

target\_url = 'https://inscription.ancientbooks.cn/docShike/shikeSublib.jspx?libId=4&nowDynasty=%E8%A5%BF%E6%BC%A2'

username = '13933881257'

password = 'Meng020516'

chrome\_path = '/path/to/chromedriver'

chrome\_service = ChromeService(executable\_path=chrome\_path)

driver = webdriver.Chrome(service=chrome\_service)

driver.get(login\_url)

username\_field = driver.find\_element(By.ID, 'login-name')

password\_field = driver.find\_element(By.ID, 'login-password')

username\_field.send\_keys(username)

password\_field.send\_keys(password)

password\_field.send\_keys(Keys.RETURN)

这三个方向都没有解决问题，于是我又去再问，最后确定是字体加密问题，需要找到对应的css映射解密，查阅相关资料和更改代码，都没有解决这个问题。