## 设计目的

正直高考之际，广大考生到了人生的转折点——志愿填报。考生们现在比较需要了解的都是全国全部高校的信息。本系统就是为了解决考生们这方面的需求而开发。

## 系统的需求及其基本功能

爬取高考网全部的高校信息，提取高校的名称、所在地、院校特色、高校类型、高校隶属、高校性质、学校网址等信息，帮助考生更好的去了解各高校，有选择性的去填报心仪的大学。本软件生成高校列表url，爬取全部高校信息，提取数据，保存到本地。

## 总体设计方案

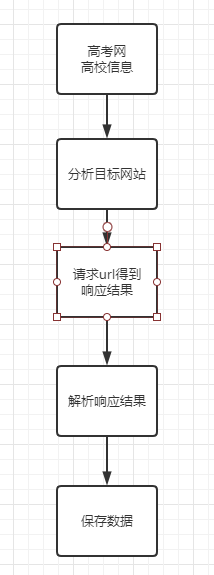
### 分析目标网站

### 生成网站请求链接队列

### 提取每条高校的名称、所在地、院校特色、高校类型、高校隶属、高校性质、学校网址等信息

### 把得到数据保存到本地表格

### 流程图



## 详细设计项目源程序代码

import requests

from lxml import etree

import openpyxl

import pymysql

headers = {

'user-agent': 'Mozilla/5.0 (Windows NT 6.1;) AppleWebKit/537.3 Chrome/85.0'

}

url = 'http://college.gaokao.com/schlist/p%d/'

data = []

db = pymysql.connect(user="root", password="123456", host="localhost", database="spider")

cursor = db.cursor()

# 内容提取

def get\_content(li):

title = li.xpath('./dt/strong/a/text()')[0]

li\_list = li.xpath('./dd/ul/li')

dizhi = str(li\_list[0].xpath('./text()')[0]).replace('高校所在地：', '')

tese\_list = li\_list[1].xpath('./span/text()')

tese = '/'.join(tese\_list) if tese\_list else '——'

leixing = str(li\_list[2].xpath('./text()')[0]).replace('高校类型：', '')

lishu = str(li\_list[3].xpath('./text()')[0]).replace('高校隶属：', '')

xingzhi = str(li\_list[4].xpath('./text()')[0]).replace('高校性质：', '')

wangzhi = str(li\_list[5].xpath('./text()')[0]).replace('学校网址：', '')

print(title)

data.append([title, dizhi, tese, leixing, lishu, xingzhi, wangzhi])

cursor.execute(

"INSERT INTO `gaokao` (`title`, `dizhi`, `tese`, `leixing`, `lishu`, `xingzhi`, `wangzhi`) VALUES ('%s', '%s', '%s', '%s', '%s', '%s', '%s');" % (

title, dizhi, tese, leixing, lishu, xingzhi, wangzhi))

db.commit()

# 索引页分析

def get\_index(url):

response = requests.get(url)

html = etree.HTML(response.text)

dl\_list = html.xpath('//div[@class="scores\_List"]/dl')

for li in dl\_list:

get\_content(li)

# 保存表格

def save():

book = openpyxl.Workbook()

sheet = book.active

sheet.append(['学校名称', '高校所在地', '院校特色', '高校类型', '高校隶属', '高校性质', '学校网址'])

cursor.execute('select \* from gaokao')

results = cursor.fetchall()

for li in results:

sheet.append(li)

book.save('学校.xlsx')

# 主程序

def main():

for n in range(1, 95):

print(n)

get\_index(url % n)

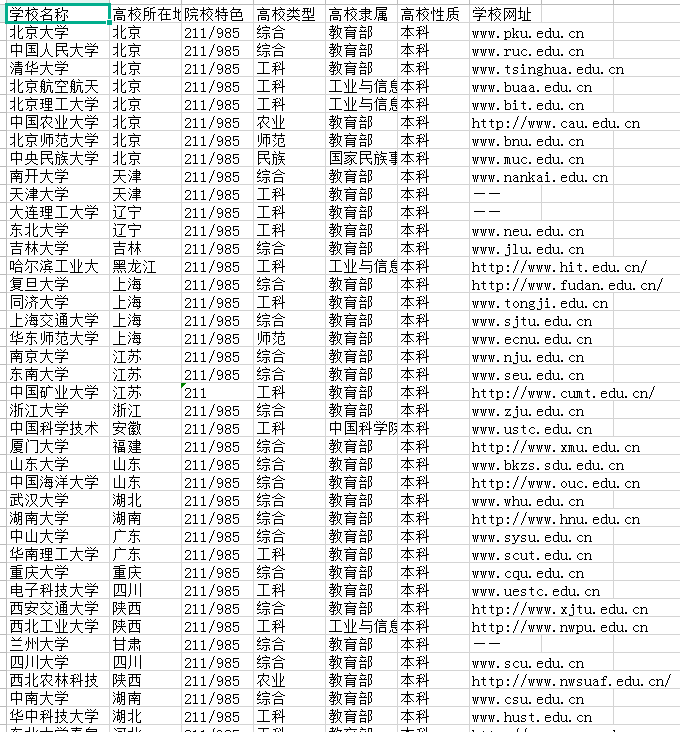
save()

if \_\_name\_\_ == "\_\_main\_\_":

main()

## 运行与测试

运行结果如下：



## 成员分工

## 总结与收获

通过这次课程设计使我懂得了理论与实践相结合是很重要的，只有理论知识是远远不够的，只有把所学的理论知识与实践相结合起来，从理论中得出结论，才能真正掌握这门技术，也提高了自己的独立思考的能力。在设计的过程遇到问题，可以说得上是困难重重，这毕竟第一次做的，难免会遇到各种各样的问题，同时在设计的过程中发现了自己的不足之处，对以前所学过的知识理解得不够深刻，掌握得不够牢固，通过这次课程设计之后，一定把以前所学的知识重新温故。

## 指导教师评语