## 设计目的

网络爬虫是一种自动搜集互联网信息的程序，通过网络爬虫不仅能够为搜索引擎采集网络信息而且可以作为定向信息采集器定向采集某些网站下的特定信息如招聘信息、房地产信息等。

网络在我们的生活中越来越重要，网络的信息量也越来越大，研究该课题可以更好的理解网络爬虫在搜索引擎中的作用以及网络爬虫的原理。

本系统是基于python的房天下二手房信息爬虫系统，目的是采集房天下全部的二手房信息，此数据非常有用，可以为用户作整体的二手房信息分析，让用户更准确的了解现在的二手房市场情况。

## 系统的需求及其基本功能

我们要爬取房天下兰州市全部的在售二手房信息，可以了解在售楼盘的户型、面积、年代、价格等信息。供用户更加方便的筛选对比，选择房产。设计之初，先经过对网页的分析，掌握了网站的请求流程，系统模拟这个流程，得到数据并保存到本地。

## 总体设计方案

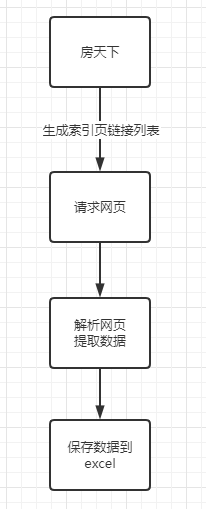
### 分析网站请求流程

### 生成网站请求链接队列

### 提取每套房产的信息（'标题', '户型', '面积', '楼层', '朝向', '年份', '小区名称', '地址', '总价', '单价'）

### 把得到数据保存到excel表中

### 流程图



## 详细设计项目源程序代码

import requests

from lxml import etree

import openpyxl

import time

import random

import re

import pymysql

class Fangtianxia():

def \_\_init\_\_(self, total):

self.headers = {

'user-agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64; rv:6.0) Gecko/20100101 Firefox/6.0',

'referer': 'https://www.baidu.com/'

}

self.url = 'https://lz.esf.fang.com/house/i3{}/'

self.total = total

self.db = pymysql.connect(user="root", password="good1234", host="localhost", database="spider")

self.cursor = self.db.cursor()

def parser(self, dl):

try:

biaoti = dl.xpath('./dd/h4/a/span/text()')[0].strip()

info = dl.xpath('./dd/p[@class="tel\_shop"]')[0].xpath("string(.)").split('|')

huxing = info[0].strip()

mianji = info[1].strip()

louceng = info[2].strip()

chaoxiang = info[3].strip()

nianfen = info[4].strip()

nianfen = nianfen if re.search('^\d+', nianfen) else ''

xiaoqu = dl.xpath('./dd/p[@class="add\_shop"]/a/text()')[0].strip()

dizhi = dl.xpath('./dd/p[@class="add\_shop"]/span/text()')[0].strip()

zongjia = dl.xpath('./dd[@class="price\_right"]/span[1]')[0].xpath("string(.)").strip()

danjia = dl.xpath('./dd[@class="price\_right"]/span[2]/text()')[0].strip()

print(biaoti)

self.cursor.execute(

"INSERT INTO `fangtianxia` (`biaoti`, `huxing`, `mianji`, `louceng`, `chaoxiang`, `nianfen`, `xiaoqu`, `dizhi`, `zongjia`, `danjia`) VALUES ('%s', '%s', '%s', '%s', '%s', '%s', '%s', '%s', '%s', '%s');" % (

biaoti, huxing, mianji, louceng, chaoxiang, nianfen, xiaoqu, dizhi, zongjia, danjia))

self.db.commit()

except Exception as e:

pass

def spider(self):

for n in range(1, self.total + 1):

response = requests.get(self.url.format(n), headers=self.headers)

html = etree.HTML(response.text)

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

for dl in dl\_list:

self.parser(dl)

time.sleep(random.uniform(2, 5))

def save(self):

self.book = openpyxl.Workbook()

self.sheet = self.book.active

self.sheet.append(['标题', '户型', '面积', '楼层', '朝向', '年份', '小区名称', '地址', '总价', '单价'])

self.cursor.execute('select \* from fangtianxia')

results = self.cursor.fetchall()

for li in results:

self.sheet.append(li)

self.book.save('房天下.xlsx')

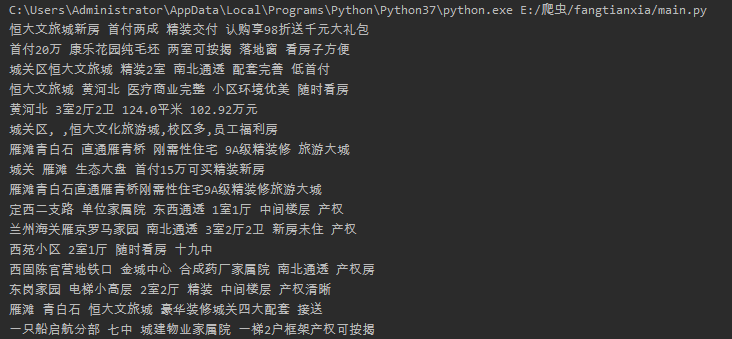
self.db.close()

f = Fangtianxia(20)

f.spider()

## 运行与测试

运行结果如下：





## 成员分工

## 总结与收获

该项目是在学习了《数据可视化技术》课程后，为了更好的展示自己的学习成果；学习本身让我对爬虫有了框架性的认识，懂得构建基本的框架。通过项目的实践，我对《数据可视化技术》课程的知识点有了更深入清晰的认识。让我终于从理论步入实践。

## 指导教师评语