## 设计目的

天天基金网实时新闻爬虫软件——本软件设计目的是为广大基金经理准备，使他们可以每天方便快捷的得到第一手的基金新闻资讯，可以敏锐的察觉市场动态，以便对基金市场的动态有全面的掌握，对于今后市场的走向有初步的分析。

本软件采用python开发，python开发爬虫相比其他编程语言有着非常大的优势，python的代码简洁，美观，上手容易，可以很方便快速的就写好一个爬虫系统，大大提高了生产效率。

## 系统的需求及其基本功能

我们的目标网站就是天天基金的新闻频道，要把频道内的全部新闻都采集到本地，并保存到excel中。目标明确后，我们就要去分析目标网站，包括 URL 结构，HTML 页面，网络请求及返回的结果等，目的就是要找到我们要爬取的目标位于网站的哪一个位置，我们如何才能获取到，从而指定爬取策略；接下来，就是 编码环节 了，使用代码去发起网络请求，解析网页内容，提取目标数据，并进行数据存储；最后，测试代码，完善程序。

## 总体设计方案

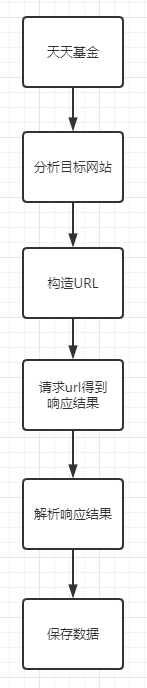
### 分析目标网站

### 按网站规则去请求URL，得到数据

### 从数据中提取解析想要的数据

### 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 pymysql

headers = {

'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.77 Safari/537.36'

}

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

cursor = db.cursor()

def create\_excel():

'''

创建一个新的excel

:return:

'''

wb = openpyxl.Workbook()

ws = wb.active

ws.append(['标题', '发布时间', '内容', '网址'])

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

results = cursor.fetchall()

for li in results:

ws.append(li)

wb.save('data.xlsx')

def spider\_index(page: int):

'''

解析索引页

:param page: 页码

:return: 索引URL列表

'''

url = 'https://roll.eastmoney.com/fund\_{}.html'.format(page)

response = requests.get(url)

html = etree.HTML(response.text)

index\_url = html.xpath('//div[@id="artList"]/ul/li/a/@href')

index\_url = [str(li) for li in index\_url if 'http' in li]

return index\_url

def parser\_content(url: str):

'''

解析内容页

:param url:网址a

:return:

'''

try:

response = requests.get(url, headers=headers)

response.encoding = response.apparent\_encoding

html = etree.HTML(response.text)

title = html.xpath('//div[@class="title"]/text()')[0].strip()

datetime = html.xpath('//div[@class="infos"]/div[1]/text()')[0].strip()

content = html.xpath('//div[@id="ContentBody"]')[0]

content = str(etree.tostring(content, encoding="utf-8", pretty\_print=True, method="html"),

encoding='utf8').strip()

print(title, datetime)

cursor.execute(

"INSERT INTO `jijin` (`title`, `datetime`, `content`, `url`) VALUES ('%s', '%s', '%s', '%s');" % (

title, datetime, content, url))

db.commit()

except Exception as e:

pass

def run(page\_star: int, page\_end: int):

'''

主程序

:param page\_star:开始页数

:param page\_end: 结束页数

:return:

'''

for page in range(page\_star, page\_end + 1):

index\_url = spider\_index(page)

for url in index\_url:

parser\_content(url)

time.sleep(0.1)

create\_excel()

db.close()

if \_\_name\_\_ == '\_\_main\_\_':

run(1, 100)

## 运行与测试

运行结果如下：



## 成员分工

## 总结与收获

在这短短的程序开发时间里，让我深深的感觉到自己在实际应用中所学专业知识的匮乏。让我真真领悟到“学无止境”这句话的涵义。我懂得了实际生活中， 专业知识是怎样应用与实践的。 在这些过程中， 把专业理论运用到具体实践中， 不仅加深我对理论的掌握和运用， 还让我拥有了一次又一次难忘的开发经历。社会不会要一个一无是处的人，所以我们要更多更快的从一个学校人向社会人转变。 为此我们将会在以后的日子里继续努力，不断激励经验，不断磨砺自己，早日走向工作岗位。

## 指导教师评语