摘要

现如今，人民群众对物质生活水平的要求已不再局限于衣食住行，对于精神文化有了更多的需求。电影在我国越来越受欢迎，电影业的发展越来越迅猛，为了充分利用互联网技术的发展，掌握电影业的态势，对信息进行挖掘和处理、提高数据库的利用率，本文采用文献分析法，对网络爬虫的相关内容以及发展现状进行简单介绍，并利用网页抓取技术爬取电影票房网站的相关数据，进行分析，为票房分析提供数据支撑。

关键词：Python 网络爬虫 电影票房

Abstract

Nowadays, the people's requirements for material living standards are no longer limited to clothing, food, housing and transportation, and there is more demand for spiritual culture. Movies are becoming more and more Fashionable in China, and the movie industry is growing rapidly. In order to make full use of the development of Internet technology, grasp the situation of the movie industry, mine and process information, and improve the utilization rate of the database, This paper introduces the content and development of web crawler by literature analysis, and use web page crawling technology to crawl and analyze the box office data related to movie websites, which provides powerful data support for box office analysis.

Keywords: Python web crawler movie box office

目录

摘 要.....1

Abstract.....1

一、绪论.....3

1.1 研究背景.....4

1.2 研究现状.....4

1.3 研究方法.....4

二、系统开发工具与相关技术.....5

2.1 Python 网络爬虫.....5

2.2 系统开发工具.....5

2.2.1 pycharm 工具.....5

2.2.2 MySQL 数据库.....5

2.2.3 Hbuilder X 工具.....6

2.3 系统后台技术.....6

2.4 系统前端技术.....6

三、系统分析.....8

3.1 系统功能分析.....8

3.2 系统功能性需求分析.....10

3.2.1 系统用户功能性需求分析.....10

3.2.2 系统管理员功能性需求分析.....12

3.3 数据获取.....14

3.4 数据分析.....13

3.5 数据展示.....13

四、系统设计.....15

4.1 文件结构图.....15

4.1.1 前端 demo 文件结构图.....15

4.1.2 后端爬虫系统文件结构图.....15

4.2 前端功能模块 16
4.3 登录与注册模块设计 16
4.4 数据库表设计 17
4.5 数据展示模块设计 18
五、系统实现 20
5.1 解决网站反爬机制 20
5.2 实现网络爬虫 23
5.2.1 找出 url 变化规则并获取链接 26
5.2.2 解析并获取网页数据 26
5.2.3 将数据存储至数据库 27
5.3 登录注册模块实现 28
5.4 数据展示模块实现 28
六、 票房网站信息数据爬取结果及分析 32
6.1 以 2019 年的票房榜单 Top20 为例分析 32
6.2 结果分析 32
七、结论与建议 36
7.1 结果分析 36
7.2 不足点 36
7.3 对未来的展望 37
参考文献 38
致 谢 39

一、绪论

1.1 研究背景

近几年，在网络 Python 语言强势的发展背景下，数据思维及数据分析方法也逐渐被运用到各个领域当中，成为人们进行分析数据，传播内在规律的有效途径。要是我们只借助人力下载有关信息，不仅需要花费很多时间，而且得到的消息也非常少。网络爬虫是个可以自己获取网页的次序，它会在拥有大量信息的信息库里十分有效率地提取有用的信息，这就让解决和剖析数据变成了现实。网络爬虫会持续提取网页上的数据储存进本地，通过剖析和筛选，在缓存完成的数据中创建好指引并且把它们储存到体系里，可以协助之后要用的人更方便地查询以及搜索。爬虫系统很好的提取出藏匿在众多数据后的信息十分有效率地搜索，在很大程度上更好地运用了信息数据库。爬虫系统节约了很多人力阅读以及储存数据信息的时间，协助研究人员以及储存众多信息，因此可以更加便捷地获取藏匿在数据之后的知识。

中国的爬虫技能探究虽然开始研发时期比国外晚，但是发展的势头十分迅猛，成果显著。对爬虫技术的研究可以追溯到 2003 年，一些以数据探索为主题的学界研讨会渐渐在中国传播开来。从此之后，国内的研究人员开始慢慢涉足爬虫领域，并逐渐深入。直到 2007 年，一名研究人员在爬虫领域取得了新的突破。他就是浙大的罗兵教授。他的研究基于对古版互联网爬虫技能的精通，对剖析领域与支撑领域分别深入调研，使下载内容的分解过程得以完善。在此基础上，越来越多的学者在爬虫领域取得了新的突破。他们已经可以获取流动的互联网信息，提高了爬虫领域的使用效能。与此同时，也减轻了使用户进行下载的压力。让下载的工作更加高效便捷。因此，更新换代之后的爬虫工具已经成为人们工作时用来信息查找，信息整理，数据分析的一大利器。爬虫工具的使用与发展不仅仅推动了爬虫技术的探究与发展，还十分有利于专家学者研究反爬虫技术。而电影行业的发展越来越快，越来越深入。电影行业的不断发光发热也引起了大量企业和国家统计部门的广泛关注。大数据的新基建的建设同时也加快了电影行业的发展，但是目前关于电影数据的采集和挖掘的技术方案还是不够完善。本文基于网络爬虫理论，开展电影票房相关数据的采集挖掘和分析，而如何从猫眼电影票房网站相爬取需要的数据，是本次项目的核心所在。本文通过 python 编写爬虫脚本以实现获取票房数据的方案，

并找出猫眼电影网的反爬机制，根据相关的反爬机制进行破解。最后把爬取到的数据以图表的形式进行分析介绍。

1.2 研究现状

网络爬虫在消息探索与数值整理进程中发挥着关键作用，上世纪初，就已有科学家对爬虫开启探究模式，现今，爬虫技能已处于成熟阶段。网络爬虫可主动获取网络界面，从而自行下载主人所需要的东西，基本实现了大幅度的数据下载模式，也更便于人们利用其进行高效工作。

在我国，爬虫技能发展的有关探究开启速度比较慢，但其后续的发展却非常迅猛。2003 年该技能得到正式发展，国内数据探索的学论会越来越常态化，在该区域中的探究也随之扩展。2007 年，浙大教授罗兵在旧版网络爬虫的基准上，增添了分析模型，使对该内容的分析越发完善。近几年，经过我国学界的专家、学者们的积极探讨与破除障碍，使得我国流动性网络消息的获得能力不断提升，爬虫体系的效能也随之增强。既减弱了人工完成的压迫感，也逐步实现了高效率的下载任务，成为了大众查找、分解与融合信息中不可或缺的手段。

1.3 研究方法

①著作了解法
②撰写程序语言：Python 语言、HTML 语言、JS 语言、css 语言
③信息库技能：MySQL 信息库技能

二、系统开发工具与相关技术

本章节主要表述该课题所开发的猫眼电影票房数据爬取系统开发所用到的工具及相关技术，还有技术介绍。

2.1 Python 网络爬虫

Python 语言是一种开源编程的语言，其强大的功能、简洁易懂的语法、系统兼容性广以及学习上手成本低的优势受到许多开发者的青睐。Python 具有高效率且简单地实现面向对象编程的优势。对于数据库也能直接方便的进行操作，在处理一些规模较大的数据分析上具有很高的效率。而网络爬虫，简言之，就是进行网页爬取，模拟普通用户去浏览网页却实际在爬取数据的过程。综合来说，python 网络爬虫就是利用 python 这个程序语言来编写爬虫程序或者脚本。基于 python 的网络爬虫程序开发分为三个步骤：首先，做充分调研确立爬虫对象，然后深入调查该网站的反爬虫机制,然后编写爬虫程序并开展爬虫工作获取数据。将获取的数据经过清洗过滤，以 png、excel、mp4 等文件类型或者保存着数据库等方式，保存爬取的数据。常见的网络爬虫有两种，分别是广度优先爬虫和聚焦爬虫。其中广度优先爬虫主要适用一般网络搜索引擎的网络爬虫对象,类似百度、谷歌以及搜狗搜索之类的搜索引擎，采用的网络爬虫主要是广度优先爬虫技术。而聚焦爬虫主要适用于垂直搜索引擎的网络爬虫对象。类似需要搜索某一领域的内容。本课题所采用的的也是这一类型的聚焦爬虫技术。

综上可知，若想依据使用者自身的意见来获取目的网络界面的内容，满足自身的要求，最佳的办法便是以自身需要为主来编写爬虫次序。此探究驻足于猫眼电影网页的体系分解，对爬虫进程中会碰到的各种难题，以 Python 语言为基准撰写了对猫眼电影网电影信息数据获得的互联网爬虫程序。

2.2 系统开发工具

2.2.1. pycharm 工具

PyCharm 是一款高效简洁的 Python 开发工具，代码分析能力强，用户在打代码的过程中可以快速补全 pycharm 所建议的代码，而且自带了多项编辑器。功能十分强大。

2.2.2 MySQL 数据库

MySQL 数据库是一款强大的数据库，体积占比不大、学习成本低且系统兼容性十分优秀。在使用上方便易懂。

2.2.3. Hbuilder X 工具

Hbuilder 是一款 HTML 的编辑器，同时也结合了 IDE。从外观上看，该工具界面清爽，而且性能敏捷使用起来很轻巧。

2.3 系统后台技术

1. flask-web 框架技术

Flask 是一个的基于 python 的 web 框架。

2.requests 库

requests 库基于 urllib，在本系统里，requests 库主要功能是请求目标网站、各种请求方法等方式。

3. Beatifulsoup 库

BeautifulSoup 一种解析器，是借助于 Python 进行开发的。该解析器将不规则标签进行整理，并且进一步建立分析树。Beautifulsoup 组件的功能相当强大，其主要功是能够检索当前页的内容，按照需要选取有用的部分，且输出时能够自动校对格式。

4. Numpy 库

Numpy 库主要用于数组运算，在本系统中，破解猫眼电影字体反爬里有用到该库来计算欧氏距离配对字体。

5. lxml 库

lxml 库是一款解析器，在解析网页内容中发挥着不可或缺的角色。

2.4 系统前端技术

1.layui 框架技术

layui 是一款前端 UI 框架，高度模块化的独特设计，使其上手学习成本大大降低。其中在本系统的演示部分，类似导航栏，主题选择等模块有涉及到 layui 技术。

2. jQuery 技术

jQuery 是一个 JavaScript 框架，接口的短小清晰、插件的丰富以及语法的独特性让使用者用起来十分的方便。而且该框架的兼容绝大多数浏览器，兼容性十分优秀。

3. Echarts 框架技术

4.ECharts是一款前端可视化框架，使用者可以使用该框架搭建自己所需要的图表，因为Echarts 提供了许多生动美观的图表供使用者使用。其中在系统的演示部分，类似折线图、词云图、柱状图等数据图表都用到了 ECharts 技术。

三、系统分析

3.1 系统功能分析

本电影信息数据爬取系统主要由后台管理模块和用户模块两大模块组成，其中用户模块的适用对象为普通用户，主要功能包括了登录注册、主题设置、个人中心、可视化展示以及信息推送功能。其中除了登录注册功能之外，其他功能需要再用户登录的情况下才能使用。接着是后台管理模块，其适用对象主要为管理者。后台管理模块的主要功能为：数据爬取、用户管理、页面管理以及数据管理。其管理权限较大。具体功能模块示意图如 3-1 所示。

其中，管理员功能用例图对应图 3-3，用户功能用例图对应图 3-2

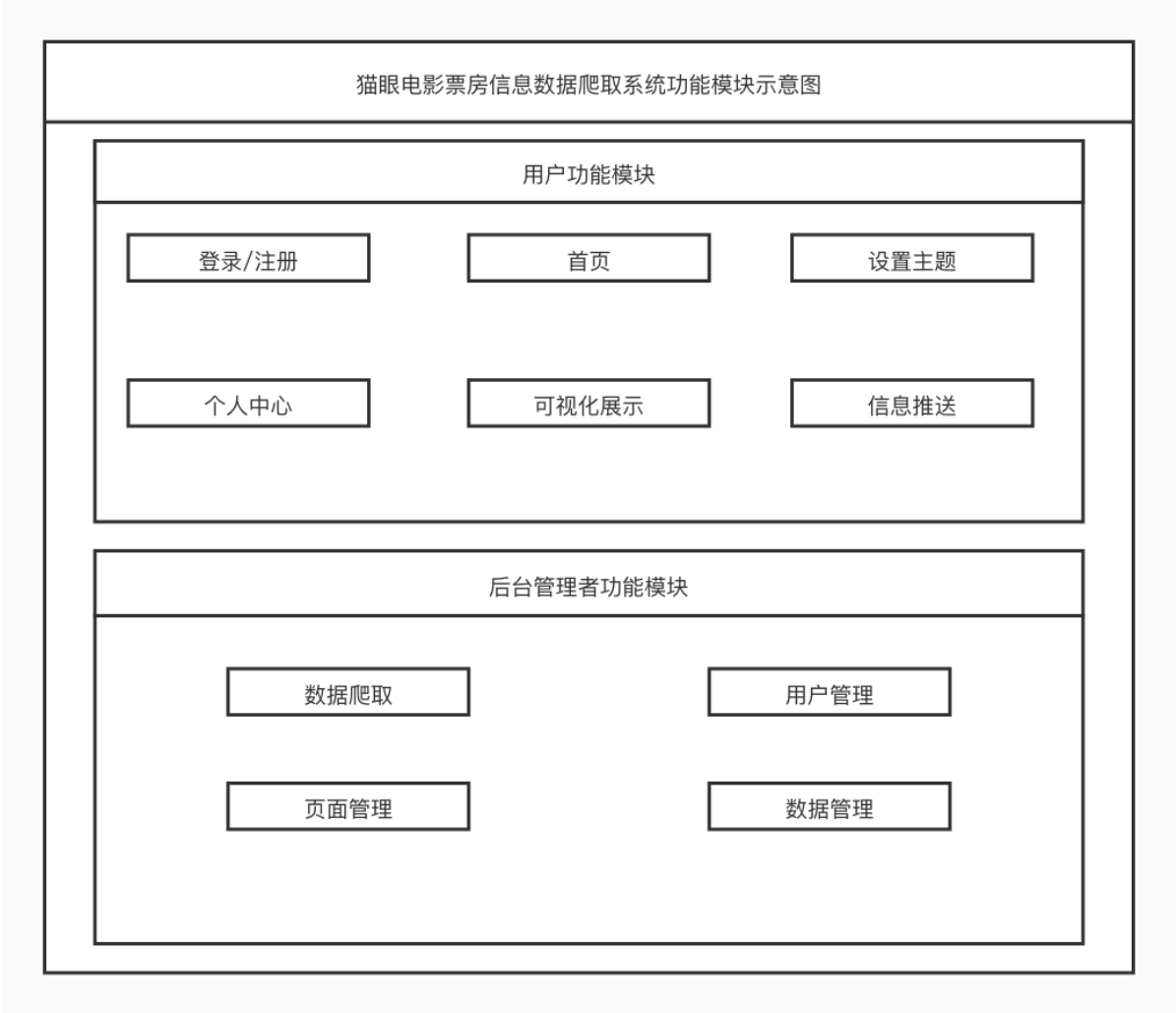

图 3-1 系统功能模块示意图

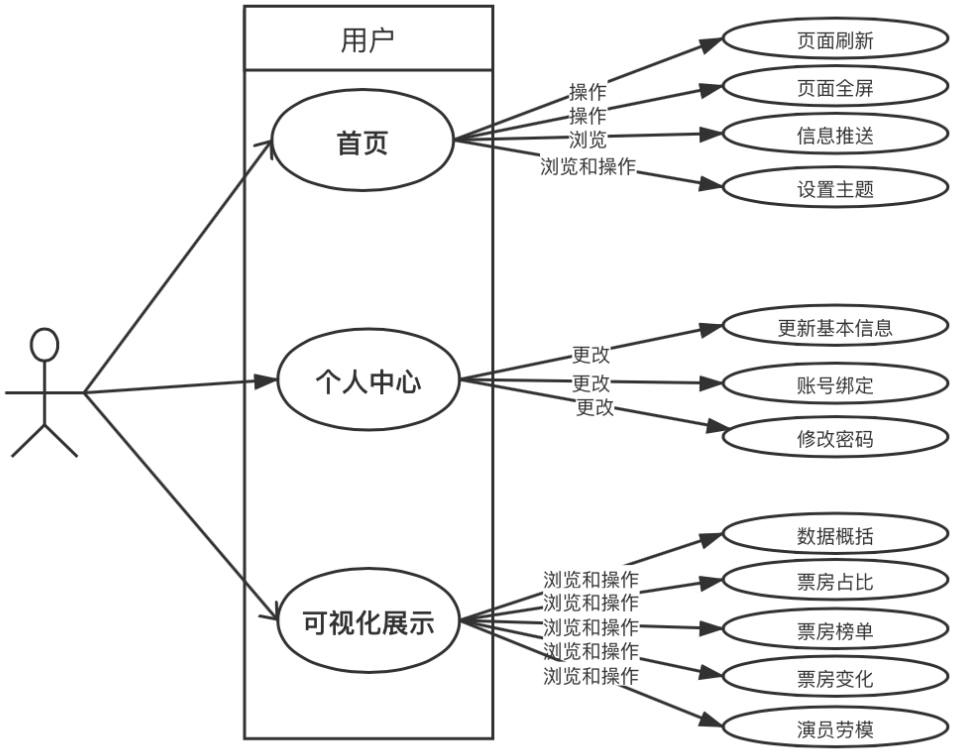

图 3-2 用户功能用例图

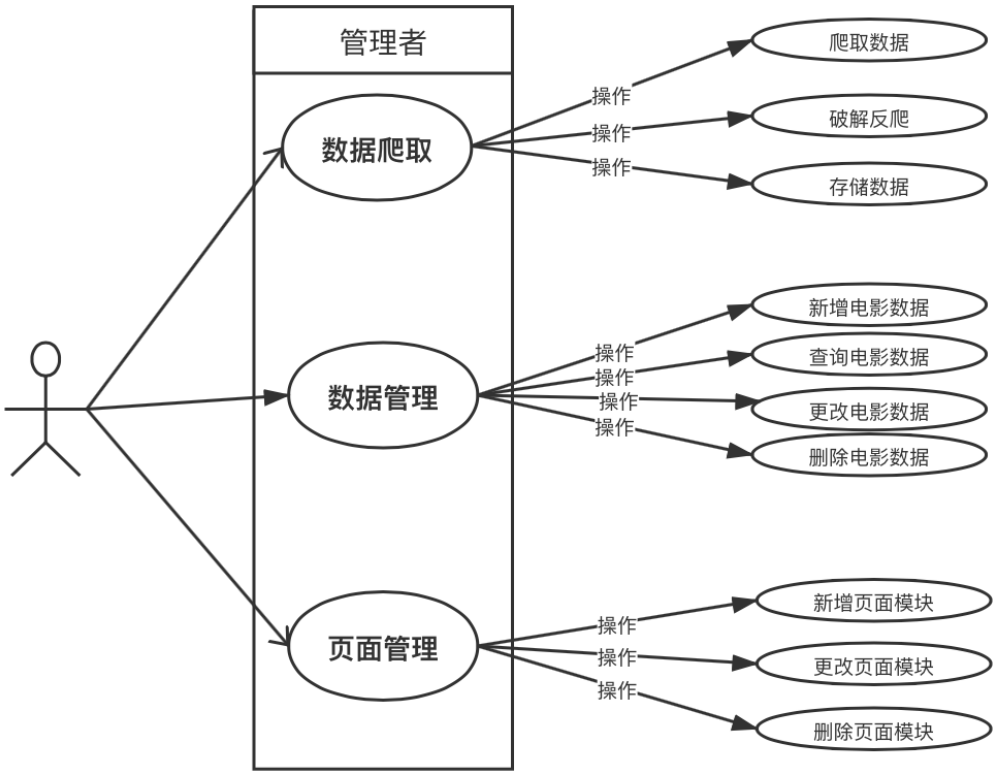

图 3-3 管理员功能用例图

3.2 系统功能性需求分析

本节从用户功能和管理员这两个模块分别阐述其功能性需求和做详细的分析介绍。通过详细的分析介绍进一步明确系统功能性需求，为接下来的系统设计与开发做好布置工作。

3.2.1 系统用户功能性需求分析

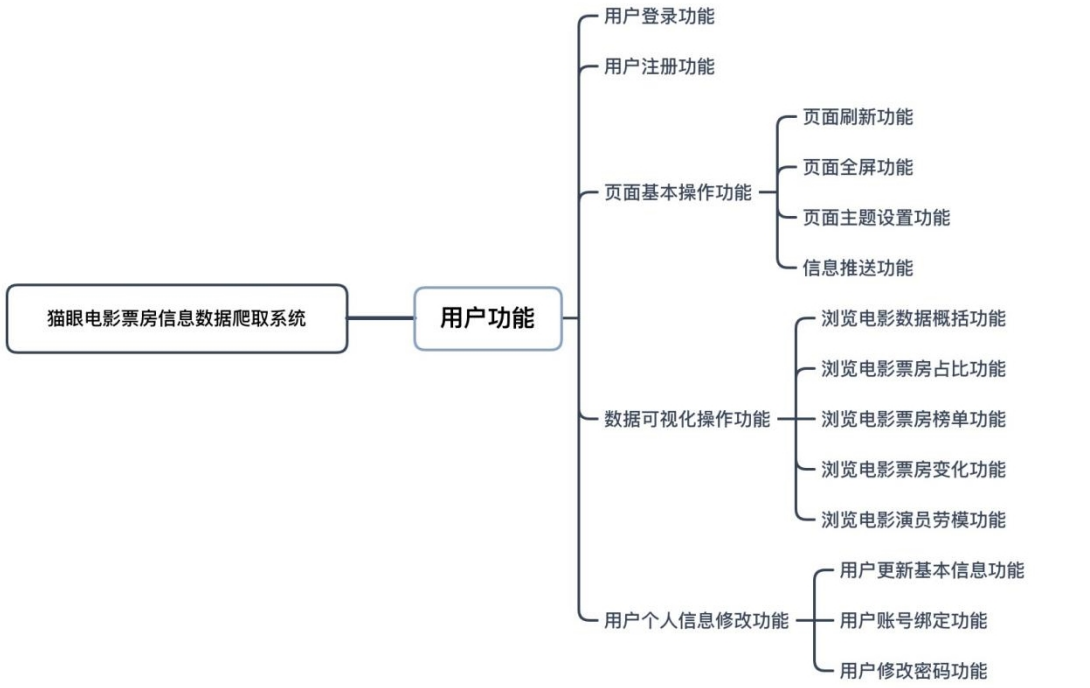

图 3-4 用户功能需求概述图

图 3-4 为猫眼电影票房信息数据爬取系统的用户功能需求的概述图，下面将对图 3-4 所列的功能进行详细的讲解和说明。

（1）用户登录功能

用户登录功能为该系统的基础功能，用户进入该系统的前提是登录账号，登录账号之后可以进入系统，并且系统会开放所有功能供用户使用。用户在未登录账号的情况下，无法进入该系统。

（2）用户注册功能

用户注册功能的作用是让用户在未拥有账号的状态下可以进行注册，获得账号，以得到更多的功能。

（3）页面基本操作功能

页面基本操作功能是该系统的基础功能，该功能具有四个子功能，分别为页面刷新功能、页面全屏功能、页面主题设置功能以及信息推送内容。以下对其四个子功能进行详细的讲解和说明。

页面刷新功能：该功能为页面基础功能的子功能之一，主要是给系统页面进行刷新，将系统页面置于初始状态。

页面全屏功能：该功能为页面基础功能的子功能之一，主要是将系统页面放至全屏状态，方便用户更详细的查看页面。

页面主题设置功能：该功能为页面基础功能的子功能之一，主要是将系统页面的主题颜色、按钮进行更改，方便用户根据自己的喜好对系统页面主题进行 DIY 设置。

信息推送功能：该功能为页面基础功能的子功能之一，主要是查看和预览用户的个人推送信息。

（4）数据可视化操作功能

数据可视化操作功能为该系统的重要功能，该功能具有五个子功能，分别为浏览电影票房变化功能、浏览电影数据概括功能、浏览电影票房榜单功能、浏览电影演员劳模功能以及浏览电影票房占比功能。以下将其五个子功能进行详细的讲解和说明。

浏览电影数据概括功能:该功能为数据可视化操作功能的子功能之一，主要是对全部电影信息数据进行概括，将其基本信息以表格形式展示出来，方便用户浏览查看。

浏览电影票房占比功能：该功能为数据可视化操作功能的子功能之一，主要是对各个电影类型票房占比情况分别以柱状图和玫瑰图的形式展示出来，用户可以选择不同的年份和月份查看不同时期时的各个电影类型票房占比情况。

浏览电影票房榜单功能，该功能为数据可视化操作功能的子功能之一，主要是将电影票房靠前的电影名字以词云图的形式展示出来，用户可以选择不同的年份和排行数量，查看不同时期时电影票房排行靠前的电影名字。票房越高的电影，其名称字号大小将会更大。方便用户对电影票房查看，一目了然。

浏览电影票房变化功能，该功能为数据可视化操作功能的子功能之一，主要是将 2015 年至 2019 年的电影票房走势以折线图的形式展示出来，用户可以选择不同的电影类型查看该电影类型的票房走势情况。

浏览电影演员劳模功能，该功能为数据可视化操作功能的子功能之一，主要是将电影演员参演次数情况以词云图和柱状图的形式展示出来，用户可以选择不同的年份和排行数量，查看不同时期时电影演员参演次数靠前的演员名字。参演次数越多的演员，在词云图里，其名字的字号大小将会更大，在柱状图里，将会更明显。

（5）用户个人信息修改功能

用户个人信息修改功能为该系统的基础功能，该功能具有三个子功能，分别为用户更新基本信息功能、用户账号绑定功能以及用户修改密码功能。以下将其三个子功能进行详细的讲解和说明。

用户更新基本信息功能，该功能为用户个人信息修改功能的子功能之一，用户在该功能上可以更改自己的邮箱、昵称、个人简介、街道地址以及联系电话信息。

用户账号绑定功能，该功能为用户个人信息修改功能的子功能之一，用户可以在该功能上可以修改密保手机、密保邮箱、绑定 QQ 以及绑定微信。

用户修改密码功能，该功能为用户个人信息修改功能的子功能之一，用户可以在该功能上修改自己的账号密码。

3.2.2 系统管理员功能性需求分析

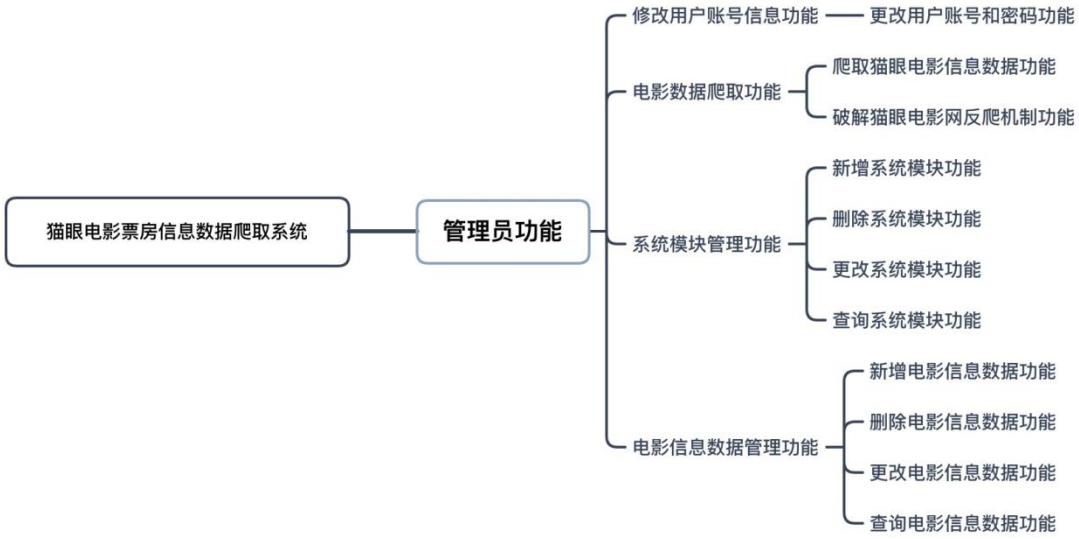

图 3-5 管理员功能需求概述图

图 3-5 为猫眼电影票房信息数据爬取系统的管理员功能需求的概述图，下面将对图 3-5 所列的功能进行详细的讲解和说明。

（1）修改用户账号信息功能

修改用户账号信息功能可以更改用户的账号和密码，管理员可以根据需求，将用户的账号或者密码进行更改

（2）电影数据爬取功能

电影数据爬取功能，该功能是整套系统的核心。该功能具有两个子模块，分别为爬取猫眼电影信息数据功能和破解猫眼电影网反爬机制功能。以下就其两个子功能进行详细的讲解和说明。

爬取猫眼电影信息数据功能，该功能为电影数据爬取功能的子功能之一。管理员可以将猫眼电影网所需的信息数据获取下来，为其他功能使用。

破解猫眼电影网反爬机制功能，该功能为电影数据爬取功能的子功能之一。利用该功能可以破解猫眼电影网的部分反爬机制，类似数字乱码这方面的技术难题。对管理员来说十分有用。

（3）系统模块管理功能

系统模块管理功能，该功能是整套系统实现可视化的关键。该功能具有四个子功能，下面分别阐述这四个子功能的内容：

新增系统模块功能，该功能为系统模块管理功能的子功能之一。管理员可用该功能新增所需的系统模块。

删除系统模块功能，该功能为系统模块管理功能的子功能之一。管理员可用该功能删除所需的系统模块。

更改系统模块功能，该功能为系统模块管理功能的子功能之一。管理员可用该功能更改所需的系统模块。

查询系统模块功能，该功能为系统模块管理功能的子功能之一。管理员可用该功能查询所需的系统模块。

（4）电影信息数据管理功能

电影信息数据管理功能，该功能具有四个子功能，分别为新增电影信息数据功能、删除电影信息数据功能、更改电影信息数据功能以及查询电影信息数据功能

新增电影信息数据功能，该功能为电影信息数据管理功能的子功能之一，管理员可以用该功能新增电影信息数据，以此来更新系统。

删除电影信息数据功能，该功能为电影信息数据管理功能的子功能之一，管理员可以用该功能删除电影信息数据，以此来清洗过滤不需要的信息数据。

更改电影信息数据功能，该功能为电影信息数据管理功能的子功能之一，管理员可以用该功能更改电影信息数据，纠正格式不规范或者错误的电影信息数据。

查询电影信息数据功能，该功能为电影信息数据管理功能的子功能之一，管理员可以用该功能查询电影信息数据，以此来查找到所需的电影信息数据。

3.3 数据获取

系统数据分析里，必不可少的一环是数据获取。因为系统的数据分析是基于数据来展开的。数据获取之前要明确什么数据是需要用到的，什么是不需要的。经过筛选之后确定数据目标，进而在获取数据。根据本次课题，需要获取的信息主要是通过 Python 爬取筛选 2015 年至 2019 年之间的评分靠前电影数据，例如电影名、评分、票房以及上映时间等内容。

3.4 数据分析

在确定获取数据的目标及得到了数据之后，进一步做的是分析数据。本系统主要是通过统计分析的分析方式去研究某个时间段的票房变化及演员的参演次数，来完备该系统。

3.5 数据展示

数据展示原理是将数据进行可视化，让用户方便清晰地了解到该系统数据的变化。此系统的数据可视化图表主要是以五种形式存在，分别是: 表格、词云图、折线图、柱状图和玫瑰图来分析结果。其中数据概述的结果使用表格形式显示，票房占比的结果使用柱状图和玫瑰图形式显示，票房榜单的结果使用词云图形式显示，票房变化的结果使用折线图形式显示，演员劳模的结果使用词云图和柱状图形式显示。

四、系统设计

系统详细设计阐述了该系统如何实现的一些较为重要的功能，该章节利用图文结合的方式，让表述更加清晰，更加方便读者了解到本系统的具体构造。

4.1 文件结构图

4.1.1 前端 demo 文件结构图

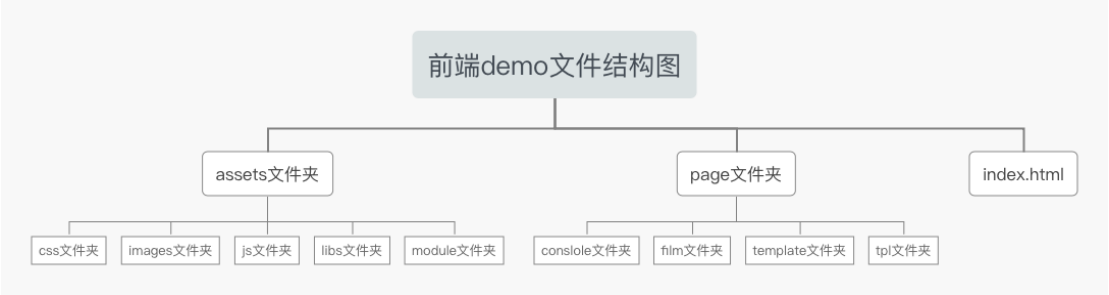

图 4-1 前端 demo 文件结构图

1. assets 文件夹是本系统的资源目录，包括 js，css 图片，依赖的库文件都在里面。

2. page 文件夹为主页面目录，各个模块的页面，需要 localhost 运行起来才能打开。

3. index.html 为根文件，里面存放了前端的主体代码，采用前端框架是 layui + jQuery+echarts。

4.1.2 后端爬虫系统文件结构图

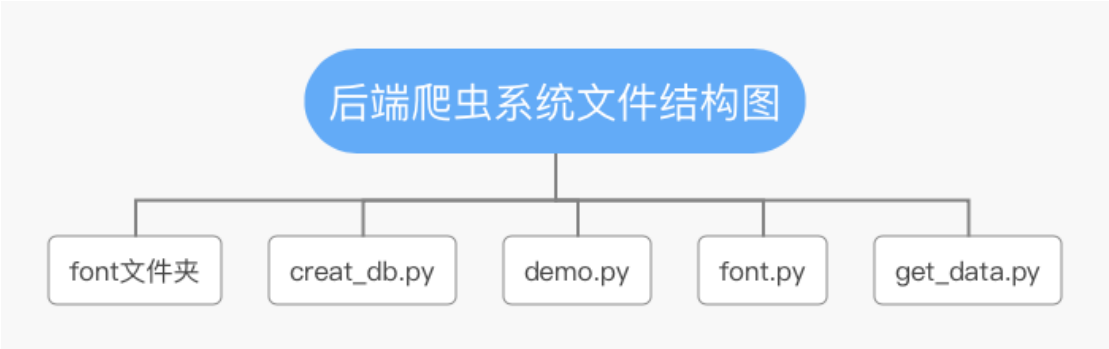

图 4-2 后端爬虫系统文件结构图

关于爬取网站的信息有这 5 个文件：

1.creat_db.py 文件，主要功能是创建数据库。

2.demo.py 文件，主要功能是调用数据到前端，实现前后端的数据交互。

3.font.py 文件，主要功能是字体反爬破解。

4.get_data.py 文件，主要功能是爬取猫眼电影网站数据。

5.font 文件夹，主要功能是字体配对。

4.2 前端功能模块

系统的展示层最主要的页面和相关解释如下：

console.html:控制台页面，用于展示。

bangdan.html：票房榜单页面，用于分析不同的时期里，电影的票房排行。

bianhua.html:票房变化页面，用于分析不同电影类型在 2015 年至 2019 年的票房走势。

data.html:数据概括页面，用于展示爬取的电影数据内容。

laomo.html:演员劳模页面，用于分析不同时期里，演员的参演次数排名。

piaofang.html:票房占比页面,用于分析不同时期里,各个电影类型的票房。

login.html：用户登录注册页，用于用户的登录与注册。

user-info.html：用户信息页面，用于用户修改自己的信息。

tpl-message.html：信息通知面板页面，用于信息通知。

tpl-password.html：用户密码修改页面，用于用户修改自己想要密码。

tpl-theme.html：主题修改页面，用于用户修改系统页面的主题、标签，按钮等操作。

index.html: 首页，用于展示和操作相关可视化界面。

以上 html 页面主要采用了 layui+jQuery 搭建主体框架,其中可视化数据图表采用的是 Echarts 图表库，对数据挖掘和整合非常友好。

4.3 登录与注册

登录和注册页面是用户进行登录和注册的地方,在登录页面中用户能够输入账号和密码进行登录，倘若用户是第一次访问并没有账号，可以点击注册按钮进行账号注册。

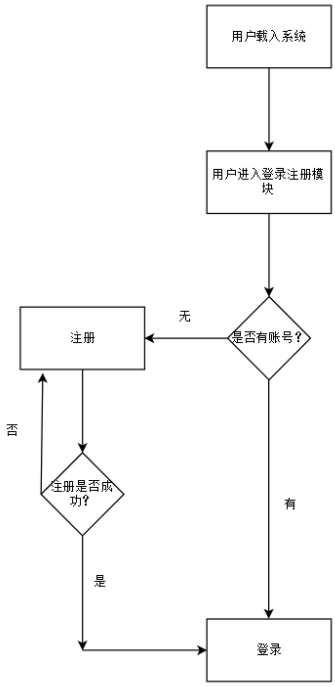

图 4-3 登录流程图

4.4 数据库表设计

通过利用 Python 抓取电影票房数据内容，进行有效的清洗、转换等操作之后保存下来。接着做数据表设计。包括表中的字段名称、数据类型、数据长度、是否为主键，字段说明等。如表 4-2 所示：

表 4-2 电影 films 表

字段名称	数据类型	长度	是否为主键	字段说明
name	varchar	255	是	电影名称
time	varchar	255	否	电影上映时间
type1	varchar	255	否	电影的类型
type2	varchar	255	否	电影的类型
type3	varchar	255	否	电影的类型
type4	varchar	255	否	电影的类型
type5	varchar	255	否	电影的类型
country	varchar	255	否	电影出品国家
length	varchar	255	否	电影长度
year	int	0	否	上映时间（年份）
month	int	0	否	上映时间（月份）
day	int	0	否	上映时间（日期）
director	varchar	255	否	导演
actor1	varchar	255	否	演员

actor2	varchar	255	否	演员
actor3	varchar	255	否	演员
actor4	varchar	255	否	演员
score	varchar	255	否	电影分数
people	int	0	否	评分人数
box_office	bigint	0	否	电影票房
type	varchar	255	否	电影类型总和

该系统数据库主要是 films 表，films 表中储存的是爬虫程序在猫眼电影网站上爬到的所有电影数据，其中的字段包括了`name`, `time`, `type1`, `type2`, `type3`, `type4`, `type5`, `country`, `length`, `year`, `month`, `day`, `director`, `actor1`, `actor2`, `actor3`, `actor4`, `score`, `people`, `box_office`, `type，name 字段表示电影名称、time 字段表示电影上映时间、type1-5 表示电影的类型、country 表示电影出品国家、length 表示电影长度、year、month、day 表示上映时间、score 表示分数、actor 表示演员、director 表示导演、people 表示评分人数、box_office 表示票房。type 表示电影类型总和。

图 4-4 为数据库的实体关系 E-R 图，主要作用是清晰地展示出系该统中各实体之间的关系。

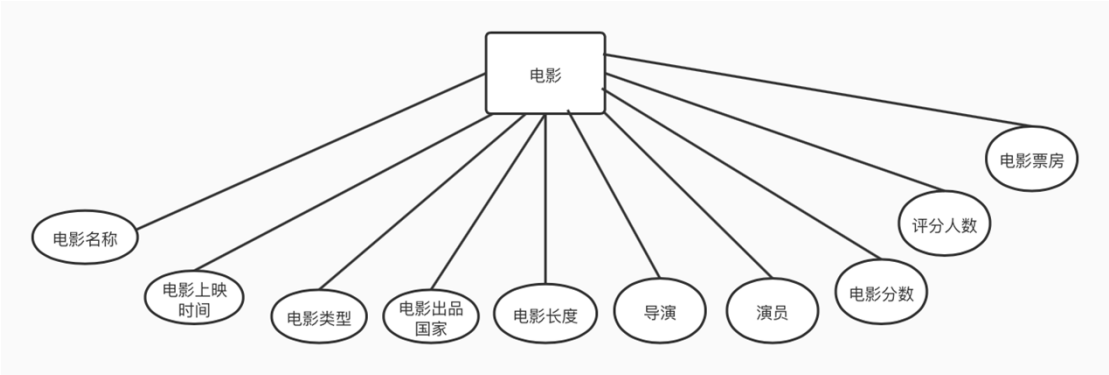

图 4-4 数据库 E-R 模型图

4.5 数据展示模块设计

本系统的功能模块以猫眼电影信息数据为主，通过构建多种多样的图表模型向用户展示爬取到的票房数据，即直观又方便。同时了方便了非专业人员对该系统的理解及使用。该系统将从五个模块对从猫眼电影网爬取回来的电影数据进行分析，分别是数据概述、票房占比、票房榜单、票房变化，演员劳模五个模块作研究分析。模块示意图如 4-5 所示：

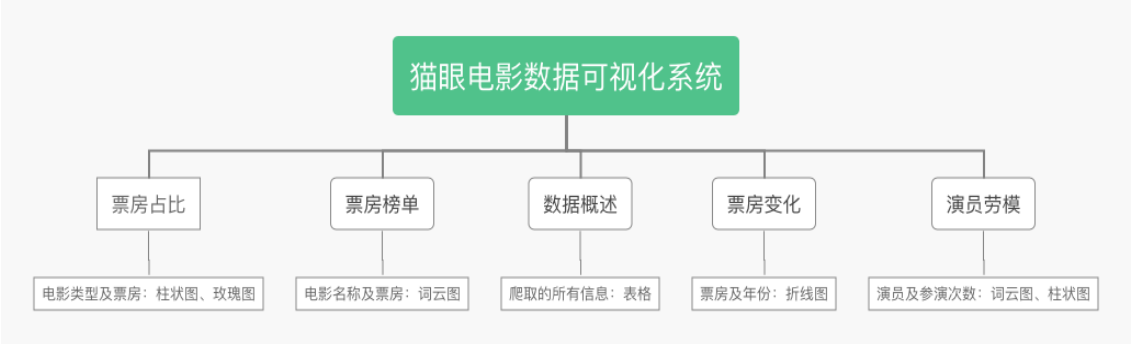

图 4-5 数据展示模块图

（1）票房占比模块：

使用柱状图展现表达出猫眼好评靠前电影中哪个类型的电影在猫眼电影网的票房最高。可以理解为哪个类型的电影更受观众的青睐。使用玫瑰图展现表达出猫眼好评靠前电影中哪个类型的电影占最大的比例。

（2）票房榜单模块：

使用词云图将猫眼好评靠前电影清单中，将票房突出的电影名字放大处理。字号越大更能凸显哪部电影的票房更高，更受观众的欢迎。

（3）数据概述模块

使用表格形式将电影的基本信息展示出来，类似电影名字、电影出品国际、票房、评分及评分人数等内容。电影的基本一目了然

（4）票房变化模块：

使用折线图展现表达出猫眼好评靠前电影中哪个类型的电影，在 2015 年至 2019 年这段时间票房的走势。通过选择不同的电影类型，直观的看到该类型电影的票房走势。

（5）演员劳模模块：

使用用词云图将猫眼好评靠前电影清单中，将演员参演次数突出的演员名字放大处理。字号越大更能凸显哪位演员参演次数更多，在电影行业里更加投入。使用柱状图展现比较出猫眼好评靠前电影中演员参演次数的高低。

五、系统实现

5.1 解决网站反爬机制

爬取猫眼电影网站的电影详情数据，首先是要解决网站的反爬机制，然后获得权限访问网站数据。否则爬取工作无法进行。所以爬取数据要绕过网站的反爬机制，通过研究该网站发现了有以下三个机制：

（1）反爬机制一：申请向猫眼电影网服务器发送访问请求时，该服务器会判断是否为用户浏览器发来的请求，这其中会有个判断识别。那么爬虫就需要绕过该识别。于是我们需要在 Python 里添加头部信息文件 headers。用这个 headers 来绕过服务器的判断识别。

```
head = """
Accept:text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8
Accept-Encoding:gzip, deflate, br
Accept-Language:zh-CN,zh;q=0.8
Cache-Control:max-age=0
Connection:keep-alive
Host:maoyan.com
Upgrade-Insecure-Requests:1
Content-Type:application/x-www-form-urlencoded; charset=UTF-8
User-Agent:Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/59.0.3071.86 Safari/537.36
"""
```

图 5-1 添加 headers 请求头

（2）反爬机制二：使用 python 进行爬虫的时候，猫眼电影网站会检测到我们的访问请求过于频繁。这时候服务器就会阻止我们的访问。为了解决该困扰，在 python 里导入 time 方法。通过 time.sleep（）降低访问请求频率。模拟打开页面以查看页面的真实用户的操作，避免被猫眼电影网站阻止或拒绝。

```
import time

time.sleep()
```

图 5-2 time.sleep（）函数

（3）反爬机制三：使用 python 进行爬虫的时候，是个自动化采集数据的过程，如果采集的方式不当，采集频率高或者数量多了，猫眼电影网的反爬机制就

会监测到我们正在使用程序爬取数据，这时候会出现一个滑动验证码。不滑动该验证码，爬取任务就无法继续进行。但破解滑动验证码的反爬机制是个比较大的挑战。到目前为止，笔者的操作方式是人工手动滑动图片。

图 5-3 猫眼电影滑动验证码

（4）反爬机制四：在猫眼电影详情页使用谷歌浏览器开发人员工具发现，猫眼电影网会有独特的文字反爬机制。致使我们没法在开发人员工具里直接获取准确的数字。而且每次刷新页面，猫眼电影网页源代码里的文字下载链接每都会随之改变。这时候为了获取准确的数字，就需要找出字体规则然后做进一步的判断，得到精准数字。下面是动态字体反爬破解处理过程：

一、网页分析

首先尝试对猫眼电影详情页的信息进行获取。

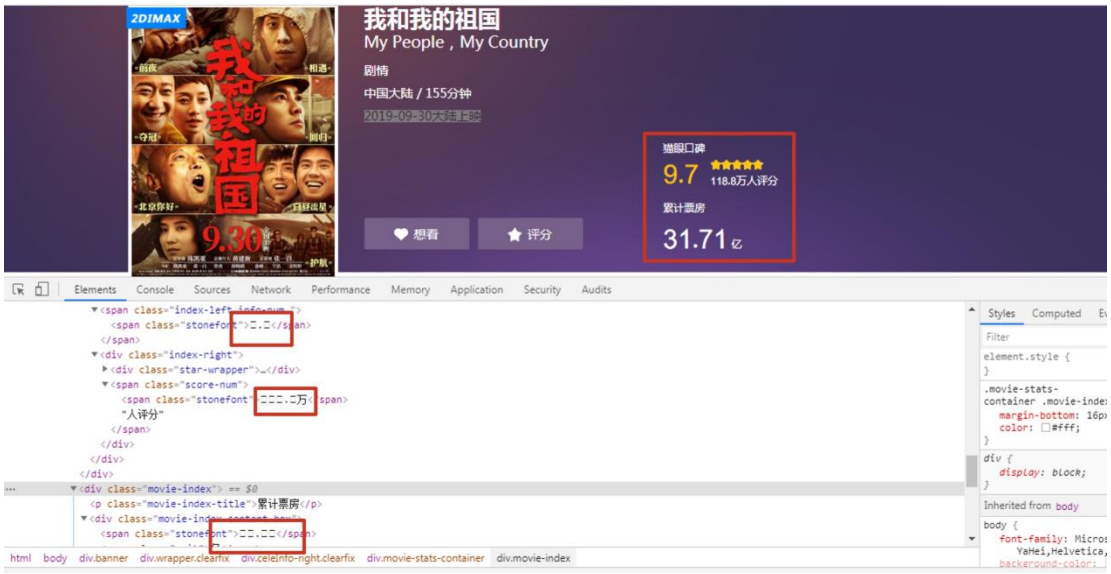

图 5-4 开发工具检查猫眼电影详情页

通过谷歌浏览器开发人员工具发现，猫眼电影网采用了文字反爬处理，导致我们在开发人员工具里看到的数据是框框，也就是所谓的乱码。

图 5-5 查看猫眼电影网页源码

用谷歌浏览器查看网页源码并刷新页面，发现图 5-5 三处编码会随之改变。

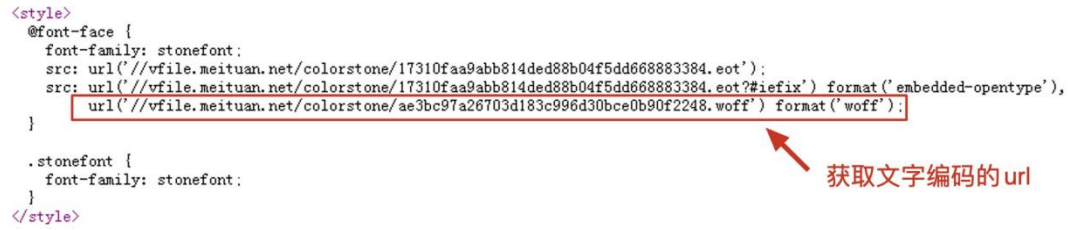

图 5-6 获取猫眼电影网页源代码的文字编码的 url

于是搜索关键字（stonefont），找到图 5-6 里的三个 url 地址，将最后一个地址的字体文件下载下来（ woff 格式）

二、处理字体

使用 Font Creator 工具 打开下载的字体文件（maoyan.woff）

.null	x	uniEFD2	uniEFD4	uniE069	uniE26F	uniEFF5	uniE7CF	uniF816	uniE9FA
	.	0	9	2	6	7	3	1	8

图 5-7 maoyan.woff 字体对应的编码

通过图 5-7 我们将每个数字的编码，输入至 Python 里，构成字典。（下面是笔者下载的字体文件对应的编码）