数据结构（Python 课程）课程项目 3-2（视频信息收集和管理）

（北京大学数学学院，2014 年 12 月 11 日）

本项目要求以项目组为单位独立完成下面工作。一个项目组由两位同学组成。项目的评分要求包括：

1. 所提交的程序应该完整（可运行并正确实现了所需功能），应包括主程序和一个 demo 程序（演示程序）。程序的模块划分合理，代码清晰，格式合理，易读易理解；
2. 所提交报告对项目工作描述应清晰准确，说明为什么采用有关的结构设计，其中的想法和解决的问题等；报告应包含对所完成工作的分析，讨论其优点和缺点；
3. 报告中应总结在完成这一项目过程中的体会和遇到的困难。

项目描述：

本项目要求实现一个能对用户喜爱的电影/电视剧/动画等进行管理与分析的程序。具体要求如下（部分功能需要参考的知识在附录中给出）：

1. 用户可以向视频库单个或批量地添加本地视频（或网络视频）；
2. 程序可以自动从网络上获取视频信息，包括但不限于：视频标题（可由用户提供，也可尝试从文件名猜测），类型（动作/剧情/惊悚/动画/治愈/……），发行年代，导演，编剧，演员/声优列表，综合评分等，同时用户也可以在本地对这些信息进行修改和补充，并留下自己的文字评价；
3. 所有视频库相关信息都应能导出到本地文件中，或者从本地文件导入；
4. 程序应提供适当且方便的接口，使用户能方便简洁地在本地视频库中完成多种搜索与分析工作，例如：
   I) 将宫崎骏导演的所有评分在8分以下的电影修改为8分；
   II) 统计周星驰参演的所有影片的平均得分；
   III) 总结出虚渊玄为编剧的所有作品中最常标记的3个视频类型；
   IV) 获取2005年以前施瓦辛格参演的所有动作片，按得分排序，生成播放列表；
5. 用户可以借助视频库中已有的信息分析演员之间的关联。例如，如果两名演员多次（次数可由用户给定）同时出演一部影片，就称两人是友人。可以实现的功能如：
   I) 用户给出演员三泽纱千香与水树奈奈，程序能试图求一条类似三泽纱千香-中村悠一-水树奈奈的演员链，使链条尽可能短且其中任意相邻两人都是友人；
   II) 给定一名演员，程序能尝试求出演员所在的一个好友团体，即团体中的演员彼此都是友人。思考能否尽可能求出人数最多的团体；
6. 高级用户应该能遵循给定的规范（应在报告中说明），在不修改项目已有代码的前提下，自行实现新的类来扩展程序的功能，比如，添加更多的在网络上获取视频信息的来源，添加更多的查询与分析策略，更多的查询结果输出方式（比如输出网页/更多格式的播放列表）等。注意本条中的举例仅说明项目系统应有这类扩展性，并不强制要求实现；

7. 项目成员认为有趣或实用的其他功能。

项目要求：

根据题目要求设计并实现所需的功能

1. 设计所需要的数据结构，根据需要定义有用的类（class），如考虑用什么数据结构存储视频信息，演员信息等；
2. 实现一个脚本文件（Python 程序文件）main.py，用户执行它便能进入交互界面进行操作，并能进行简单的查询与分析，这里的“简单”指交互可实现的功能可以不覆盖项目描述 4 中的全部功能的任意组合。实现中应注意网络操作，文件操作，用户错误输入等等各种可能导致错误的情况，并友善且健壮地处理错误；
3. 实现另一个脚本文件 demo.py，展示项目中开发的全部功能，包括类似项目描述 4 中的高级查询，以及所实现的项目描述 6 中的扩展范例；
4. 提交作业时，应随程序文件一起提供一个供测试用的本地视频信息库文件。显然，所提供的 demo.py 文件应包含装入这个库的操作。

可以根据情况和需要，把整个系统分别实现为几个模块，以利于系统的开发。

报告的要求：

报告大致可以分为几个部分：

1. 对问题的分析和整体系统的设计概述；
2. 具体的数据结构和程序结构设计；
3. 实现中的关键问题和技术分析；
4. 系统完成的情况和实际效果的说明；
5. 重要算法的时间复杂性分析，并说明自己的程序没有不合理的空间浪费；
6. 完成了这个系统之后的回顾和分析：优点和缺点，改进可能性。

报告可以参考以上结构组织。

附注：

本附录提供一些可能对项目开发有用的信息。进一步的信息可以需要查阅 Python 语言和标准库文档，或自行从网络获取。

1. 关于如何在 Python 程序直接获取互联网上的信息：标准库中的 urllib 库实现了读取给定网址的网页内容的功能。所读取的网页内容可能是 html/xml/json 等格式的纯文本（串），可以用标准库提供的 html.parse，xml 和 json 等模块进行网页内容解析。请自行查阅标准库的帮助文件，进一步了解这些概念的具体使用方法。

   有一些第三方开发的库提供了类似上述标准库的功能，可能提供了更好的功能并已被广

泛使用，例如 requests/beautifulsoup/pyquery/lxml 等。也可以直接利用正则表达式提取 html 等格式文件里的有用信息，但相对而言，这种手法比较“脏”，写代码更麻烦，代码的意义可能更不清晰，并且（或者）更难阅读难修改。

如果在程序的执行中使用了第三方的库，应在报告里说明情况，并在上交的项目文件里包括所用到的代码文件。

2. 部分视频网站提供了 api（应用程序开发接口）供编程开发者调用。本项目推荐使用豆瓣电影的 api，因为一般开发者不需要特别申请授权就能直接使用它提供的部分公开接口的功能（限制每分钟不超过 40 次请求）。例如，以 fate 为关键字搜索豆瓣视频，可以通过访问“http://api.douban.com/v2/movie/search?q=fate”，豆瓣将返回 json 格式的数据。具体信息可以参考 http://developers.douban.com/wiki/?title=movie_v2

3. 程序输出播放列表时，可以直接将其输出为一个文本文件，在一行输出一个视频文件的路径即可。事实上，这已经是合法的 m3u 格式播放列表，可被大量视频播放器识别并直接播放了。有关的具体信息可自行检索。

4. 项目描述 4 中的接口的一种可能的设计如下：

```
moviedb.all().byRate(less_than=8).setRate(8)
moviedb.all().byCast("周星驰").getAverageRate()
moviedb.all().byWriter("虚渊玄").getMostType(count=3)
moviedb.all().byCast("施瓦辛格").byType(
                "动作").sort(by="rate").savePlaylist(input("Filename: "))
```

只是举例说明，并不强制要求按这种方式设计，但应尽可能将接口设计的清晰灵活。

另外，在完成这一项目时需要了解一些处理中文的技术，处理网页时还可能遇到 bytes 和 str 的区别和转换方式。这些问题需要自己设法解决，作为作业内容。请自己设法弄清相关的技术（例如查看系统文档，查阅相关书籍或检索网络等）。