# 第四阶段 Python数据分析

# Python 数据分析库1

## 网校相关视频

## 明确本次课知识点，明确重点难点

### 【知识点目标】

* Pandas简介
* Series对象基本操作

### 【重点】

* Series对象基本操作

### 【难点】

* Series对象基本操作

## 复习巩固作业讲解

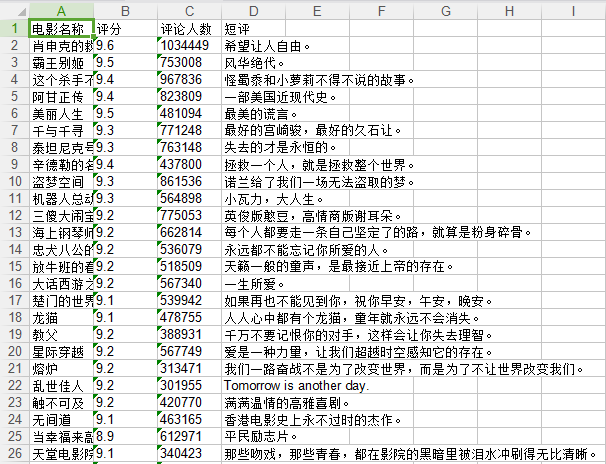
## 本次课程任务讲解

### 【课程案例】使用requests和Beautiful Soup抓取数据

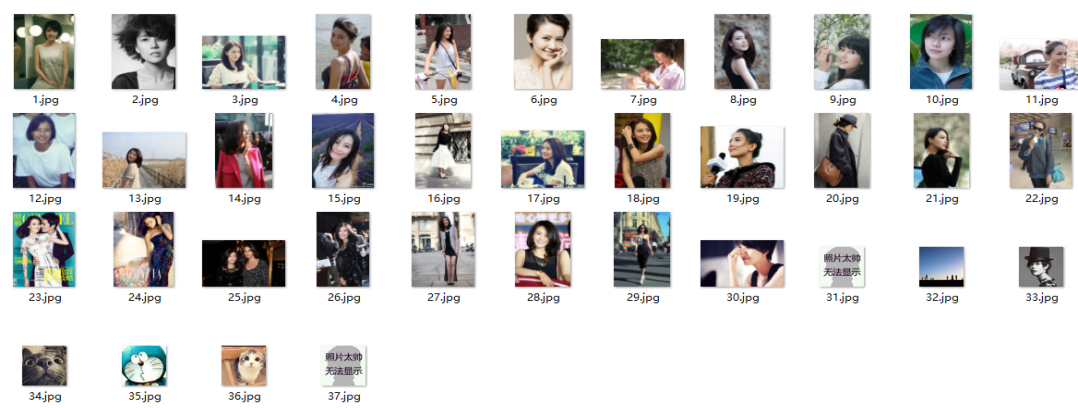
1. 爬取豆瓣电影top250的信息，并写入EXCEL

<https://movie.douban.com/top250>





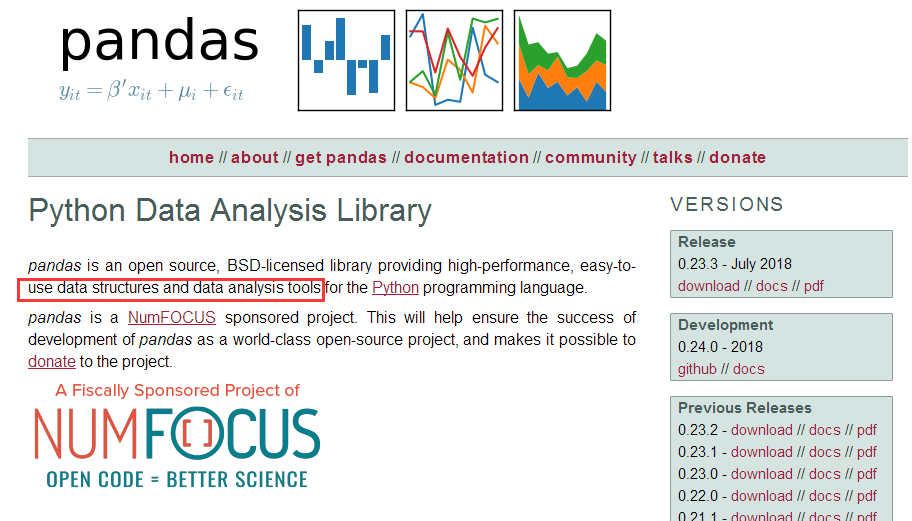
1. 爬取豆瓣艺人信息中的高圆圆照片，并存入指定文件夹



### 【知识点1】Pandas简介

1. Pandas库介绍

<http://pandas.pydata.org>



Pandas是一个Python软件包，提供快速，灵活和富有表现力的数据结构，旨在使“关系”或“标记”数据的工作变得简单直观。

Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具以及大量能使我们快速便捷地处理数据的函数和方法。

Pandas是建立在Numpy之上，并与其他第三方库在科学运算环境中有较好兼容性的 Python包。其最大特色在于提供灵活直观的数据结构来处理关联数据和有标签的数据。

Pandas提供了两大数据结构：一维结构的Series和二维结构的DataFrame。这两大数据结构都有数据标签这一独特性质，数据标签能够对不同变量的数据按照相同数据标签进行汇总，为多个变量的特定数据元素提取、变量合并等操作分析提供了极大的便利。因而，Pandas被广泛运用于金融、统计以及社会科学领域的数据处理。

Pandas基于NumPy实现，常与NumPy和Matplotlib一同使用。

### 【知识点2】Series对象类型

1. Series对象简介

Series类型的数据由一列数据及与之对应的标签(索引，位于数据的左侧)两部分组成。

Series对象本质上是一个NumPy数组，因此NumPy的数组处理函数同样适用于Series对 象。每个Series对象实际上都由两个数组组成，具有index和values两大属性。

•index:保存标签信息，是从NumPy数组继承的Index对象；

•values:保存值，是一维NumPy数组对象。

Series对象看起来由两列数据构成，实则是一维数据结构。

1. Series对象的创建

Series对象的创建通过Series类的构造函数Series()函数来实现。在创建Series对象之前，要先导入pandas包，我们一般使用简写pd来代指pandas

创建一个Series对象时, 可以同时为其index属性和values属性赋值，如果没有对这两个属性赋值，则创建一个空的Series对象。

* 1. 同时指定index属性和values属性的具体值
  2. 仅制定values属性的取值

1. Series对象的元素提取与切片
   1. 调用方法提取元素
      1. head
      2. tail
      3. take
   2. 利用位置或标签提取元素与切片
      1. 提取元素

Series对象与一维数组最大的不同在于其可以指定标签作为索引值。数组只能通过元 素所在位置的索引来提取元素，而对Series对象来说，除了位置索引以外，更方便更有特点的是通过标签索引来提取元素。

* + 1. 切片

Series也支持标签切片和位置切片。两种切片方式有一定的区别：

•标签切片同时包括起始标签和结束标签位置；

•位置切片遵循Python的切片规则，即包括起始位置，但不包括结束位置。

* 1. 时间序列的操作
     1. 创建时间序列
     2. 截取时间段数据
     3. 滞后或超前操作

## 自主学习作业讲解

无

## 课程总结

#### Pandas简介

#### Serires基本操作

## 下次自主学习任务布置

作业：爬取51job上的数据分析师职位信息，将爬取到的内容写入EXCEL

