数据分析与实践实验报告

姓名：陈鸿绪　　学号：PB21000224　　完成日期：4.1

实验题目：数据获取与管理实验（实验 3：王者荣耀网站数据爬取）

具体要求：给定数据网站 http://db.18183.com/wzry/，需要设计一个网站遍历策略，爬取 50 个英雄的详细信息，记录与 json 文件中，爬取的英雄属性数据在以下一类网页中全部有展示：

需求分析：要求利用 python 语言，需要对每个英雄的基础属性和星级属性进行数据爬取，爬取后将数据写入 json 文件。爬取过程中需要遵守该网站对于爬虫行为的规范准则。爬虫程序需要具有稳定性，对爬取所用时长需要进行一定的优化。

程序实现：

1. 首先导入爬虫相关的模块：re 模块、json 模块、BeautifulSoup 模块、urlopen 模块。
2. 程序利用 urlopen 函数打开 http://db.18183.com/wzry/，生成 prime_html 对象，再对 prime_html 使用 BeautifulSoup 函数生成 prime_bsobj 对象

3. 对 prime_bsobj 对象使用 findAll 方法，主网页 html 代码中寻找"href"形为'\/wzry\/hero\/[0-9]*\.html'的<a>tag，创建关于其的对象列表，这些都是每个英雄具体所在网页链接内容。创建打开 json 文件 data.json。
4. 打开第 i 个英雄的网页，获取该英雄的名字，创建该网页的 beautifulsoup 对象 temp_bsobj。对每个基础属性，如最大生命，可以通过关键词“最大生命”利用 temp_bsobj 的方法 find 查询含有关键词“最大生命”的 tag 内容，对该关键词进行正则表达式处理，利用 re 库中 findall 函数得到该属性后的具体数值。对星级属性，例如生存能力，该属性的具体值在网页代码中处于tag 的 class 属性中，对 temp_bsobj 利用 findAll 函数寻找 class 属性正则表达式型为"star star-[0-9][0-9]*"的tag，并提取其中的数字即可。最后将所有属性以键对值形式写入字典 dict_，并利用 str 函数将其转化为 str 对象，利用 json 中 dumps 将该 str 对象转化为 json 对象，写入 json 文件。
5. i 加一，如果已经到达 50，则结束，否则返回至 4 步骤

程序运行测试：网速正常情况下，对 50 个英雄的爬取时间在 40 秒以内，爬取内容在 data.json 文件中，经过对照，检验了爬取数据的正确性与 json 文件格式的正确性。