**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<span>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<span>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。