**雷速体育足球资料库数据爬取**

根据对网站页面结构分析，得出足球资料库结构分为两种，国际赛事以及各大洲组织的赛事页面结构一致（如世界杯、欧洲赛事等）为一种，各大洲国家级赛事页面结构一致（如英格兰、意大利等）为另一种；并且网站的url后缀都是由唯一的id拼接而成的，如联赛id 、赛事id、球队id、球员id。

**爬取思路**：

1. 每一个联赛都有自己唯一的链接，在源代码中即可找到，获取到联赛的名称和链接，存到redis数据库中，作为爬虫的url仓库使用（可以根据需求按照地域分，或者页面结构分）。
2. 获取到联赛的url进行拼接，访问该联赛页面，如丹麦中的丹麦超，发现今年的赛事数据都在里面。(以下都以丹麦超为例)。
3. 设计数据存储结构，以及爬取逻辑，定义好爬取数据的字段，第一层首先爬取:
4. 丹麦超中的所有球队数据，以及对应球员数据，根据每个球队和球员的固定id数字，拼接url 获取。
5. 因为涉及到年份问题，所以需要在对赛事年份进行获取，每个年份都有一个唯一的的连接和一个相对应的数据页面，获取之后拼接，依次访问并爬取所需数据。
6. 这时已经进入到相应年份的页面，第二层爬取的数据：
7. 联赛信息（球队数量、球员信息、市值等）。

(2) 对应的赛程赛果（至于里面的数据分析和数据指数，因为每场比赛都有固定的id值，可以通过id值进行url拼接，另行存储，通过id关联查询，建议使用mysql数据库）。

(3) 积分榜（能够直接爬取总积分，若想要每一场的积分在赛果里面能够计算得出）。

(4) 球队球员数据（射手榜、助攻榜、球员防守、球队数据）。

5、设置logging日志模块来对爬虫进行监听，记录程序运行状态。

6、如果想做到实时更新，另写一爬虫，获取未完成比赛的时间，转化成时间戳进行判断，到时间后进行数据抓取，没有比赛的时间段内sleep。

