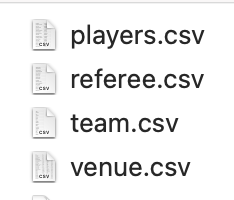
**最终目的：给每一个球员，每一个球队，每一个球场，每一个裁判的Wikipedia的内容**

**分成两步：**

1. **找到所有wikipedia的对应网址**
2. **对于每个wikipedia爬取内容，生成一个python dict，后续我来操作就行。**
3. 找到所有wikipedia的对应网址

我会给您一些csv文件，需要你在这些csv文件里面都生成新的一列，标题行是”wiki\_link”，里面的内容就是这一行的内容对应的wikipedia的链接。



这四个文件分别是足球里面的球员、裁判、球队、足球场的名称，在这一步需要您先确认每一个词条对应的链接。

为了确保搜索的链接是正确的，可以加一些信息：

* 在搜索player的时候，就搜索第一列Full Name里的内容，建议在搜索词条的时候加上football player，比如”Leo Stulac”这位球员，可以在搜索的时候搜索”Leo Stulac football player”，这样子一般就能找到，
* 在搜索referee的时候，同样的，就在referee的detected\_name后面加上一个“referee”去搜索，准确率会比较高。
* Team的时候就在后面加一个“football club”来搜索
* Venue就不用加了，直接搜索就行。

总的来说，希望提高搜索到正确的对应的wiki链接是第一要务，您可以根据我上面的建议做一些微调。

在完成第一步以后，把所有这些包含有wiki链接的csv文件重新保存下来，进行后面第二步的操作。

1. 找到所有wikipedia的对应网址

这一点的要求我有一定的预期，就是需要分门别类的爬去网页里面的内容。但是具体怎么爬，能不能全部都爬，需要根据情况而定，比如梅西的wiki界面为例：<https://en.wikipedia.org/wiki/Lionel_Messi>



您需要帮我生成一个python dict，形式大概是

{

“summary”: ………,

“meta-info”: ………,

“early career”: ………,

……

}

这种感觉。我期望得到的是：首先我需要一个summary，就是最前面这些总结的话（不要超链接）、我还需要右边这个卡，记录了他的很多信息什么的，把它作为meta-info存为一个dict。然后对于后面出现的每一个板块，我都需要爬到具体的文字内容，如果有分几个自然段的就把他们合起来作为value保存在他们小标题作为key的里面，就放在上面展示过的那个key中。

因为里面涉及到一些不同的形式，可能需要您先看一看，不同的形式，方便怎样进行杂糅，比如，里面会有很多表格，我希望您通过一些方法准确的爬出来，而不是爬下来以后变成一团乱乱的叠在一起的信息。总之，就是要表达出表格本身表达的意思即可。

还有一个额外要求，我希望把wiki里面出现的图片，都给我留下来一个链接。不管您以什么方式保存，我需要得到的是图片的下载地址，和图片的名字。比如梅西的封面图<https://en.wikipedia.org/wiki/Lionel_Messi#/media/File:Lionel-Messi-Argentina-2022-FIFA-World-Cup_(cropped).jpg>点开来以后就有Lionel Messi playing for Argentina at the 2022 FIFA World Cup.这样一个标签，应该也是嵌入在网页信息里。在wiki里面出现的所有图片内容，我都需要留存记录：他们在哪个词条出现、网址是什么，名字是什么。

这个项目时间不是最关键的因素，价格也不是，上次的合作比较愉快，我们可以方便讨论价格。我的期望是我们能够在你研究这个爬虫的形式的时候，我们能保持交流，确保不会出一些差错。价格、时间，我这里相对限制不多，导师会报销，质量是最重要的。麻烦您啦，有空我们保持联系！