17/6/3

完成功能：当前比赛日结果抓取，前一比赛日结果抓取

已解决问题：bs4抓取的话没什么问题，但是需要点击向前按钮的话，就要用selenium抓取

selenium包含webdriver，职责是抓取和动态地去点按钮，webdriver需要与浏览器结合，来请求网页内容，代码中使用phantomJS，它没有图形界面，是轻量的浏览器。

webdriver的使用中遇到问题。首先是为webdriver绑定phantomJS浏览器。首先弄明白了，executable\_path要输入浏览器exe的路径（包括exe本身），一般方法是把浏览器执行文件目录加入系统变量，或者是把浏览器执行文件拷到python的相关目录下，配置python的时候要加入系统路径变量path的那个。

这里我考虑到最好还是能够在不同的计算机上能够便捷运行的问题，把phantomJS放到脚本同目录下，用path = os.getcwd()+'\phantomjs.exe'来指定phantomJS的位置。

其次是在webdriver中获得相关按钮、相关文本信息。按照bs4的方法指定会报错。Compound class names not permitted。

http://blog.csdn.net/Jywangkeep\_/article/details/45568723 (要注意这个是java代码)

根据上面的文章，发现是webdriver无法识别带空格class，会认为是复合class不支持。解决办法是用CSS\_SELECTOR来迂回，把空格换成点。这个挺不方便的。要区分复合class么让用户传个链表不就好了，不允许有空格很困扰。

driver.find\_element(By.CSS\_SELECTOR,".flecha\_jornada\_anterior.pointer.flecha\_fecha")

之后不小心搞错了按钮位置，让代码点另外个按钮去了，结果抓取输出的没有变，因为这样网页刷新的部分不是抓的部分。

下一步工作：这样的话干脆就弃用bs4，全部用selenium来抓，因为selenium必要，bs4可以替代。接下来要抓前一比赛日和后一比赛日，前一比赛日和后一比赛日都有可能不存在，所以要考虑到这个情况，加入处理的代码。将抓下来的信息最好放进结构的列表中，方便使用。

17/6/4

完成功能：本比赛日、前一比赛日、后一比赛日的抓取

已解决问题：用异常处理的方式防止第一比赛日和最后比赛日的情况无法点击向前和向后箭头抓取。放弃bs4，全面使用selenium和webdriver。使用github做了版本控制。

未解决问题：积分榜没有办法抓取。初步研究，我认为是表格中的类型在css中表现方式不同，导致没法用by.cssSelector来筛选。

下一步工作：解决没有办法抓取积分榜的问题。抓不到名次的话，想办法按名次顺序抓取各队，自己加上名次。

17/6/14

完成功能：意甲球队排名、本比赛日的抓取

已解决问题：1、意甲的官网在家抓取时会发生自动跳转，先访问一个赌球网站，然后再跳转到排名页面。这个问题导致抓取不到排名表格，再载入页面之后再time.sleep(6)能够解决问题

2、意甲官网具体的显示排名和队名比分之类的元素有的没有指定类和类名，导致难以抓取，我选择了抓取上级元素，从它的text属性里提出包含的所有文本，再进行筛选处理。

3、python选择列表元素用[]，函数调用用()，和matlab的语法容易混淆

未解决问题：抓取意甲的前后比赛日数据，意甲官网没有向前和向后的按钮，成为一个问题。

下一步工作：抓取前后比赛日，之后尝试存入sql数据库以供网页和远程应用访问。

17/6/15

完成功能：意甲前一比赛日抓取

已解决问题：1、通过寻找active标签找到当前比赛日是第几比赛日，再通过xpath寻找上一级，再在上一级中搜索box\_Ngiornata\_andata标签（上半程）和box\_Ngiornata\_ritorno标签（下半程）组成的列表，再通过active标签的text属性指出的比赛日加减得出上下比赛日对应在列表中的哪一个。

未解决问题：1、由于比赛日列表分上下半程，链接获取逻辑复杂化了。要抽出逻辑，将当前比赛日先算出，写出前一比赛日后一比赛日的标签获取函数。再调用函数，点击函数返回列表（其实长度为0或1）中的标签。根据当前比赛日是第几天来点击。（1和38及其他）

下一步工作：写出前一比赛日后一比赛日的标签获取函数。抓取前后比赛日，之后尝试存入sql数据库以供网页和远程应用访问。

17/6/19

完成功能：？

已解决问题：把获取前一比赛日后一比赛日的逻辑和打印比赛信息的逻辑封装到函数中，但还没有好好测试。

下一步工作：测试新的脚本，之后尝试将数据存入sql数据库以供网页和远程应用访问。

17/6/22

完成功能：调试前后比赛日获取函数

已解决问题：调试前后比赛日获取函数

下一步工作：尝试将数据存入sql数据库以供网页和远程应用访问。

17/6/24

完成功能：建立MySQL的新用户，建立MySQL里的爬虫数据库并转换成unicode字符集，建立表，将队伍数据存入MNySQL。

已解决问题：将队伍数据存入MySQL。MySQL中表的某一列即使是INT，用python插入的时候，命令中也应该用\”%s\”作为占位符。另外mysql需要存储操作结束后commit()才生效

未解决问题：由于添加了MySQL的部分，需要为不同使用者或不同抓取主机初始化MySQL的指导或者输出以初始化的MySQL爬虫数据库以供导入，或者需要在github上建立readme，介绍MySQL中的表结构。

下一步工作：尝试将本轮、上轮、下轮比赛结果存入数据库。

17/6/25

完成功能：将西甲、意甲数据存入数据库

已解决问题：西甲向前翻页后会抓取空标签，把他们筛选掉了。西班牙语的文字存入mysql后在控制台里查看，没法显示拉丁注音文字，显示乱码。所幸在python中取出数据后没有变成乱码，是正常状态。怀疑可能是windows控制台不支持unicode。

未解决问题：输出mysql

下一步工作：导出mysql，接下来要迁移到新电脑上了。再用asp.net搭建一个网站，显示数据，再写一个c#的工具，方便为mysql中存储的外文队伍名字对应中文名字以供中文读者查看。