网络爬虫期末作业

1. 内容

以“2022年国际足联世界杯”为主题，设计数据采集与分析的代码，完成相关数据的采集与分析，按要求从数据和文本中提取目标内容，具体提取内容见第二部分报告要求。

该次作业允许组队完成，组队人数1-4人不限

作业的提交内容包括：

1. 数据采集与分析的源代码，请使用一个命名为源文件的文件夹存储
2. 具体的分析报告，直接存储为一个word文件
3. 相关的采集结果，使用一个命名为采集结果的文件夹存储
4. 分组信息，可以用word文件也可以使用txt文件，介绍小组成员及简要分工

请打包成以小组所有成员名字命名的压缩文件提交

1. 报告要求

报告内容分为三部分

1. 数据采集与分析
2. 数据采集的方式

介绍你将使用什么方式采集这些数据，要求至少采集15个资源文件。

1. 数据采集的结果

介绍采集到的资源文件的格式，资源文件中主要包含哪方面的信息。

1. 基本要求

从采集的文件出发，分析出当届世界杯在哪举办，冠亚季军分别是那支球队，请介绍你从文件中提取出这些信息的方法，并展示对应的程序执行结果

1. 可选要求：

从你采集的文件出发，从中提取世界杯淘汰赛阶段所有比赛的对阵情况，并制成如下所示的对阵表



对阵表中包含每一轮的对阵双方及最后比分

请在实验报告的对应部分介绍你提取这些对阵表的方式，并展示程序提取结果之后的对阵表，

1. 程序介绍

1.介绍源文件文件夹中每个源文件的具体作用

2.介绍每个源文件中声明的函数，介绍方式参照数据挖掘作业中要求的介绍方式（即数据挖掘书上283页表8.4），但请在该方式的基础上，加上一列用于记录函数编写者。

3. 介绍代码的运行方式以及输出结果中各部分的作用是什么（有中文描述的可以不用介绍作用），如在cmd中使用什么命令运行代码或是在IDE中从哪一部分开始执行，提供相关的截图介绍。（如选择完成附加要求，请将附加要求和基本要求的执行结果分开展示）

1. 小组成员分工介绍（如小组成员只有一人该部分可省略）

介绍小组各成员在作业中完成的内容，包括分析报告部分及代码部分的工作

1. 评分标准
2. 采集数据的方法越通用，评分越高
3. 数据分析的方法越通用，评分越高
4. 前两项结果类似条件下，小组成员越少，分数越高
5. 不得互相抄袭
6. 小组成员分工需合理，不得出现小组分工中一位成员工作超过另一位成员工作两倍的情况
7. 报告中如需展示多行代码请以截图展示，报告文字不得少于800字，
8. 请按照附件中的格式重新创建报告文件，不要在本文件下直接填写

请各位同学在6月19日前完成

**报告的基本格式：**

1. 数据采集与分析：
   1. 数据采集的方式
   2. 数据采集的结果
   3. 提取冠亚季军的方式
   4. 提取淘汰赛对阵表（可选）
2. 程序介绍

2.1 程序文件介绍

2.2 各程序文件的函数介绍

2.3 程序执行方式及执行情况

2.3.1基本要求

2.3.2附加要求（可选）

1. 分组介绍

介绍小组各成员在作业中完成的内容，包括分析报告部分及代码部分的工作