# 2022 年“泰迪杯”数据分析技能赛

# A 题

# 竞赛作品的自动评判

## 一、背景

在各类学科竞赛中，常常要求参赛者提交 Excel 或/和 PDF 格式的竞赛作品。本赛题以某届数据分析竞赛作品的评阅为背景，要求参赛者根据给定的评分准则和标准答案，使用 **Python 编程**完成竞赛作品的自动评判。

## 二、目标

1. 使用 Python 解压压缩文件，从中读取指定的文件。

2. 使用 Python 解析 PDF 文件，获取其中的图片信息。

3. 使用 Python 解析 Excel 和 PDF 文件，对数据进行处理与统计，根据评分准则对每份作品打分，并输出报表。

## 三、任务

根据提供的评分标准及要求，对每份作品进行自动评分，并撰写报告，在报告中详细描述各项任务的处理思路、过程及必要的结果。同时，将 Python 源代码保存为 py 文件，文件名为任务编号，例如“task1_1.py”“task1_2.py”“task2_1.py”等。

**任务 1　基本处理**

压缩文件“DataA.rar”中包括所有待评分的作品，每份作品是以作品号为文件名、包含若干结果文件的压缩文件，文件格式可能是 rar、zip 或 7z。

任务 1.1　将压缩文件“DataA.rar”中的所有作品解压到当前文件夹的同名子文件夹（即以每份作品的作品号为子文件夹名）中。

任务 1.2　在当前文件夹中新建“summary”子文件夹，在每份作品文件夹中新建“image”子文件夹。

任务 1.3　判断每份作品中是否包含文件“task2_1.xlsx”“task2_2.xlsx”

“task2_3.pdf”及“task3.xlsx”，每包含一个文件得 2 分，满分 8 分。

任务 1.4　对每份作品提取文件“task2_3.pdf”中的图片，保存在“image”文件夹的“XXXX_n.png”文件中，其中“XXXX”为作品号、n 为图片在文件“task2_3.pdf”中的图片序号。提取所有作品中的图片信息，按照表 1 的格式保存在文件夹“summary”的“result1_4.xlsx”文件中。将含有作品号 A118~A120 的结果截屏放在报告中。

**表 1　result1_4.xlsx 的格式**

| 作品号 | 图片编号 | 保存路径 | 图片分辨率 | 图片文件大小 |
|---|---|---|---|---|
| A001 | A001_1 | C:/ProblemA/A001/image/A001_1.png | 600*800 | 124KB |
| | A001_2 | C:/ProblemA/A001/image/A001_2.png | 600*900 | 164KB |
| A002 | A002_1 | C:/ProblemA/A002/image/A002_1.png | 500*600 | 174KB |
| | A002_2 | C:/ProblemA/A002/image/A002_2.png | 1200*860 | 386KB |
| | A002_3 | C:/ProblemA/A002/image/A002_3.png | 600*700 | 197KB |
| A003 | 无 | | | |

**任务 2　数据分析**

任务 2.1　通用名称评分

以“criteria2_1.xlsx”为标准，根据正式登记证号对每份作品中的“task2_1.xlsx”进行匹配，按以下规则统计错误数：

(1) 对匹配的记录，判断“产品通用名称”是否一致，如不一致，错误数 s 加 1。

(2) 对“criteria2_1.xlsx”中的每条记录，查找“task2_1.xlsx”，如没有匹配的记录，错误数 s 加 1。

(3) 对“task2_1.xlsx”中的每条记录，查找“criteria2_1.xlsx”，如没有匹配的记录，错误数 s 加 1。

对错误数 s：$s = 0$ 得 15 分，$1 \le s \le 10$ 得 10 分，$11 \le s \le 20$ 得 5 分，$s \ge 21$ 得 0 分。

任务 2.2　分组标签评分

以“criteria2_2.xlsx”为标准，根据正式登记证号对每份作品中的“task2_2.xlsx”进行匹配，按以下规则统计错误数：

(1) 对匹配的记录，判断“分组标签”中的数值和顺序是否一致，如不一致，错误数 s 加 1。

(2) 对“criteria2_2.xlsx”中的每条记录，查找“task2_2.xlsx”，如没有匹配的记录，错误数 s 加 1。

(3) 对“task2_2.xlsx”中的每条记录，查找“criteria2_2.xlsx”，如没有匹配的记录，错误数 s 加 1。

对错误数 s：$s \le 5$ 得 15 分，$6 \le s \le 15$ 得 10 分，$16 \le s \le 30$ 得 5 分，$s \ge 31$ 得 0 分。

任务 2.3　读取每份作品“task2_3.pdf”中产品登记数量及排名的表格，针对每个排名判断“分组标签”和“产品登记数量”的数值与表 2 中的标准答案是否一致，每个匹配的数值得 2 分，满分 12 分。

**表 2　产品登记数量排名**

| 排名 | 一 | 二 | 三 |
|---|---|---|---|
| 分组标签 | 7 | 6 | 5 |
| 产品登记数量 | 2012 | 1501 | 1038 |

**任务 3　相似矩阵评分**

以“criteria3.xlsx”为标准，对每份作品“task3.xlsx”中的相似矩阵（以下简称相似矩阵）按以下规则进行评分。

任务 3.1　判断相似矩阵的维数与“criteria3.xlsx”中的是否一致，如一致得 5 分，否则得 0 分。

任务 3.2　公司 ID 匹配

对相似矩阵进行匹配，按以下规则统计错误数：

(1) 对“criteria3.xlsx”中的每个公司 ID，查找“task3.xlsx”，如没有匹配的公司 ID，错误数 s 加 1。

(2) 对“task3.xlsx”中的每个公司 ID，查找“criteria3.xlsx”，如没有匹配的公司 ID，错误数 s 加 1。

对错误数 s：$s = 0$ 得 15 分，$1 \le s \le 2$ 得 10 分，$3 \le s \le 5$ 得 5 分，$s \ge 6$ 得 0 分。

任务 3.3　判断相似矩阵的对角线元素是否均为 1（允许误差 $10^{-6}$），如均为 1 得 5 分，否则得 0 分。

任务 3.4　判断相似矩阵的元素是否关于主对角线对称（允许误差 $10^{-6}$），

每出现一组不对称的元素，错误数 s 加 1。对错误数 s：$s = 0$ 得 10 分，$1 \le s \le 5$ 得 7 分，$6 \le s \le 10$ 得 4 分，$s \ge 11$ 得 0 分。

任务 3.5　以“criteria3.xlsx”为标准，统计相似矩阵上三角元素的错误数 s。对每个匹配的元素计算绝对误差 e，$0.005 \le e < 0.01$，错误数 s 加 0.5；$e \ge 0.01$，错误数 s 加 1。对每个匹配不上的元素，错误数 s 加 1。对错误数 s：$s = 0$ 得 15 分，$1 \le s \le 5$ 得 10 分，$6 \le s \le 10$ 得 5 分，$s \ge 11$ 得 0 分。

**任务 4　评分结果汇总**

对每份作品，完成任务 1~3，对得分进行统计，按表 3~5 的格式生成“score.xlsx”，保存在文件夹“summary”中。

任务 4.1　按表 3 的格式，在工作表“report4_1”中保存所有作品的得分明细，计算各作品的总分，并进行排名。按总分降序排列。作品号放在第 1 列，字段名放在第 1 行，从第 2 行第 2 列开始列出明细数据。字段名单元格底色设置为蓝色，总分与排名单元格底色设置为红色，其他单元格底色设置为黄色。

**表 3　作品得分明细汇总表**

| 作品号 | 任务 1.3 | 任务 2.1 | 任务 2.2 | 任务 2.3 | … | 任务 3.5 | 总分 | 排名 |
|---|---|---|---|---|---|---|---|---|
| A010 | 8 | 15 | 10 | 12 | … | 10 | 90 | 1 |
| A067 | 4 | 15 | 10 | 12 | … | 15 | 86 | 2 |
| A035 | 8 | 10 | 5 | 12 | … | 10 | 80 | 3 |
| A091 | 8 | 5 | 15 | 8 | … | 15 | 78 | 4 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |
| 最高分 | 8 | 15 | 15 | 12 | … | 15 | 88 | |
| 最低分 | 0 | 0 | 0 | 0 | … | 0 | 0 | |
| 平均分 | 6.1 | 9.7 | 9.4 | 7.8 | … | 10.2 | 67 | |

任务 4.2　在明细数据之后列出每项评分的汇总数据：最高分、最低分和平均分。汇总数据单元格底色设置为绿色。将含有作品号 A417~A419 及统计结果的截屏放在报告中。

任务 4.3　在工作表“report4_3”中，以 10 分为一段（参见表 4）分段汇总各段的作品数。得分区间在第 1 列，作品数在第 2 列，字段名在第 1 行。同时，将结果放在报告中。

**表 4　总分得分区间数量汇总表**

| 得分区间 | 作品数 |
|---|---|
| 0~9 | 2 |
| 10~19 | 7 |
| 20~29 | 3 |
| 30~39 | 3 |
| ⋮ | ⋮ |
| 80~89 | 17 |
| 90~100 | 3 |
| 合计 | 100 |

任务 4.4　基于任务 4.3 的结果，在工作表“report4_4”中绘制各得分段作品数的柱形图。同时，将柱形图放在报告中。

**任务 5　嵌套压缩文件处理**

任务 5.1　将压缩文件“DataB.rar”中的所有作品解压到当前文件夹的同名子文件夹中。

任务 5.2　判断每份作品中是否包含文件“task2_1.xlsx”“task2_2.xlsx”“task2_3.pdf”及“task3.xlsx”，每包含一个文件得 2 分，满分 8 分。按照表 5 **错误!未找到引用源。**的格式，将所有作品的得分保存在文件夹“summary”的“result5_2.xlsx”文件中。

**表 5　result5_2.xlsx 得分表的格式**

| 作品号 | 得分 |
|---|---|
| B001 | 4 |
| B002 | 6 |
| ⋮ | ⋮ |

任务 5.3　基于任务 5.2 的结果，对每份作品提取到的文件，按照表 6 的格式列出各文件的路径，保存在文件夹“summary”的“result5_3.xlsx”文件中。

**表 6　result5_3.xlsx 的格式**

| 作品号 | 文件名 | 路径 |
|---|---|---|
| B001 | task2_1.xlsx | C:/ProblemB/B001 |
| | task2_2.xlsx | C:/ProblemB/B001/result2 |
| | task2_3.pdf | C:/ProblemB/B001/result2 |
| B002 | task2_2.xlsx | C:/ProblemB/B002/Result2 |
| | task2_3.pdf | C:/ProblemB/B002 |

## 四、关于竞赛成果提交的说明

### 1. 登录方式

请使用**队长**的账号登录数睿思网站（www.tipdm.org），进入第五届技能赛页面。为保证成功提交，**请使用谷歌浏览器无痕模式**。

### 2. 作品提交

报告以 PDF 格式提交，文件名为“**report.pdf**”，要求逻辑清晰、条理分明，内容包括每个任务的完成思路、操作步骤、必要的中间过程、任务的结果及分析。

### 3. 附件提交

3.1　将任务 1、任务 2、任务 3、任务 4、任务 5 的 **Python 源程序**分别保存到“**program1**”“**program2**”“**program3**”“**program4**”“**program5**”文件夹，然后存放到“**program**”文件夹中。

3.2　将任务 1、任务 2、任务 3、任务 4、任务 5 所得到的结果文件存放到“**result**”文件夹中。

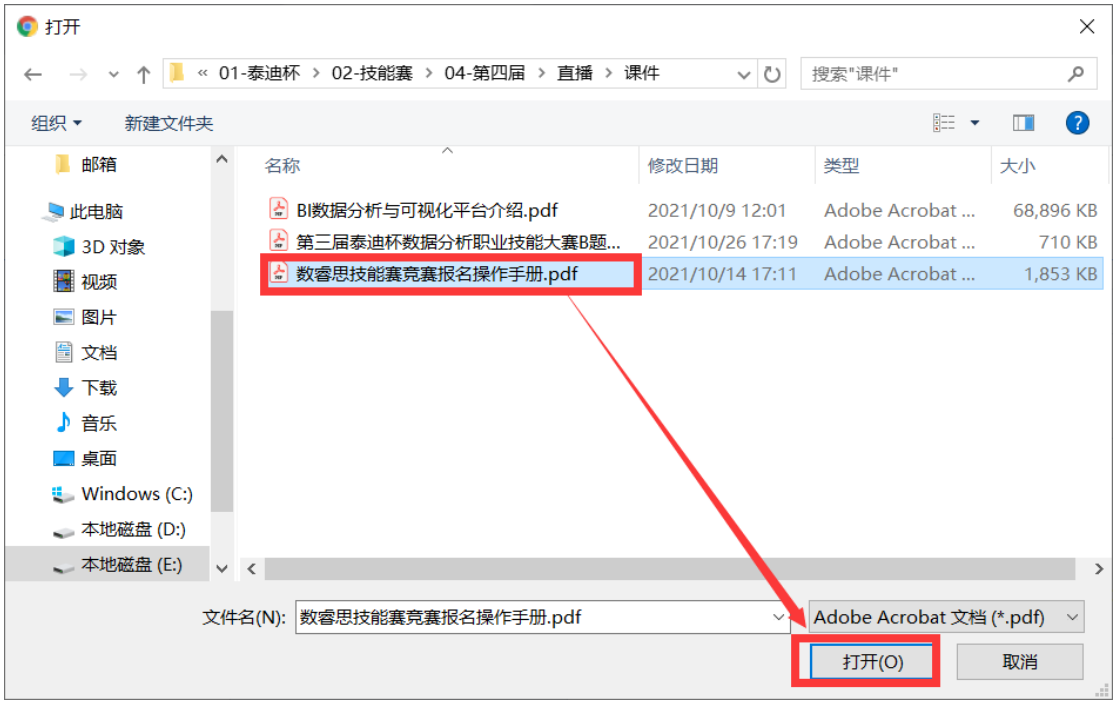
3.3　将程序文件夹“**program**”、结果文件夹“**result**”以及报告的 **Word** 版本打包成“**appendix.zip**”，作为附件提交。

### 4. 提交界面

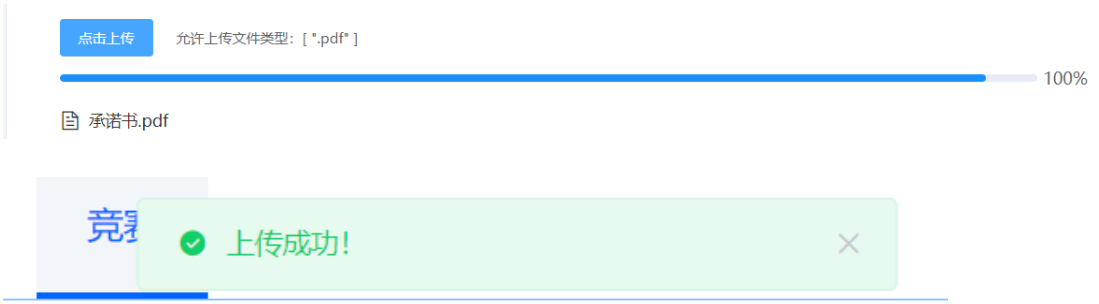
4.1　找到赛题提交入口。

4.2　点击“点击上传”按钮。


4.3　选择需要上传的对应文件，点击“打开”。



4.4　进度条加载完成后会有“上传成功”提示。




4.5　页面如下图即为上传提交成功，多次提交会以最后一次为准。