**南京大学《数据科学基础》**

**课程大作业说明**

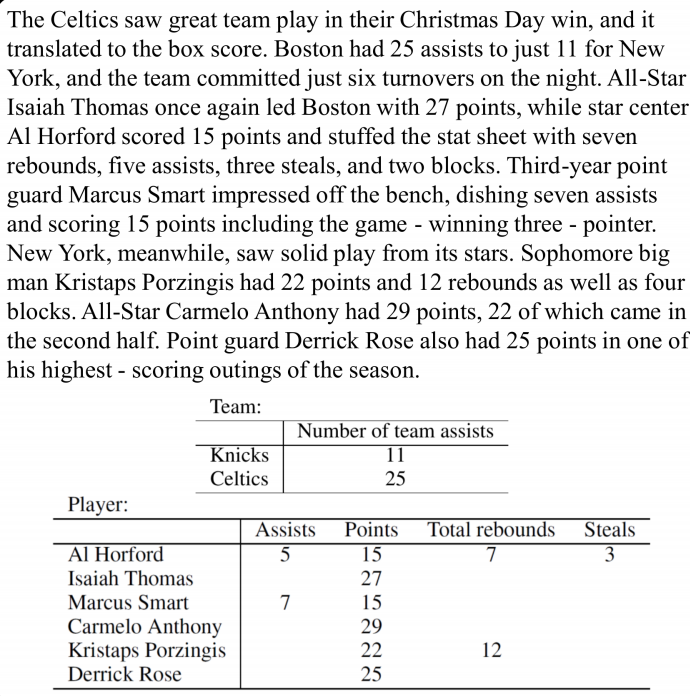
**——基于text生成的table数据分析**

### 任务目标

对于我们给出的text，请将其转换为table，并使用课程上学习到的知识，对table进行可视化分析。

### 任务内容

现实生活中，text数据非常多样丰富，对其进行分析处理是一项非常重要的工作。有一种思路是将text数据转换为table数据，text-to-table是一种自然语言处理技术，它可以将自然语言文本转换成结构化的表格数据。



我们将给出一个**数据集rotowire**。Rotowire数据集它里面每条文本都描述了一场足球比赛，如上图所示。你需要调用或者自己实现一个将text转换为table的NLP模型，将rotowire数据集中的文本转换为table。

转换好的table是非常好的结构化数据，你需要对其进行分析处理工作。请使用课程上所学到的数据分析工具对其进行分析和可视化。例如：**python 、excel、R语言、SPSS**。

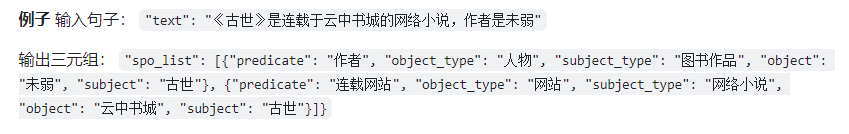
### 任务要求

此任务需要3人组队完成。总的来说，请尽可能地从text数据中挖掘出结构化信息组成table。在table中，善用课程上学到的数据分析工具，对table数据进行分析，并且将数据分析结果进行可视化。

任务可以分为2个阶段：第一阶段为text-to-table转换任务（40分），第二阶段为table分析任务（60分）。

**text-to-table阶段**：

1. 你可以调用开源API实现text-to-table任务，代表性开源API有**OpenAI、hugging face**，如PPT中展示的情况。
2. 你也可以在本地训练模型来完成这一任务。如果你能使用自己的模型完成，你将获得更高的分数。
3. 如果你觉得直接实现text-to-table模型具有一定困难，也可以将这个任务简化为文本信息抽取任务。你可以自己定义schema约束，使用schema定义关系P以及其对应的主体S和客体O的类别，然后本地训练一个模型提取文本信息中的实体关系，得到实体关系之后对其进行手动创建表格。

数据集rotowire中包含728条文本，请根据你想要进行的分析需求至少从rotowire中选择20条文本转换为table。得到的table请用excel格式存储，需提交excel源文件。

**table分析阶段**：请回顾课堂上所学到的数据分析知识，并使用至少2种方式实现5个分析内容。分析内容请用文字详细说明，并且分析内容需要可视化。

可以从player以及team等角度进行分析，提供一些思路（并不限制按照以下思路进行分析）：

1. 根据球队和球员的信息，比较不同球队和球员在得分、助攻、盖帽等指标上的表现。可以使用柱状图或折线图来进行对比和分析。
2. 根据时间信息，分析比赛中各个指标的变化趋势，例如每场比赛的得分平均值或总得分等。
3. 根据得分、助攻或其他指标，计算球队和球员的排名和排行榜，以识别表现出色的球队和球员。

#### 在本地训练模型完成text-to-table任务的一些提示

**训练深度学习模型对电脑配置有一定的要求，在电脑配置不高的情况下可以选择使用免费GPU开放工具，例如Google Colab。**

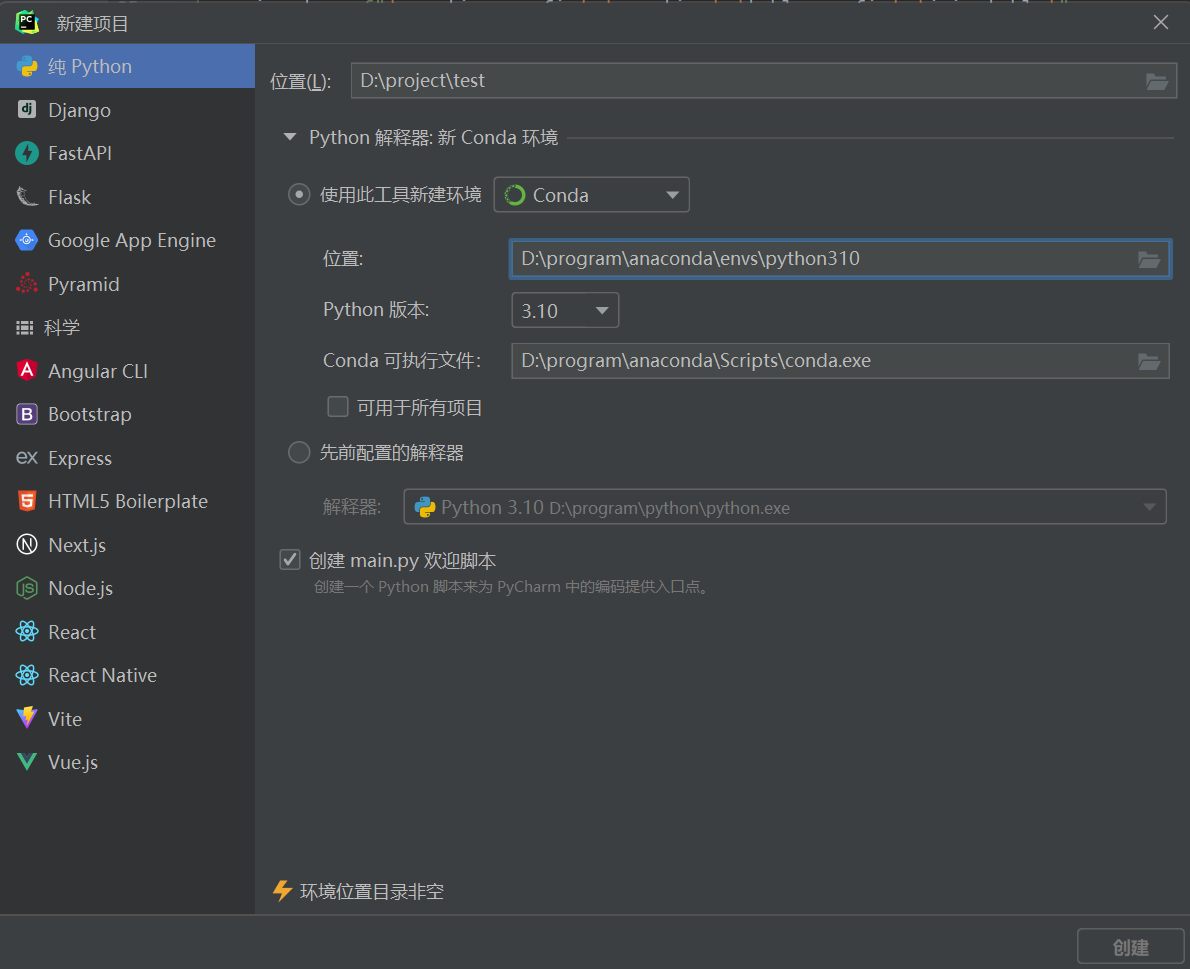
目前，深度学习模型已经成为文本转表格领域的主流方法。以下是一些常用的深度学习模型：

1. Transformer模型：Transformer是一种使用自注意力机制进行序列建模的神经网络模型，可以处理长序列数据并具有良好的泛化能力。一些成功的文本转表格模型，例如Google的TAPAS和Facebook的TAT，都是基于Transformer模型的。
2. BERT模型：BERT是一种基于Transformer模型的预训练语言模型，可以用于文本分类、命名实体识别、问答等任务。在文本转表格任务中，可以使用BERT模型进行文本向量化和表格数据生成。

无论是Transformer模型还是BERT模型，它的原理都是提取文本信息中的实体与关系。再将提取到的实体关系建立生成table。正如上面任务要求阶段提到的，如果你觉得直接实现text-to-table模型具有一定困难，也可以将这个任务简化为文本信息抽取任务。你可以本地训练一个模型提取文本信息中的实体关系，得到实体关系之后对其进行手动创建表格。

如果你需要在本地训练模型，需要安装python解释器，并推荐使用anaconda来管理python解释器。现介绍在pycharm里面使用anaconda环境的python解释器，请确保已成功安装anaconda，以Windows系统为例：

1. 进入cmd，命令“conda --version”，查看安装的conda版本，确认已成功安装。
2. 进入pycharm选择新建项目，在位置中输入项目路径与项目名，python解释器选择使用conda新建环境，在python解释器位置输入解释器名字，例如python310。(这一步是使用conda新建了一个python环境，并且pycharm新建的项目使用此python环境)



1. 进入cmd，命令“conda env list”可以看到conda所管理的所有python解释器环境，每一个conda都有一个base环境。此时conda下应该有刚刚创建的python310环境。
2. 命令“conda activate python310”激活python310环境。

PPT中给出了一个开源的text-to-table模型**https://github.com/shirley-wu/text\_to\_table**，可以参考此模型。

### 提交材料

中期提交材料（6月1日晚上20：00截止）：

1. 文档：请在文档中描述组内分工，所选择的text-to-table任务的方法，挑选的文本为哪些（**至少挑选20条文本**），准备从什么角度进行分析以及其他已完成的工作。

最终提交材料（7月1日晚上20：00截止）：

1. 文档：请在文档中详细描述自己的贡献和工作量。包括将text转换为table的具体过程，是使用的开源API或者自己训练的模型。如果是自己训练的模型，请详细描述训练过程和使用的数据集。对于转换成功的table，请对其进行分析处理，**至少使用2种分析方式得到5个分析结果**，并给出文本详细描述相应过程和分析结果的可视化。
2. 开源链接：如果是自己训练的模型，请给出训练好的text-to-table模型的开源链接。
3. Excel格式的table文件：文本转换得到的table。
4. 演示视频：整个任务过程的演示视频，即如何从text得到table和如何对table进行分析。请将视频上传到b站，给出b站链接，不要直接发mp4文件。

提交方式：邮箱：522022320046@smail.nju.edu.cn

### 评分标准

1. Text-to-table任务的准确性和精确度与工作量
2. Table分析结果的多样性和信息挖掘程度