

## 第四届杭州电子科技大学

### “滨创杯” Python 语言程序设计挑战赛

Python是一门有条理的和强大的面向对象的程序设计语言，不仅已经成为数据分析、人工智能、金融等领域必不可少的工具，还被越来越多地公司用于网站搭建。Python 在设计上坚持了清晰划一的风格，这使得 Python 成为一门具备大数据处理能力、拥有惊人数量的库、可靠并且高效，并且被大量用户所欢迎的、用途广泛的语言。

为了提高我校学生对程序设计语言方面理论与实践相结合的能力，促进商科与计算机学科之间的交叉与融合，培养大学生团队合作能力，鼓励其勤于求知、学以致用，我校特此举办第四届Python 语言程序设计挑战赛

本次比赛由管理学院发起，并由管理学会联合杭州电子科技大学滨江创新中心负责此次大赛的所有承办工作。我们谨代表管理学院与管理学会在此感谢所有参与本次大赛的同学，同时也感谢所有对本次比赛提供支持帮助的老师。

以下为本次大赛的三道题目。请大家认真阅读题目，了解项目要求，并在符合题意的基础上发挥自己的创意，展示出队伍中每个人的能力。预祝大家取得优秀的比赛成绩。

## 题目

- 一、高速公路交通状况分析与预测.....3
- 二、水流量的预测与分析.....4
- 三、新闻人物实体对象和相关事件文本抽取.....6



## 一、高速公路交通状况分析与预测

### 【问题背景】

随着社会经济的不断增长，智能交通系统(ITS, intelligent transport system)在近几年蓬勃发展。智能交通系统主要目的是在大规模的综合交通中实时、准确、高效的感应和控制交通情况，而交通流量的准确预测是ITS的基础，也是ITS的核心研究点之一。准确的交通预测信息可以为交通管理者提供有力的交通决策依据，同时也可以让驾驶员选择更为畅通的道路出行，进而避免或缓解交通拥堵的状况。

### 【问题说明】

请根据提供的交通流量数据、速度等传感器数据，利用数据挖掘和可视化方法探索、分析、展示交通规律，并利用机器学习方法预测多尺度未来时刻（未来5min、30min、1hour、2hour、1天）的交通流量与速度。

### 【数据说明】

1. 不同的文件表示不同车检器的数据
2. 每个文件下数据的含义如下

变量	变量含义	备注
deviceId	设备Id	
property	属性	流量、车速、车长、占有率、间距等
timestamp	时间戳	5分钟一个轮次
value	取值	

3. property中各变量名含义如下

变量	变量含义
small_car_traffic_{lane}	小型车车流量_{车道}
medium_car_traffic_{lane}	中型车车流量_{车道}
large_car_traffic_{lane}	大型车车流量_{车道}
average_speed_of_small_cars_{lane}	小型车平均车速_{车道}
average_speed_of_large_cars_{lane}	大型车平均车速_{车道}
average_car_spacing_{lane}	平均车间距_{车道}
average_car_length_{lane}	平均车长_{车道}
occupancy_{lane}	占有率_{车道}
average_speed_{lane}	平均车速_{车道}
total_traffic_{lane}	总车流量_{车道}

注意：数据请勿外传，严禁用于商业用途！

**【提交标准】**

提交文件为：交通分析预测报告（pdf/word 各一份）、相关代码文档（压缩包形式上交）

备注1：最终的预测目标为total\_traffic 和average\_speed

备注2：分析代码建议以 Jupyter Notebook的ipynb 文件格式上交。

备注3：交通分析预测报告中须包括数据预处理部分

备注4：交通分析预测报告中须明确展示的评价指标包括但不限于MSE, RMSE和相对预测偏差率，相关计算公式如下

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad \#(1)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad \#(2)$$

$$\text{相对预测偏差率} = \frac{1}{n} \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{y_i} \times 100\% \quad \#(3)$$

其中 $y_i$ 表示真实值， $\hat{y}_i$ 表示预测值

## 二、水流量的预测与分析

### 【问题背景】

由于我国信息产业的迅猛发展，科技的日新月异，资源的大量开采，尤其是与人们休戚相关的水资源问题尤为凸出，因此，科学高效地利用水资源尤为重要。实现水流量的方便、快捷、精准地监测为安全供水和节水降耗提供了科学的保障和手段。

水流量监测的一项重要任务是流量预测，即根据历史数据来预测某一时刻的水流量。在此基础上，分析当前时刻水流量是否正常并对发生的异常的原因和类型进行分析，及时采取相应措施，对供水系统进行调整或通知相关维护人员现场检修，减少损失。

### 【问题说明】

要求参赛队员研发多维度时空感知的趋势分析模型，分析各种指标(水流量、水压、水浊度、水含氯量等)的变化趋势、时空规律，进而深入挖掘各种指标之间的关联关系并进行展示、分析和建模。在此基础上分析不同供水调度策略和管道设置对于各种指标的影响程度，为制定供水调度策略提供决策指导。

### 【提交标准】

提交文件包括：

- 水流量预测分析报告一份(Word/PDF均可);
- 算法说明文档和相关代码文档(压缩包形式上交);
- 爬取到的数据(如果数据自行爬取，需提交)。



## 三、新闻人物实体对象和相关事件文本抽取

### 【背景说明】

在新闻中，往往会包含对一个或多个个人物，以及对他们在新闻事件中参与情况的描述，对人物的描述存在指代、省略、交叉和时空跨度等情况。梳理人物关系和他们行为、属性和状态是理解一篇新闻的重要环节，在这基础上，我们可以进一步构建事件图谱和人物关系等知识结构。

### 【问题说明】

在一篇新闻中。我们需要对其中的人物实体进行识别，并将他们参与的相关事件描述文本抽取出来，形成围绕中心人物的事件描述。如果有多个个人物，则分别将他们各自参与的事件抽取出来。包括以下2个任务：

(1) 识别新闻中的人物实体对象

(2) 结合人物实体对象，提取围绕对象的事件描述，包括行为、状态和属性等句子或段落文本

### 【数据说明】

THUCnews 是根据新浪新闻 RSS 订阅频道2005~2011年间的历史数据筛选过滤生成，包含 74 万篇新闻文档（2.19 GB），均为 UTF-8 纯文本格式。划分出 14 个候选分类类别：

财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐。我们针对娱乐新闻进行人物实体和相关事件文本的抽取。新闻数据可以从<http://thuctc.thunlp.org> 获取

### 【提交标准】

提交相关代码和文档（压缩包形式上交），以及如下格式的抽取结果

序号	人物实体名称	新闻高编号（文件名）	事件列表（多个事件以' '进行分割
1	张黎	131611	<p>由张黎执导的电视剧《圣天门口》正在横店热拍</p> <p>让黎叔信任地将《圣天门口》中麦香这个角色交给了她（练束梅）</p>
2	练束梅	131611	<p>在片中出演重要女性角色麦香的则是青年演员练束梅</p> <p>练束梅在由张黎执导的《人间正道是沧桑》中曾有过短暂的亮相，并出演范希亮的妹妹——护士“范媛”练束梅在片中饰演天门口镇的酒馆老板娘麦香</p>
...			