2019 大数据科学导论大作业说明

- 本作业占总成绩的 40%
- 3-5 人组队，每个队伍从下面 2 题中选择一题， 请组长在 11-7 号之前在填写分组信息：https://shimo.im/sheets/kdyt8CjwjCqdpH3c/MODOC/ （组长组员信息填写学号姓名）
- spark 集群使用：同学们需在本地伪分布式环境完成初步调试（建议抽取一部分数据），如果需要 spark 集群，请发送邮件到助教邮箱申请 spark 集群帐号:wangcaimeng@nlsde.buaa.edu.cn 写清楚工作进展

1.中文垃圾邮件过滤

本题目布置在 http://contest.mooc.buaa.edu.cn/competitions/43
选择本题目的同学，停在网站注册帐号密码，参加竞赛，下载数据，提交结果进行评测。

数据说明：

- train.csv：共两列，content 为邮件内容，label 为邮件类别（1:垃圾邮件，0:正常邮件）
- test.csv：一列，content 为邮件内容

任务说明：

- 使用机器学习算法或者设置规则判断 test.csv 中的邮件是否为垃圾邮件
- 程序主体应使用 spark，可以结合其他工具

温馨提示：

- 算法：
  - 基于情感辞典、关键词等方法
  - 基于机器学习的方法：词向量+分类模型（如 tf-idf+lr）
  - 工具：
    - 邮件内容预处理: beautifulsoup4
    - 分词：jieba
  - 文本分类任务网上有大量参考内容，请自行查阅

提交方式：

在作业截止日期之前，每位同学可以多次提交作业结果，但每天最多只能提交两次。网站会实时给出同学们得分反馈，最后同学可以在网站上选择众多提交结果中得分最高的一次显示在 leaderboard 上

在网站上同学们只提交标注结果即可，要求正常邮件标注数字 0，垃圾邮件标注数据 1。每个标注结果占 1 行，以回车分割，注意，请不要有任何多余字符，请确保你已经标注了所有的邮件（标注结果行数和邮件总数相同），即最后你提交的 txt 应该是

1\n0\n0\n0\n1\n……

在大作业完成后，要求同学们提交压缩包到助教的邮箱:wangcaimeng@nlsde.buaa.edu.cn 压缩包应包含以下内容：

1. 程序代码
2. 标注结果 txt 文件
3. 此平台用户名和成绩截图
4. 小组分工情况
   郑重提示：一旦发现雷同作业，大作业按 0 分处理。

2. 滴滴出行数据分析

本题目为开始放型题目，选择该题目的同学需要使用 spark/hadoop 完成数据分析，提交分析报告

数据下载：

数据来源于滴滴盖亚数据开放计划：
https://outreach.didichuxing.com/research/opendata/
从北航网盘下载数据：
https://bhpan.buaa.edu.cn:443/link/11DE23805669FDBFC17E4C0D8C5EB721

数据说明：

2017 年 5 月 1 日至今海口市每天的订单数据，包含订单的起终点经纬度以及订单类型、出行品类、乘车人数的订单属性数据。

开放城市：海口

数据内容：上述时间范围内的海口市每天订单数据，包含订单的起终点经纬度以及订单类型、出行品类、乘车人数的订单属性数据。其中所有涉及个人信息的数据都经过了匿名化处理。

数据字段详情

字段 ID	字段名称	字段样本描述
order_id	订单 ID	string 类型且已脱敏
product_id	产品线 ID	1 滴滴专车， 2 滴滴企业专车， 3 滴滴快车， 4 滴滴企业快车
city_id	城市 ID	选取海口当地
district	城市区号	海口区号
county	二级区县	记录区县 id
type	订单时效	0 实时，1 预约
combo_type	订单类型	1 包车，4 拼车
traffic_type	交通类型	1 企业时租，2 企业接机套餐，3 企业送机套餐，4 拼车，5 接机，6 送机，302 跨城拼车
passenger_count	乘车人数	拼车场景，乘客选择的乘车人数
driver_product_id	司机子产品线	司机所属产品线
start_dest_distance	乘客发单时出发地与终点的预估路面距离	乘客发单时，出发地与终点的预估路面距离
arrive_time	司机点击‘到达’的时间	司机点击‘到达目的地’的时间
departure_time	出发时间	如果是实时单，出发时间(departure_time) 与司机点击‘开始计费’的时间(begin_charge_time)含义相同；如果是预约单，是指乘客填写的出发时间
pre_total_fee	预估价格	根据用户输入的起始点和目的地预估价格

字段 ID	字段名称	字段样本描述
normal_time	时长	分钟
product_1level	一级业务线	1 专车，3 快车，9 豪华车
dest_lng	终点经度	对应乘客填写的目的地对应的经度
dest_lat	终点纬度	对应乘客填写的目的地对应的纬度
starting_lng	起点经度	对应乘客填写的起始点对应的经度
starting_lat	起点纬度	对应乘客填写的起始点对应的纬度

任务提示：

可以参考一下题目或者自行发挥

1. 出租车区域推荐：分析出行数据，为司机提供行车区域建议
2. 城市规划建议：分析出行数据，提供城市交通规划建议
3. 用户级别的目的地预测：根据用户当前所在地，时间等信息，预测用户出行目的地
4. 自行发挥

必须满足以下要求：

- 使用 spark/hadoop 进行数据分析
- 工作量足够
- 提交的报告能清楚的说明分析的问题及结果，有图表进行数据可视化，有使用工具和算法的说明。

提交方式：

在大作业完成后，要求同学们提交压缩包到助教的邮箱:wangcaimeng@nlsde.buaa.edu.cn 压缩包应包含以下内容：

1. 程序代码
2. 小组分工情况
3. 分析报告：形式自选（ppt，word，pdf，jupyter notebook…）
   郑重提示：一旦发现雷同作业，大作业按 0 分处理。