

**大数据计算基础大作业**

**作业题目 纽约出租车出行记录分析**

**学 号 1190202405**

**姓 名 林逸灏**

**日 期 2021.11.15**

1. 绪论

本次大作业研究的问题是纽约市2013年的出租车出行情况。出租车是现代都市中人们常用的出行方式，由此产生的出行数据十分可观。对于纽约这样一个国际性大都市来说，每小时、甚至每分钟的出租车出行量都十分之大。本次大作业着重分析纽约出租车的各项数据，得到每天、每个时段出租车运行的数量，乘客数量，运行里程等，并使用相应的聚类算法得出乘客乘车的中心点。分析结果对于纽约市出租车规划等具有一定的指导意义。本次大作业的技术难点主要在于对Spark大数据处理框架的熟悉程度以及大数据的可视化。

1. 方法框架

本次大作业所用的软件/框架如下：

* Spark 3.2.0
* Hadoop 3.3.1
* Scala 2.12.15
* JDK 1.8
* Python 3.8.5
* PySpark 3.2.0
* PyCharm

本次大作业使用Python编写，使用Spark框架进行大数据处理。在乘客乘车中心点聚类中，使用了Kmeans聚类的方法，该方法是最常用的基于距离的聚类方法，使用迭代求解的思想得到聚类中心。

1. 实验过程
2. 实验环境配置

实验环境为Intel i7-9750H、macOS 11.2。

使用包管理软件homebrew安装Spark与Hadoop，并配置好相应的环境变量，并配置好相应的JDK、scala。

Text

Description automatically generated

由于本次实验使用Python编写，因此也需要配置相应的Python环境，安装PySpark库。

Text

Description automatically generated

本次实验环境配置较为繁琐，有许多细节需要注意。

1. 实验数据下载与数据内容

下载的数据为2013年纽约市12个月份的出行数据。

Text

Description automatically generated

可以看到，在每个csv文件中，都包含了同样格式的出行数据。第0列为medallion，第1列为hack\_license，第2列为vendor\_id，第3列为rate\_code，第4列为store\_and\_fwd\_flag，第5列为上车时间，第6列为下车时间，第7列为该趟出租车所载的乘客数量，第8列为运行时间，第9列为运行距离，第10列为上车位置的经度，第11列为上车位置的维度，第12列为下车位置的经度，第13列为下车位置的纬度。本次大作业中分析第五列之后的内容。

1. 大数据分析代码编写
2. Spark初始化与文件读取

初始化PySpark并读取文件存储为RDD[str]：

Text

Description automatically generated

由于读取的文件的第一行为表头，需要将第一行删除，留下剩下的数据内容，并缓存到内存中：

Text

Description automatically generated

1. 出行数据数量

调用RDD的count函数，得出总的数据数目。

Text

Description automatically generated

结果为：

Text

Description automatically generated

1. 乘客数量

使用MapReduce的思想，将每一行的出行数据映射为(本次出租车人数，1)，再使用reduceByKey函数将所有出租车载客人数相同的加起来，进行统计。计算总人次，再计算乘客的平均数量。将运行的结果存入文件中。

Text

Description automatically generated

保存的文件为：

Text

Description automatically generated

1. 不同时间运行次数

使用python datetime库中的strptime函数进行解析，将csv文件中的时间字符串存储为python的datetime类型的数据。同样使用前述的MapReduce的思想，得出不同时段的出租车运行的数量。

Text

Description automatically generated

结果分别为：

一个月中每天的出租车运行数目：

Text

Description automatically generated

一天中不同时段的出租车运行次数：

Text

Description automatically generatedText

Description automatically generated

1. 出租车平均运行时间

读取数据进行求和、求平均运算



结果：



1. 乘客乘车时间的分布

按每一分钟来划分乘客的乘车时间，统计不同乘车时段的分布，同样使用MapReduce的方法：

Text

Description automatically generated

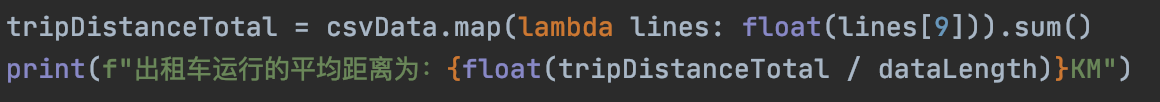
结果：

Text

Description automatically generated

1. 乘客乘车距离的分布

计算乘车平均距离，求和取平均



结果：



按每一千米来划分乘客的乘车距离，统计不同距离范围内的车次数量：

Text

Description automatically generated

结果：

Text

Description automatically generated

1. KMeans算法求乘车中心点的聚类

观察到原始的数据集中有许多位置为0，0的脏数据，一开始未过滤掉这些数据时，产生了错误的聚类结果，因此，在之后的改进中，进行了数据的过滤，忽略了这些脏数据，再进行聚类，得到比较满意的结果。

聚类中心的数量设为5。

Text

Description automatically generated

结果：

Text

Description automatically generated

1. 可视化与数据分析

使用pyecharts库进行数据的可视化，将图片渲染为html的文件，并进行相应的分析。

1. 每辆出租车的乘客数

Chart, line chart

Description automatically generated

分析：载1名乘客的出租车出行数据的数量最多

1. 每日的车次

Chart, line chart

Description automatically generated

每日的车次大致相同。

1. 每日中不同时段的乘车数

Chart, line chart

Description automatically generated

分析：从图中可以看出，从清晨6点开始，乘车数不断上升，白天（8点到17点）乘车数基本稳定，17点以后开始晚高峰，乘车数激增，夜晚8点以后，乘车数不断下降。

1. 乘客乘车时间分布

A picture containing shape

Description automatically generated

分析：可以看到乘客乘车时间大致成一个正态分布，乘车时间在5-15分钟的人数最多。

1. 乘客乘车距离分布

A picture containing graphical user interface

Description automatically generated

分析：可以看到乘客乘车距离在0-8KM之间的最多。

1. 乘客乘车中心点分析

调用gmplot库与Google Map API进行地图的绘制，标注乘客乘车中心点。

Map

Description automatically generated

分析：我们可以清晰地看到红色标记所标注的乘车中心点。

其中有三个点位于纽约曼哈顿的中心地带，这里是纽约市的城市中心，人流密集，出行量大：

Map

Description automatically generated

另外的两个聚类中心点分别位于纽约市皇后区East Elmhurst街区的LaGuardia机场和约翰·肯尼迪国际机场，都位于机场附近。这些机场是全纽约市乃至全美、全球的交通枢纽，人流量大，从机场乘出租车出发的人多。

Diagram

Description automatically generated Map

Description automatically generated

1. 相关工作

本次大作业开始前查阅了相关的文献与前人的工作。Nivan Ferreira, Jorge Poco, et al. 在2013年研究了纽约市2011年与2012年的出租车出行的时序数据，对这两年的数据做了详尽的分析，并设计了相关的大数据处理的模式与可视化的方法[1]。文中也分析了从Lower Manhattan 到JFK和LGA机场的相关出行数据，这三个位置与本次大作业中聚类所得到的中心点吻合。 Graphical user interface, application

Description automatically generated

1. 结论

本次大作业分析了纽约市的出租车出行数据，分析发现，乘客数为1位、出行距离为0-8KM、乘车时间在5-15分钟等范围内的车次最多。并且通过聚类，得到了乘车的中心点。

本次大作业中，我对大数据分析有了更加深入的理解，并能使用Spark与Python来进行一些数据分析，同时也能将分析数据进行可视化。

参考文献

[1] Ferreira N, Poco J. Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips[J]. IEEE TVCG, 2013, 19(12):2149-2158.

[2] 颜亮. 基于出租车载客数据的纽约市交通热点分析与挖掘[D]. 山东科技大学, 2020.