

数据挖掘第一次作业实验数据说明文档

学号：1452822

姓名：洪嘉勇

文件列表

- homework1.py
- CoordTransform.py

作业概述

本次作业中，我完成了以下老师布置的任务

- 将所给的1000条路径栅格化之后放入了数组grid
- 调用lsh程序包将栅格数组grid进行局部敏感hash，查询编号为15，250，480，690，900的路径，将得到的结果写入文件
- 调用knn程序包中的nn算法对栅格数组grid进行最近临近点分析，分别选取了k=3,4,5的情况，差选编号为15，250，480，690，900的路径，将得到的结果写入文件
- 调用谷歌地图api将lsh和knn结果显示在地图上进行对比，并将网页部署在自己的服务器上，完成了可视化

代码说明

homework1.py

作业处理数据的代码文件

pack_data

```
def pack_data(index):
    '''
    将原先的轨迹打包装入列表中
    :param index: 路径编号
    :return: 打包完的路径
    '''
    data = []
    doc = pd.read_csv('Traj_1000_SH_UTM')
    doc = doc.groupby(doc['Tid'])
    for line in doc.get_group(index).iterrows():
        point = utm.to_latlon(line[1][2], line[1][3],
51, 'U')
        point = wgs84togcj02(point[1], point[0])
        data.append([point[0], point[1]])
    return data
```

该函数通过轨迹编号将轨迹打包

输入: 轨迹编号

输出: 打包完的路径列表

write_json_lsh

```

def write_json_lsh(hash_size, grid):
    '''
    将生成的lsh路径放入json并存储
    :param hash_size: hash size列表
    :param grid: 处理完的栅格数组
    :return: none
    '''
    data_lsh = {}
    for size in hash_size:
        print size
        print 'list'
        data_lsh[size] = []
        lsh = LSHash(size, 44107)
        count = 0
        for line in grid:
            lsh.index(line, extra_data=count)
            count += 1
        for id in road_id:
            roads = []
            res = lsh.query(grid[id])
            print len(res)
            for r in res:
                roads.append(pack_data(r[0][1]))
            data_lsh[size].append({id: roads})

    with open('result_lsh.json', 'w') as f:
        f.write(str(data_lsh))

```

将lsh的结果写入json文件, hash_size选取10,11,12,13,14,15

输入: hash size, 栅格化数组grid

输出: json结果文件

write_json_nn

```

def write_json_knn(knn, grid):
    '''
    将生成的knn路径放入json并存储
    :param knn: k的数值列表
    :param grid: 处理完的栅格数组
    :return: none
    '''
    for nn in knn:
        data_knn = []
        neigh = NearestNeighbors(n_neighbors=nn)
        neigh.fit(grid)
        print 'nn:' + str(nn)
        for id in road_id:
            print 'id:' + str(id)
            roads = []
            distances, indices =
neigh.kneighbors(grid[id])
            for r in indices[0]:
                roads.append(pack_data(r))
            data_knn.append({'id': roads})

        with open('result_knn' + str(nn) + '.json',
'w') as f:
            f.write(str(data_knn))

```

将knn的结果写入json文件,k选取3,4,5

输入: knn中k的参数, 栅格化数组grid

输出: json结果文件

栅格化处理

```

df = pd.read_csv('Traj_1000_SH_UTM')
df['X'] = ((df['X'] - 346000) / 20).astype(int)
df['Y'] = ((df['Y'] - 362800) / 20).astype(int)

storage1 = np.zeros((3448600 / 20, 3463800 / 20))

k = df.groupby((df['X'], df['Y']))
grid = np.zeros((44107, 1000))

index = 0
for line in k:
    storage1[line[0][0], line[0][1]] = index
    index += 1

for line in df.iterrows():
    grid[storage1[line[1]['X'], line[1]['Y']],
line[1]['Tid'] - 1] = 1

grid = grid.T

```

utm数据减去最小经纬度除以20进行栅格化最后制作成44107*1000的数组同时制作一个索引数组storage1用来存储索引以备后续还原轨迹。


CoordTransform.py

用于转换火星坐标的代码文件

可视化展示

[网页](#)

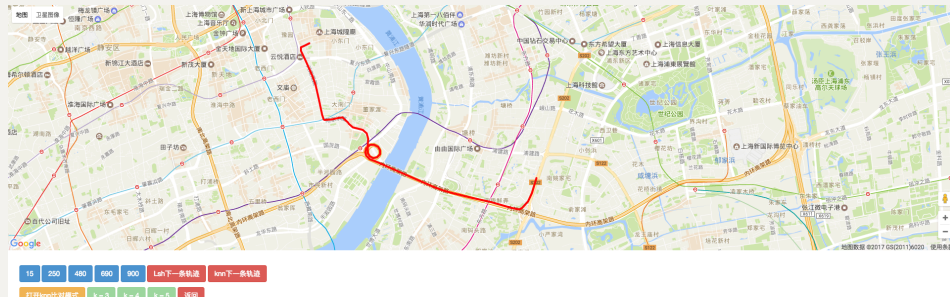
数据展现页面

数据挖掘																									
<div>  1403822 张嘉勇 </div> <div>第一次作业</div> <div>第二次作业</div> <div>第三次作业</div> <div>第四次作业</div> <div>第五次作业</div> <div>第六次作业</div>	<div>> 第一次作业成果展示</div> <div>> Hash Size: 10</div> <div>查看结果</div> <table> <thead> <tr> <th>查询序号</th><th>结果 (条)</th></tr> </thead> <tbody> <tr><td>15</td><td>1</td></tr> <tr><td>250</td><td>1</td></tr> <tr><td>480</td><td>1</td></tr> <tr><td>690</td><td>2</td></tr> <tr><td>900</td><td>3</td></tr> </tbody> </table> <div>> Hash Size: 11</div> <div>查看结果</div> <table> <thead> <tr> <th>查询序号</th><th>结果 (条)</th></tr> </thead> <tbody> <tr><td>15</td><td>1</td></tr> <tr><td>250</td><td>2</td></tr> <tr><td>480</td><td>3</td></tr> <tr><td>690</td><td>2</td></tr> <tr><td>900</td><td>1</td></tr> </tbody> </table> <div>> Hash Size: 12</div>	查询序号	结果 (条)	15	1	250	1	480	1	690	2	900	3	查询序号	结果 (条)	15	1	250	2	480	3	690	2	900	1
查询序号	结果 (条)																								
15	1																								
250	1																								
480	1																								
690	2																								
900	3																								
查询序号	结果 (条)																								
15	1																								
250	2																								
480	3																								
690	2																								
900	1																								

打表显示查询得到的数据

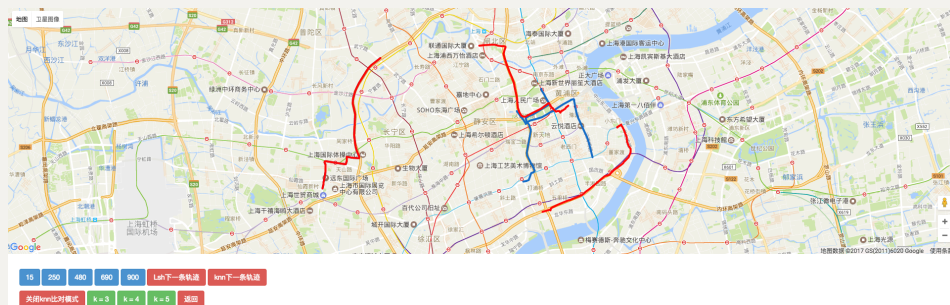
地图页

lsh,不做knn比对



选择合适的参数并打印在谷歌地图上

lsh,做knn比对



knn结果和lsh结果一起打印在地图上，红色为lsh，蓝色为knn

注意事项

在提交的作业中的实验结果文件仅仅是由代码跑出来的众多结果中的一组，因为局部敏感哈希得到的结果每次都不一样，所以我只选取了一组。

在[网页](#)中显示的结果也仅为实验中的一组，见谅。

谢谢~