Python\_codes

1.form\_basic\_datas

1)air.py 生成CK\_AIR XD\_AIR

2)hotel.py 生成CK\_HOTEL XD\_HOTEL

3)wangba.py 生成 CK\_NET XD\_NET

4)railway.py 生成 CK\_RAIL XD\_RAIL

5)dianwei.py 生成 CK\_DW XD\_DW, 但这两个表只有经纬度信息，没有时间

6)test\_tosql.py 测试文件，不用管

7) dianwei for cluster.py 生成XD\_DW\_C,提取吸毒人员经纬度信息

2. form\_statis\_datas

1) rail\_statis.py 生成CK\_RAIL\_STATIS XD\_RAIL\_STATIS

2) net\_statis.py 生成CK\_NET\_STATIS XD\_NET\_STATIS

3) air\_statis.py 生成CK\_AIR\_STATIS XD\_AIR\_STATIS

4) hotel\_statis.py 生成CK\_HOTEL\_STATIS XD\_HOTEL\_STATIS

5) hotel\_ave\_days.py 计算每个人入住酒店的平均天数，并加入至CK\_HOTEL\_STATIS XD\_HOTEL\_STATIS中

6) cluster one dw.py 将一天分为6个时间切片，然后对人6个时间切片进行聚类

7) form\_ckrail\_matrix.py 失败的尝试，无视

8) anjian\_keyword\_ lda.py 案件内容提取关键词, 案件内容运行lda主题模型

9) cluster cluster center.py 对于所有人员的聚类中心集合再次进行聚类，得到群体所有人员常用出入点。

10)center HMM.py根据人员的聚类中心以及LAI来推测该人员下一个可能落脚点。其中聚类中心按时间排序，作为观测序列，LAI作为状态序列。

11) form hotels.py根据TONGJI.CK\_HOTEL\_STATIS和TONGJI.XD\_HOTEL\_STATIS，来形成吸毒人员常住酒店列表,生成hotel\_list.npy和hotel\_list.txt。这两个文件其实是一个东西，只是.npy读取比较方便而txt方便展示

12) form xd\_airport.py 生成最重要的机场--------云南列表airport\_yunnan.npy，重要机场--------湖南、贵州列表airport\_important.npy，以及次重要机场---------江西、四川、柳州列表airport\_less.npy

13) form rail stations.py 同机场生成rail\_yunnan.npy，rail\_less.npy，rail\_important.npy

14)test JDC.py 测试各个特征在两类人间的占比，具体参见“其他可能特征.xlsx”和“特征向量.xlsx”

15) form labels.py构造模型的特征向量，重点！具体13维特征向量见“特征向量.xlsx”

16) concat.py 融合新的特征

datas:各种特征向量

3. dahei\_features

1) form neo4j data.py 为neo4j数据库准备数据

2) form cluster center.py 暂时没用，可无视

3) form import nodes.py 暂时没用，可无视

4) duanxin.py 提取短信中的关键词，暂时没用

5) form dataset.py 根据案件名称中的关键词，建立聚众赌博库SH\_JZDB\_LIST,强迫交易库SH\_QPJY\_LIST,敲诈勒索库SH\_QZLS\_LIST,寻衅滋事库SH\_XXZS\_LIST,组织卖淫库SH\_ZZMY\_LIST

6) select shehei qzls.py 筛选出敲诈勒索库中符合“三人三起”, 案件简要案情包括 "%借款%" "%非法拘禁%" "%费%" "%装修材料%" "%沙%石%" "%砂%" "%水泥%" "%工地%""%承包%" "%啤酒%" "%施工%"的记录，并建立SELECT\_QZLS

7)select xxzs.py 同上筛选出寻衅滋事库中符合“三人三起”, 案件简要案情包括"%装修材料%" "%沙%石%" "%砂%" "%水泥%" "%工地%" "%损坏%" "%被砸%" "%指使%" "%邀约%" "%洋镐%" "%洋搞%"的记录，并建立SELECT\_XXZS