1. 基本数据

本地文件

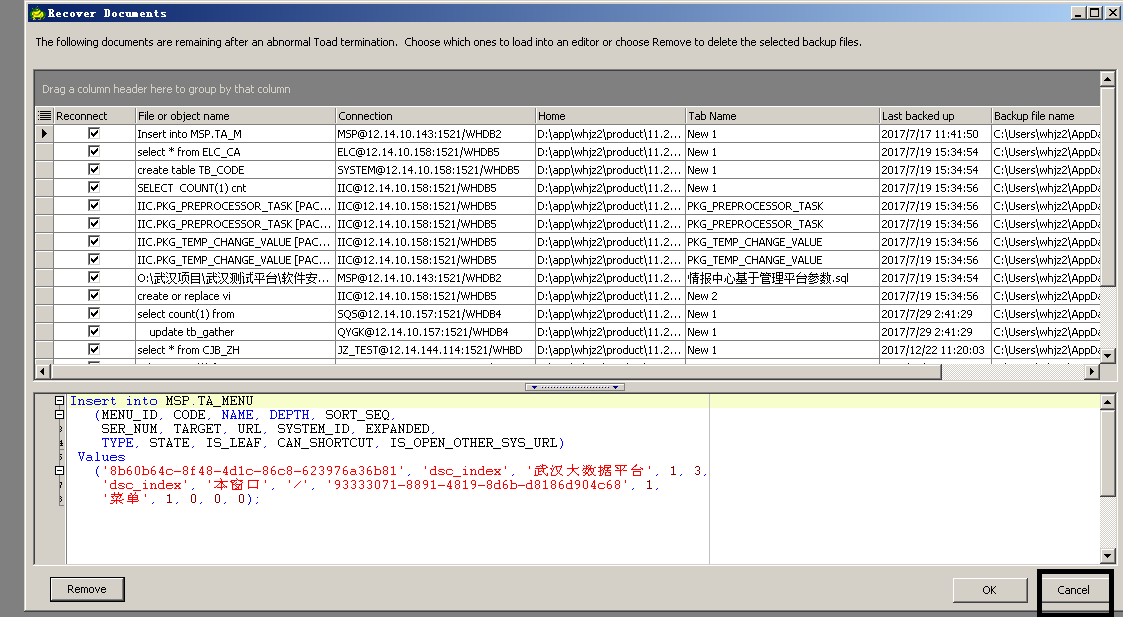
D:/python\_codes/xdry\_1818.pkl ------------1818个涉毒人员的基本信息，其中有用的是RY\_GMSFHM（身份证号）

D:/python\_codes/ck\_1953.pkl -------- 1953个常住人口的基本信息，其中有用的是CK\_GMSFHM（身份证号）

数据库

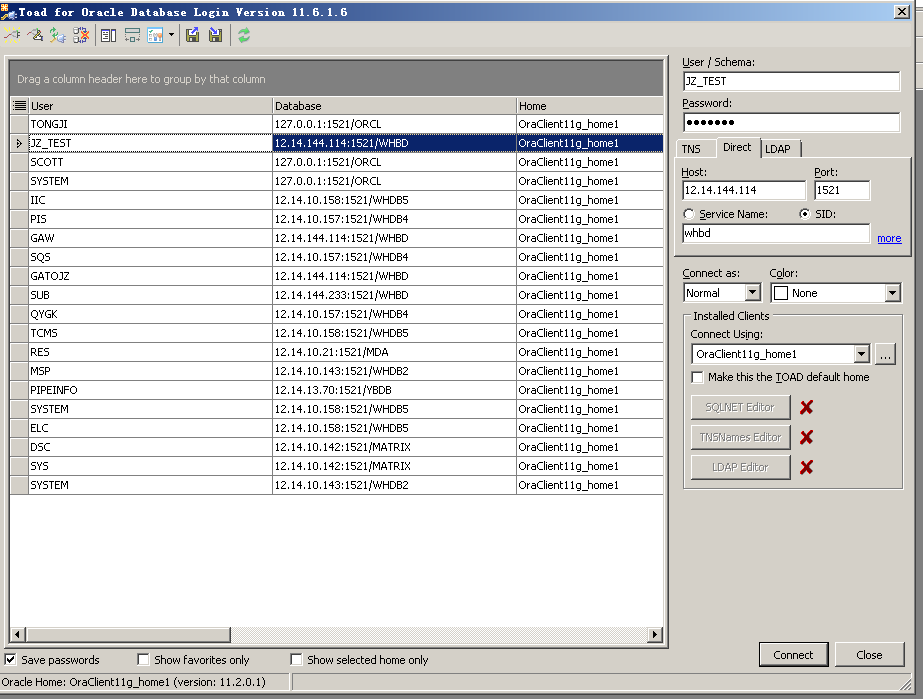
汇智给我们的数据库

1)打开toad for oracle，第一个界面点cancel

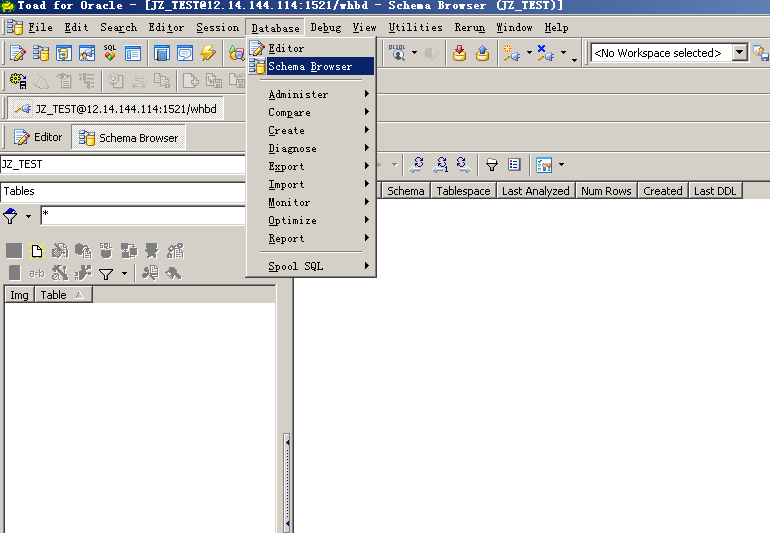


2)选择JZ\_TEST那个，一般情况下直接双击即可进入，但如果是新电脑则要手动输入ip地址等参数，其具体参数为User/schema:JZ\_TEST Password: JZ\_TEST

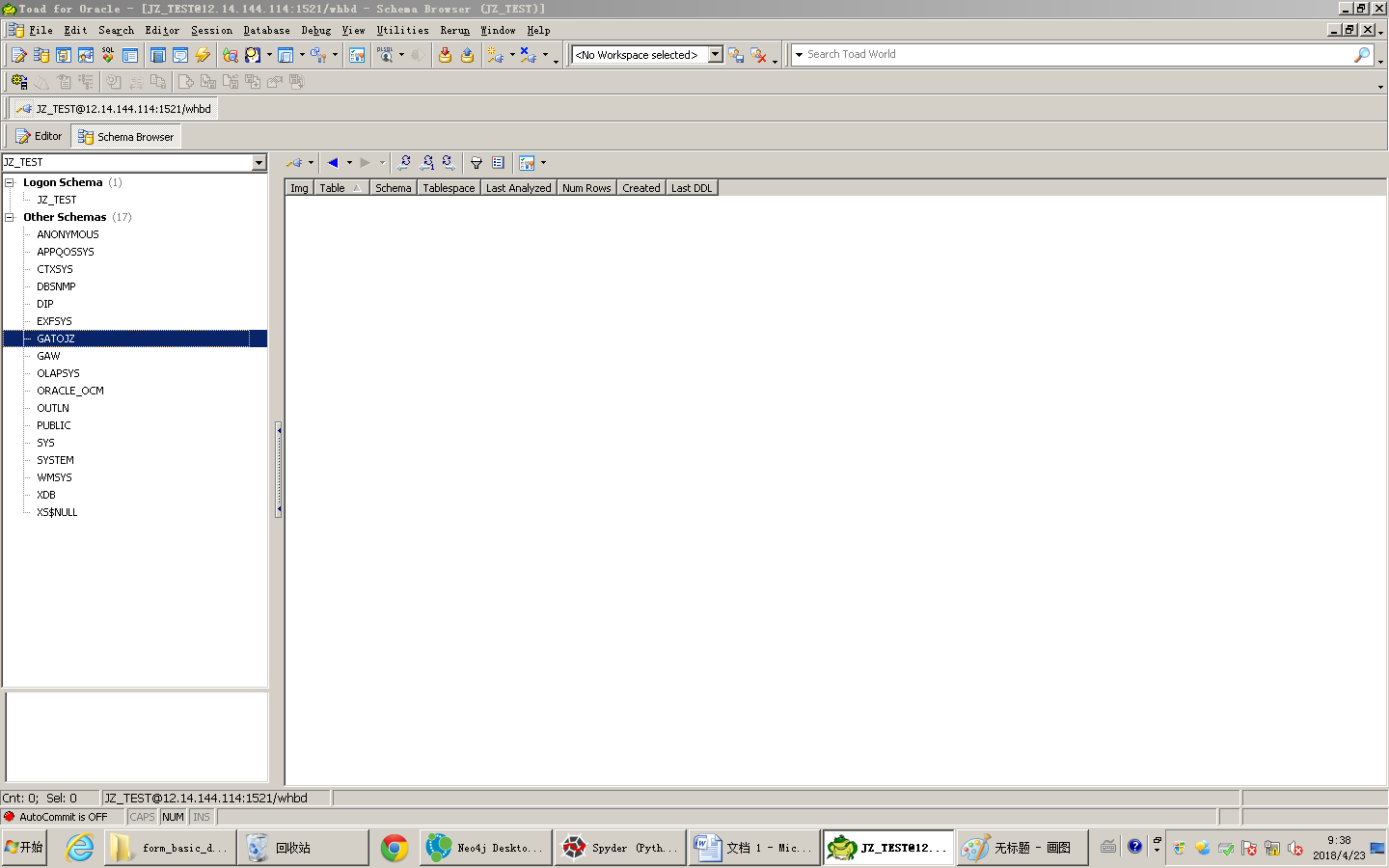
Host: 12.14.144.114 Port:1521 SID: whbd



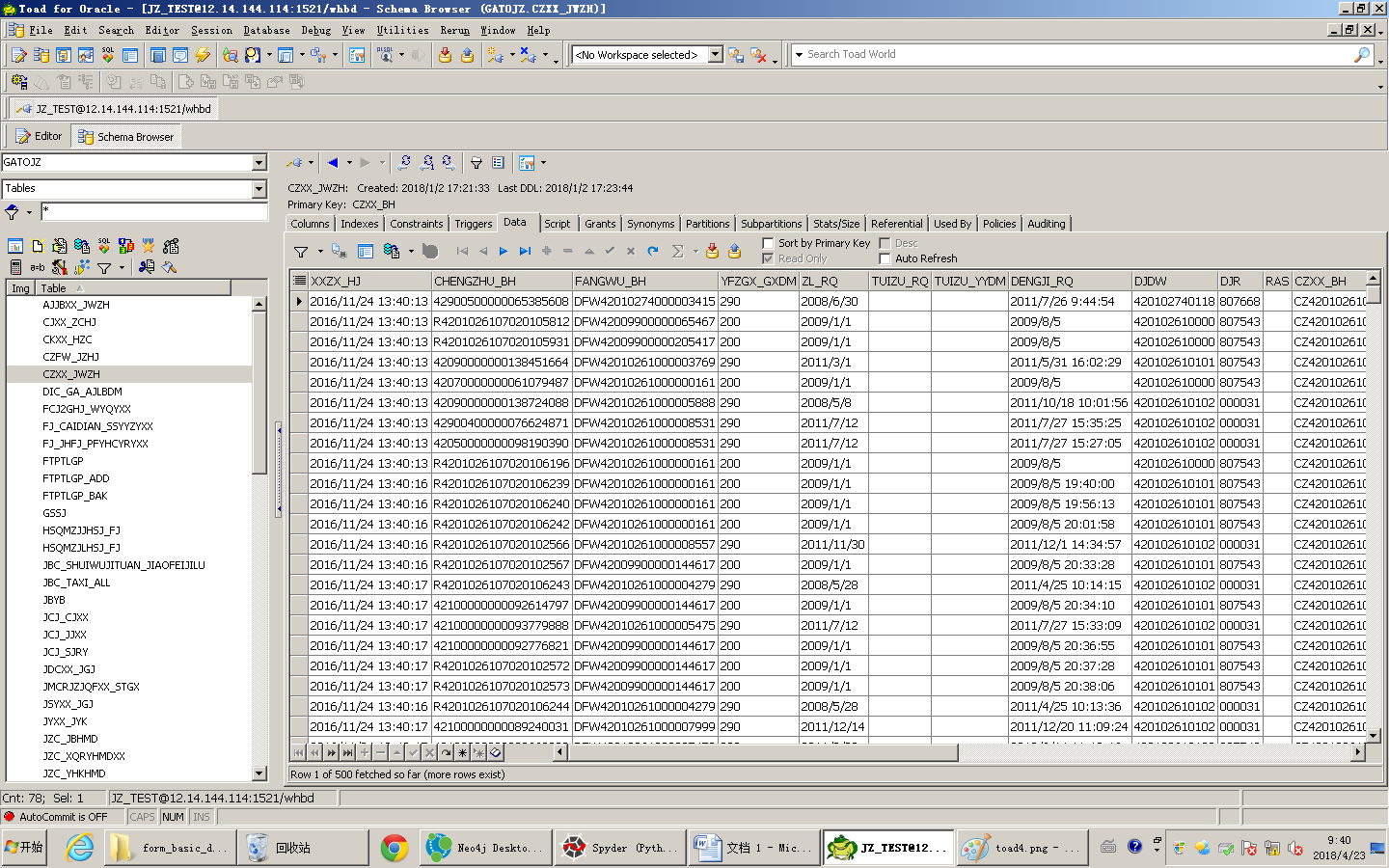
3）进入Schema Browser查看数据库中表的具体信息

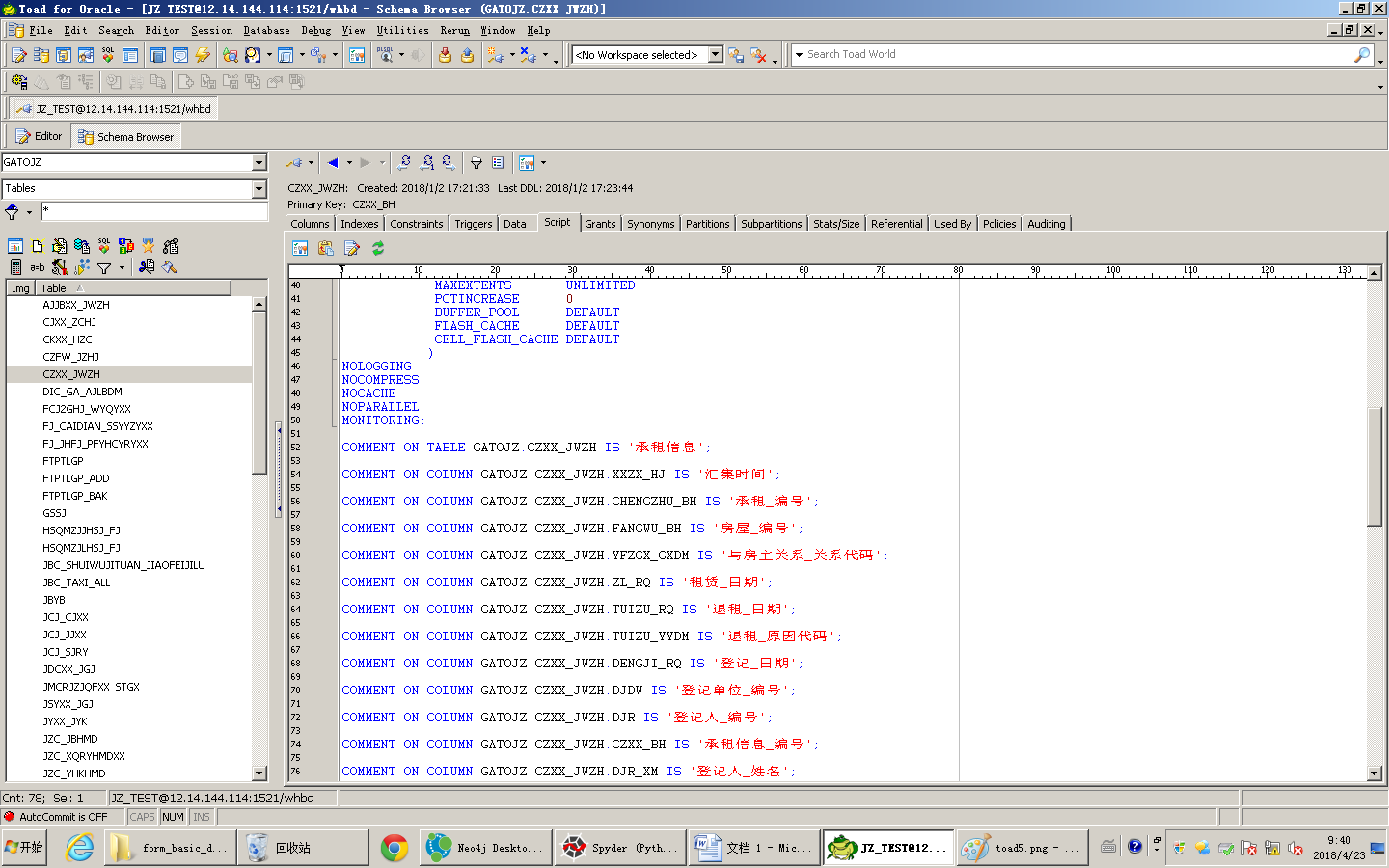


第一个下拉框选择GATOJZ



之后选择data标签看数据，选择script标签看表的注释





我们自己建的本地数据库

基本操作同上，具体参数为User/Schema:TONGJI Password:501430

Host: 127.0.0.1 Port:1521 service name:orcl

1. 从汇智数据库构建基本公安网信息-------------------代码文件夹D:/python\_codes/form\_basic\_datas

这一步我已经做完，并形成了本地TONGJI数据库，以后可能会要构成新的数据库，那时也可以找汇智寻求帮助。如果实在不行再自己建。这步的主要思路是读汇智数据库并写入同济数据库，主要操作是使用cx\_Oracle读取数据库以及from sqlalchemy import types,create\_engine create\_engine来写入到新的数据库。

读数据库主要步骤为：import cx\_Oracle

connection = cx\_Oracle.connect('TONGJI/501430@127.0.0.1:1521/orcl',encoding='utf-8')

query = “具体的sql查询语句” 如

query="SELECT \* FROM TONGJI.CK\_RAIL"

df = pd.read\_sql(query,con=connection)

connect主要格式为connect(‘用户名/密码@ip地址/service name或者sid’,encoding=’utf-8’)一般来说都要加这个utf-8，以防出现乱码。这一步的本质是把数据库中unicode编码的文件编码为utf-8编码。

理论上从汇智数据库抽出1953常住人口和1818涉毒人员应该对于列表中的每个身份证号进行遍历，对于每个身份证提取相关数据再写入数据库。但是这样效率很慢，这里我们用了sql查询中IN 命令，将整个列表的1953个身份证都查询出来。问题在于oracle IN命令只支持1000个数据，因此要构建多个列表，然后用 字段 IN list1 or 字段 IN list2 or …. 字段 IN listn 来查询

查询出结果后，需要写入数据库。写入dataframe之前要统一代码格式，即把非数值类型的列统一成str。再通过ck\_df.to\_sql("ck\_air",tj\_con,if\_exists='append',index=False,dtype=railn)命令将查询所得内容写入新数据库。

1. 从基本信息表中构建统计信息 文件夹D:/python\_codes/form\_statis\_datas

rail\_statis.py net\_statis.py air\_statis.py hotel\_statis.py 分别是对于四种数据统计其常用top10。主要操作是from collections import Counter 然后对于列表air\_count = Counter(combine).most\_common(10) 提取前十的most\_common 内容。然后将统计的内容写入数据库。这里的写入方式不同于整个dataframe的写入方式，这里主要用sql语句Insert写入而不是df.to\_sql()函数。具体写入方式为：

connection = cx\_Oracle.connect('TONGJI/501430@127.0.0.1:1521/orcl',encoding='utf-8')

c=connection.cursor()

query = "INSERT INTO CK\_AIR\_STATIS(GMSFHM,XM,TOP\_ROUTE\_LIST,TOP\_ROUTE\_TIMES, TOP\_AIR\_COMPANY,TOP\_COMPANY\_TIMES) VALUES ('%s','%s','%s','%s','%s','%s')"%(sfz,xm, air\_list,air\_times,company\_list,company\_times)

result= c.execute(query)

connection.commit()

cluster one dw.py 将一天分为6个时间段分别是:深夜0-4,清晨5-7,上午8-11,中午12-13,下午13-18，傍晚19-24。然后分别将单个人员根据六个时间段的数据分别聚类，得到其6个时间段的聚类中心。同时对于群体中所有人员，将每个人员的聚类中心进行叠加，形成6个时间段群体内所有人员的聚类中心集合。

cluster cluster center.py对于上述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。

cluster 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。

form hotel.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

form xd\_airport.py form rail stations.py 类似，也是构造吸毒人员常去的地点

form labels.py 构造模型的特征向量，重点！目前构造的特征向量为是否火车去云南、是否火车去湖南贵州、是否火车去四川柳州江西、是否有火车搭乘记录、是否飞机去云南、是否飞机去湖南贵州、是否飞机去四川柳州江西、是否有飞机搭乘记录、是否去涉毒人员常住旅店、是否出没于涉毒人员聚类中心附近、是否夜晚通话记录多于白天记录。以以上特征构造人员的特征向量。