**分类技术---二分网络上的链路预测**

姓名:吕思勤 学号:19030130448

**1、实验内容**

1. 采用二分网络模型，对ml-1m文件夹中的“用户---电影”打分数据进行建模，考虑将用户信息、电影详细信息、以及打分分值作为该网络上的边、点的权重；

2. 根据网络结构特征给出节点相似性度量指标；

3. 基于相似性在二分网络上进行链路预测；

4. 采用交叉验证的方法验证预测结果；

5. 画出ROC曲线来度量预测方法的准确性。

**2、分析及设计**

1. 采用二分网络模型，对ml-1m文件夹中的“用户---电影”打分数据进行建模；用户对自己看过的电影打分1-5分，其中1分表示最不喜欢，5分表示最喜欢。假设分数大于3分的，表示用户喜欢这部电影，在二部图中构建一条从用户到该电影的连边。考虑由m个用户n部电影构成的电影推荐系统。用户i对电影j打分超过3分，就在i和j之间连接一条边aji=1,否则aji=0。

2. 计算资源配额矩阵；

计算资源配额矩阵W中的元素wij表示产品j愿意分配给产品i的资源配额。假设一个用户选择过的商品j都有向该用户推荐其他产品i的能力



kj表示产品j的度（被多少用户评价过），kl表示用户l的度（用户选择过多少产品）。

3. 对给定用户，按照其喜欢程度，对电影进行排名，进行电影推荐；

目标用户的资源分配矢量f。初始时，将他选择过的电影对应项资源设置为1，其他为0，得到初始n维0/1向量。则最终的资源分配矢量:



将用户所有没看过的电影按照中对应项的得分进行排序，推荐排序靠前的电影给该用户。

4. 算法预测准确性预测；

将二部图中的边随机分为两部分，期中90%归为训练集，10%归为测试集。

对给定用户i，假设其有Li个产品是未选择的，如果在测试集中用户i选择的电影j，而电影j依据向量被排在第Rij位，则计算其相对位置：



越精确的算法，给出的rij越小。对所有用户的rij求平均值来量化评价算法的精确度。

5. 画出ROC曲线来度量预测方法的准确性。

选取不同的算法阈值，计算相应的真阳性率以及假阳性率，画出ROC曲线。

**3、详细实现**

1. 数据导入

将data中的ratings.dat数据利用pandas的read\_csv函数导入

一張含有 文字 的圖片

自動產生的描述

2.数据预处理

分配训练集和测试集

一張含有 文字 的圖片

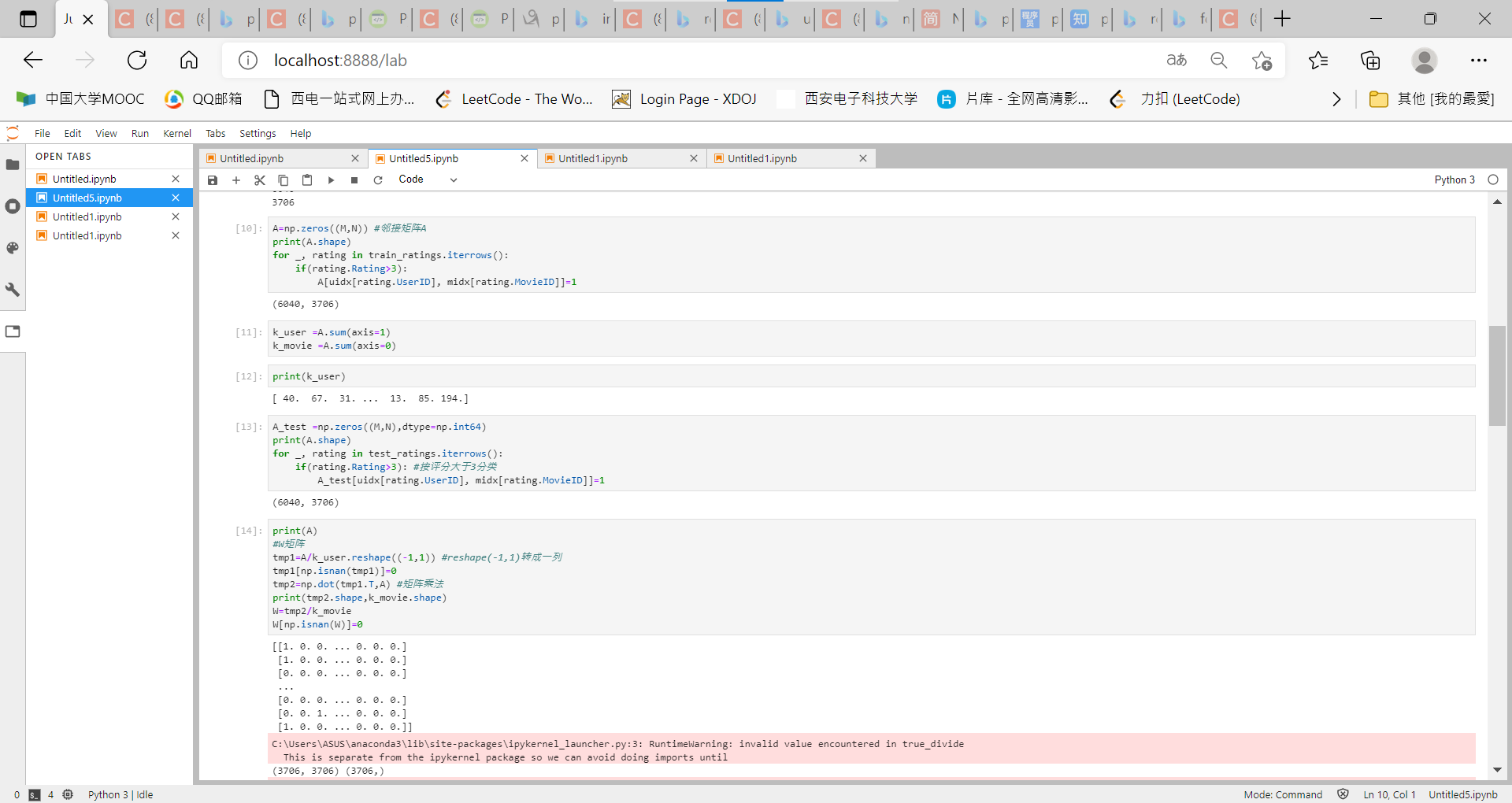
自動產生的描述

数据离散化

一張含有 文字 的圖片

自動產生的描述

建立邻接矩阵(以评分大于3分类)



3.训练模型

计算资源配置矩阵W (以向量化的方式加速计算)

一張含有 文字 的圖片

自動產生的描述

计算资源分配矢量F

一張含有 文字 的圖片

自動產生的描述

4.评价算法

计算平均推荐相对位置，评价算法精度

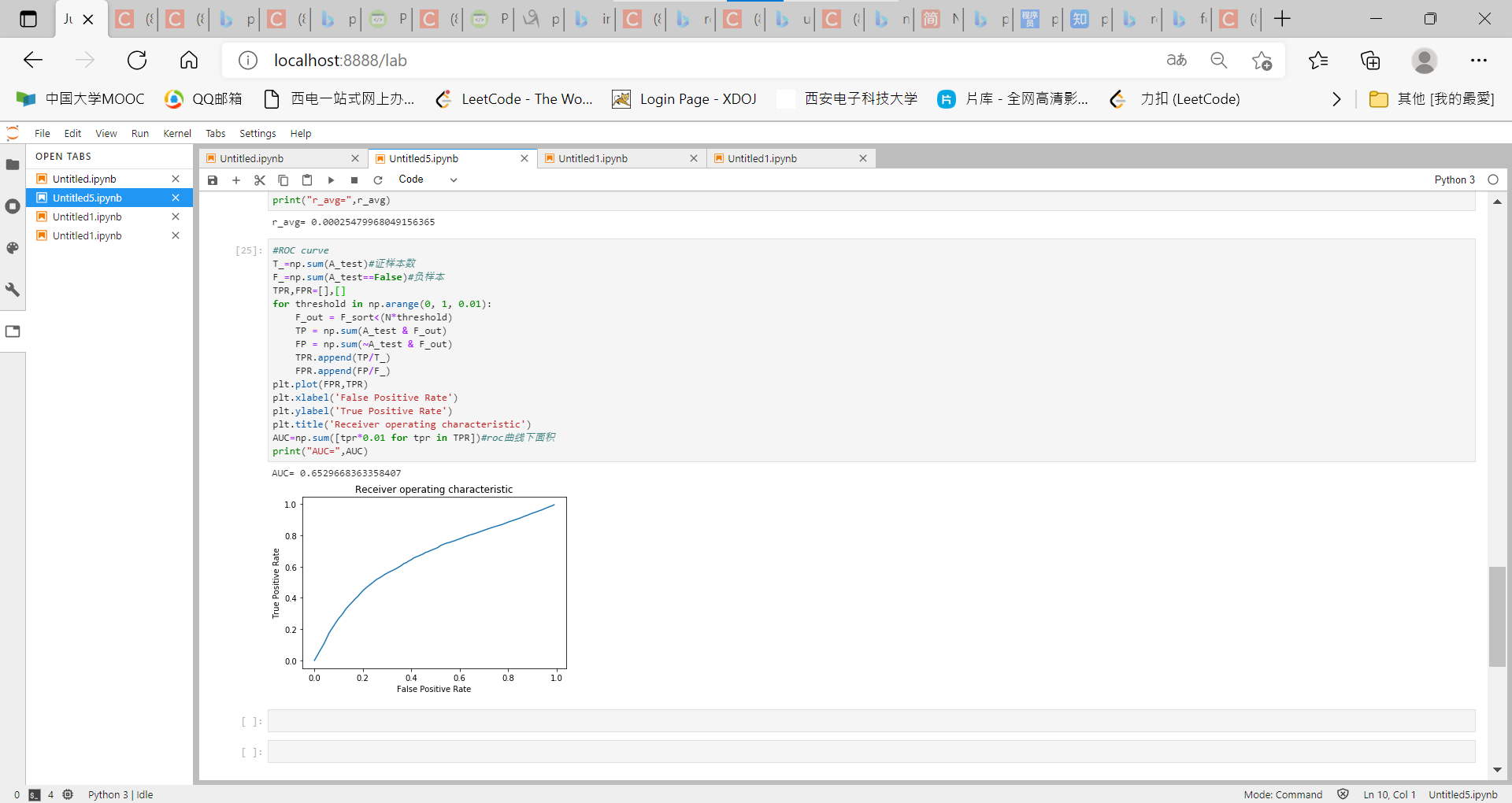
**一張含有 文字 的圖片

自動產生的描述**

**5.** 画出ROC 曲线

计算TPR(真阳性率)以及FPR(假阳性率)，画出ROC 曲线

计算AUC，评价算法准确性

****

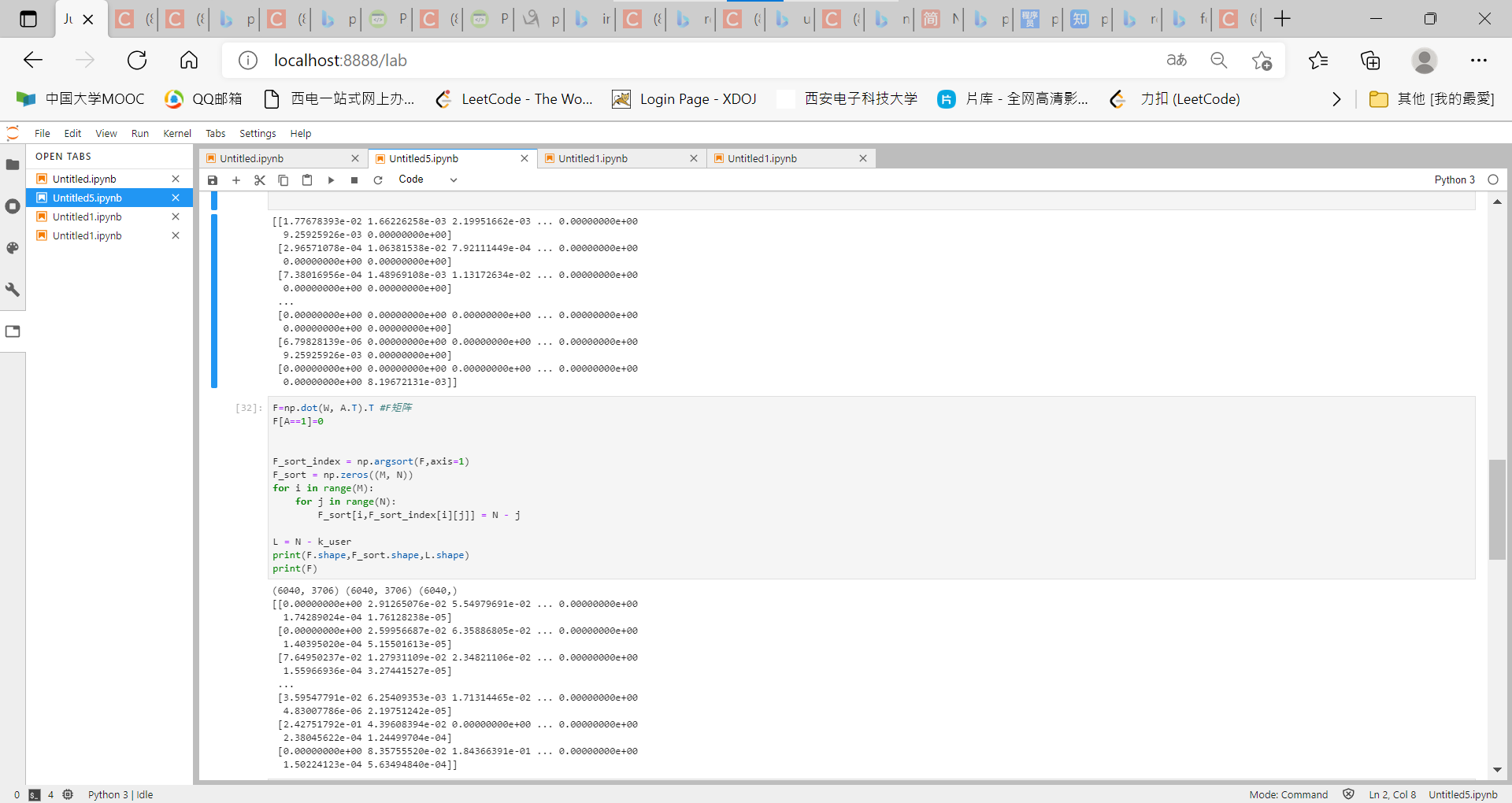
**4、实验结果**

1. W矩阵:

**一張含有 文字 的圖片

自動產生的描述**

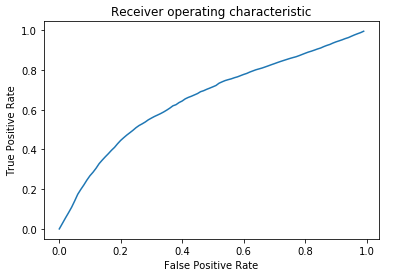
1. F矩阵:

****

1. r值:

r\_avg= 0.00025479968049156365

1. ROC曲线:



**5、心得体会**

通过本次实验，我提高了自己的动手能力，查阅资料与实践能力，学习到了一种简单有效的推算法以及python中numpy库与pandas库中的各种该函数的应用方式，这些函数的功能确实很强大，能够将复杂的数据转化成矩阵的形式做处理。在实验过程中一开始没有考虑到本实验较大的数据量，用循环方式运算，编译后运行了很长一段时间都没出结果，后来改用矩阵乘法的方式，成功得到结果。

由于刚接触数据挖掘这方面的知识且对python一些库函数使用方式还是比较生疏，这次的实验花了比较多的时间才完成，成就感十足，也更加了解数据挖掘知识的应用领域与方法，以后我会尝试学习更好的算法不断地提升自己。