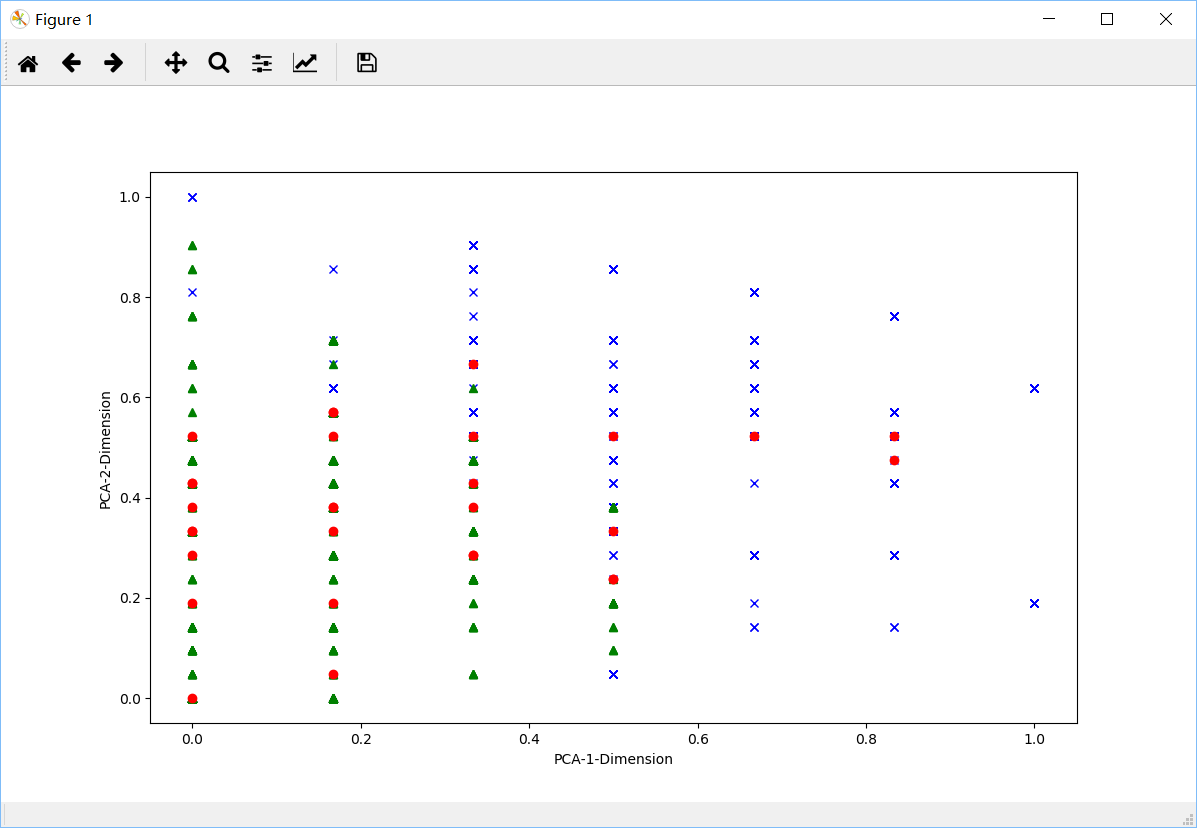
2018年9月28日星期五

下午5时24分

已经下班，自动加班。

随着数据处理开始，自己也着手继续下一步分析。昨天基于CERT5.2中用户离职/解雇的关系数据进行了初步的KMeans聚类分析，但是结果表明无论是PCA还是MinMax都无法挽救该阶段的KMeans，结果如图：



上图为PCA=2时两个维度的图

也许自己不死心吧，决定再次尝试一次，因为论文说“relationship to laid off employees”影响着Job Satisfaction，那么如果直接依据离职人数进行排队呢，不知道场景二的30个用户能排在哪个位置？

首先分析X\_0，即仅考虑同一团队的离职人数，则有：

开始输出Insiders\_2在排序后的索引...

如果考虑X\_1即同一部门内离职人数，依旧不明显

如果考虑LaidOff，依旧结果不明显，此方法作罢。

首先输出X\_0...

30

BYO1846 775

CHP1711 1042

CIF1430 1404

CKP0630 1648

DCC1119 349

GWG0497 1458

HIS1394 463

HMS1658 769

HSN0675 1452

HXP0976 1986

ICB1354 1034

ITA0159 570

JAL0811 1138

KSS1005 1825

LVF1626 105

MCP0611 1555

MDS0680 1654

MGB1235 1964

NAH1366 1557

d'dOSS1463 1063

RRS0056 1702

SIS0042 1383

SNK1280 1459

TMT0851 480

TNB1616 447

TRC1838 1155

VCF1602 557

WDT1634 1577

ZIE0741 1830

......<<<<<<分割线>>>>>>......

2018年9月29日星期六

上午10时42分

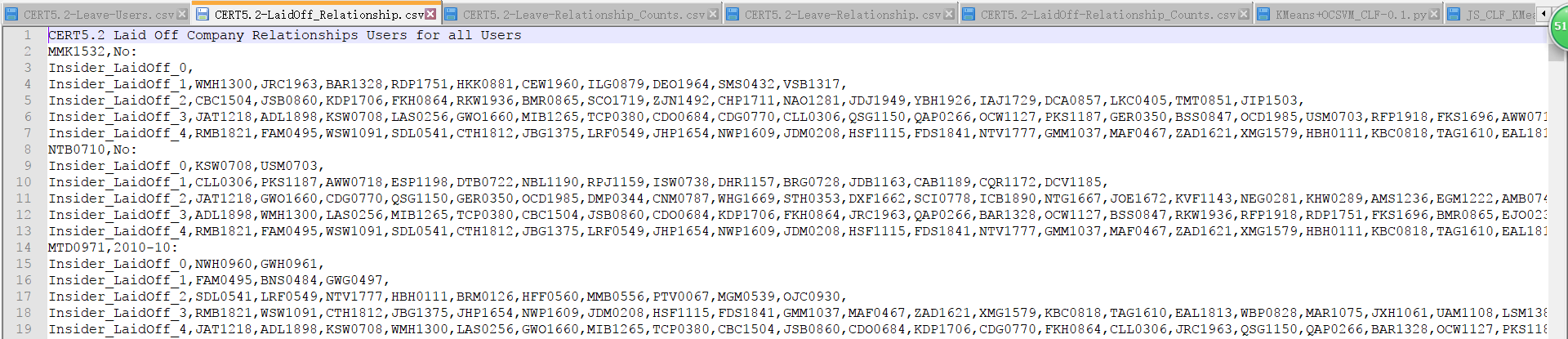
国庆假期调休。

昨天下班时进一步考证了单独从离职人员数量角度无法有效区分，那么看来需要考虑离职人员的亲密度了。

再继续进行下一步前，简单回顾下自己分析离职人员关系的步骤。

1. 统计了LDAP中所有的部门架构层级关系，进而统计了每个部门架构（细化到团队）中的用户成员，得到了CERT5.2-LDAPUsers.csv
2. 在上述基础上，从2009-12的LDAP文件开始，逐个分析到2011--05的LDAP，统计列出这段时间所有离职（工作时间未干全）的员工列表，以及其离职时间（未离职就是NO），得到了CERT5.2-Leave-Users.csv
3. 在上述基础上，重点统计2000个CERT5.2用户的各个层级架构中，出现的离职时间不晚于其离职的用户成员，并分别列入各个层次的列表中，得到CERT5.2-Leave-Relationship.csv；当过滤掉场景二的攻击者后，得到的是CERT5.2-LaidOff-Relationship.csv

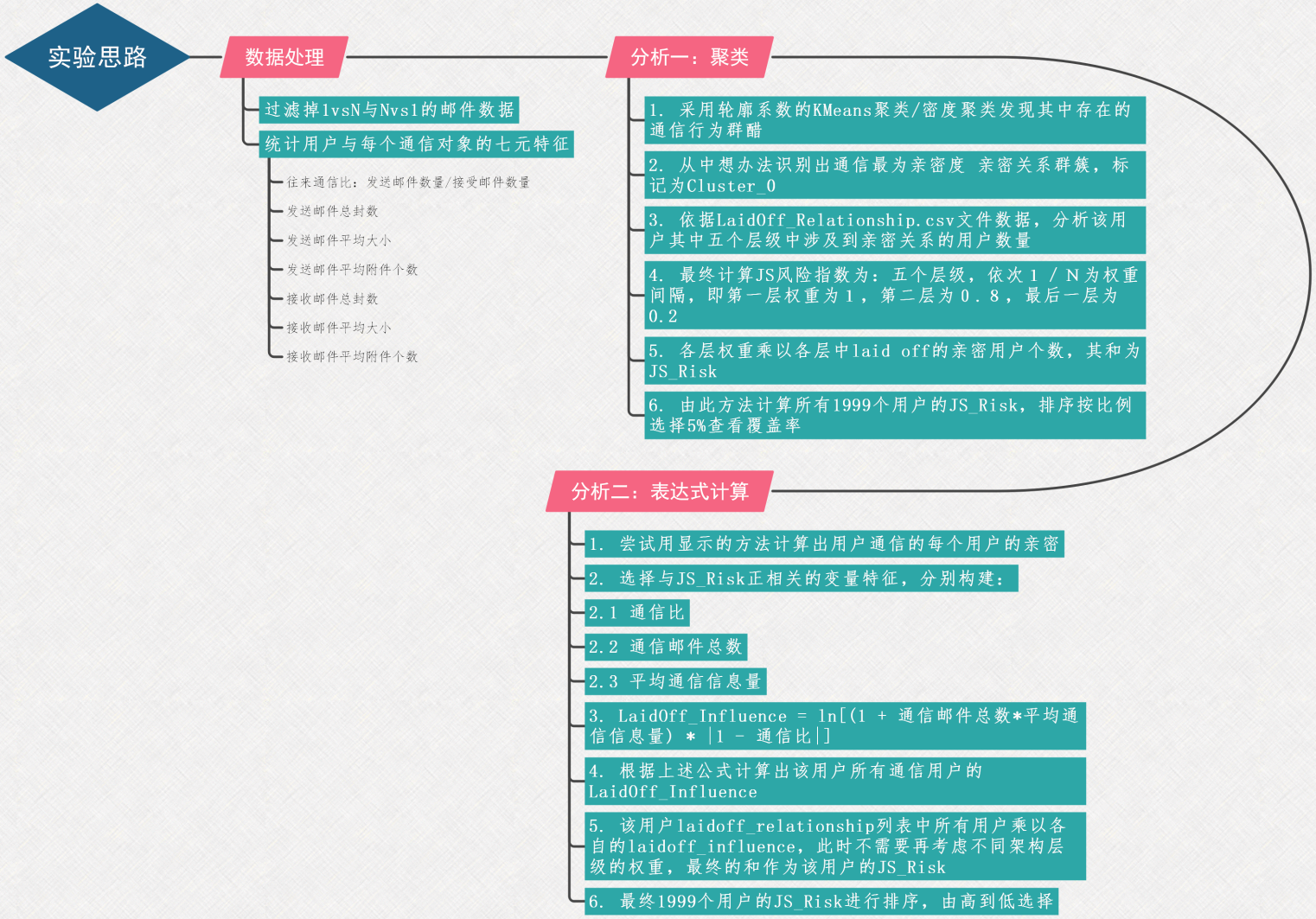
一个简单的示例如：



2018年9月29日星期六

下午2时51分

下午初步整理了下接下来分析的思路，在正式实验前，首先将整体思路放在首位。



完成了单一用户的发送与接收邮件列表，可以看出发送与接收的邮件用户十分不对等

首先统计用户的有效邮件通信列表...

BYO1846 119 发送的邮件用户列表分析完毕...

Macy\_Patterson

Hannah.Mollie.Callahan

Eugenia.Shafira.Mercado

Gareth.Ross.Becker

Xanthus.Kamal.Berger

Micah-King

Tatyana\_A\_Johnston

Jonathan\_Buck

WAS1

Tana\_P\_Orr

BYO1846 42 接收邮件用户列表分析完毕...

Buffy.Yolanda.Ortiz

Kitra.Alexa.Mack

Buffy\_Ortiz

Rose.Maisie.Blackwell

Jonathan.Kasimir.Buck

Olympia.N.Bonner

Kirby.Violet.Pena

Wang.Clark.Rice

Belle.Georgia.Frost

Jemima.Brynne.Pratt

晚上重新修改了程序，结果得到

BYO1846 发送邮件与接收邮件分析完毕...

发送邮件统计示例...

**119 119 （两次结果一样）**

['Macy\_Patterson', 2.0, 5687130.0, 7.0]

['Hannah.Mollie.Callahan', 6.0, 6778336.0, 7.0]

['Eugenia.Shafira.Mercado', 8.0, 11621806.0, 10.0]

['Gareth.Ross.Becker', 5.0, 15104163.0, 12.0]

['Xanthus.Kamal.Berger', 7.0, 14001504.0, 13.0]

['Micah-King', 4.0, 6512878.0, 10.0]

['Tatyana\_A\_Johnston', 1.0, 5938693.0, 6.0]

['Jonathan\_Buck', 2.0, 1627944.0, 2.0]

['WAS1', 6.0, 10403150.0, 10.0]

['Tana\_P\_Orr', 1.0, 2930252.0, 4.0]

接收邮件统计示例...

74 74

['Nelle.Regan.Roman', 14.0, 16731087.0, 20.0] 经过实际验证，NRR确实存在receive邮件

['Nelle.Zephr.Hammond', 6.0, 5106777.0, 7.0]

['Herrod.Bevis.Pearson', 1.0, 1576042.0, 1.0]

['Driscoll.Lewis.Hurst', 1.0, 9091311.0, 8.0]

['YDM6', 4.0, 10587019.0, 12.0]

['Hannah\_M\_Callahan', 11.0, 22461830.0, 24.0]

['WCR66', 1.0, 2053416.0, 2.0]

['Roman-Nelle', 6.0, 9075406.0, 9.0]

['Robert.Merritt.Velez', 26.0, 35519771.0, 41.0]

['Rebekah\_Santos', 15.0, 15771213.0, 17.0]

2018年10月12日星期五

下午2时26分

时隔两周，继续国庆假期前的工作。

由于原有程序终未考虑没有附件信息的邮件数据，因此导致处理的邮件数据长度不一致，重新添加了对于邮件数据长度的判断，程序通过。

新处理得到的Email关系数据，以一个用户为例：

CWW1120,-1.0,1.0,41547.0,0.0,0.0,0.0,0.0,

ZIN1633,1.0,9.0,38190.6666667,0.0,9.0,474093.111111,0.555555555556,

NRM1656,1.0,7.0,28107.8571429,0.0,7.0,186062.857143,0.571428571429,

用户ID表示与目标用户关联的邮件联系人，第一个特征表示发送/接收比，如果是-1则表示只有向该用户的发送没有接收。

必须重申我们现在的研究方向：通过relationships to laid off employees来分析出哪些用户的JS发生了变化。基于“奥卡姆剃刀假设”，在满足实验检测要求的基础上，越少的假设便越好，因此，这里我们首先不考虑用户初始静态人格OCEAN对于反生产行为影响的基础性影响，而是先从场景二用户的“LaidOff Employees Relationship”开始。

接下来有两种分析思路：

1. 对每个用户的邮件通讯用户集合进行聚类，目的是发现通信行为亲密程度不同的用户群簇，然后从其中依据一个正变量公式定性标记出最亲密的关系群簇；然后与用户的离职关系集合比对，按照不同组织层级，分别乘以对应系数，得到离职用户对该用户JS的影响风险，由高到低排序；
2. 第二种思路不再使用聚类，而是直接利用定性公式计算该用户邮件通讯集合每个对象的通信数据量，通过该变量标记其亲密度，然后依据亲密度由高到低对邮件通讯集合排序；最后比对离职用户关系，计算其中离职用户的加权和作为该用户的JS风险指数；

2018年10月18日星期四

下午4时41分

再次回首，发现上次的实验记录竟然是12日，上周五的了！

这周自己考出了摩托驾照，祝贺自己！

但是毕业研究才是自己的头等大事，专业技术的培养材质自己事业发展的关键，因此，自己还是需要着重回到研究层面。

即便自己毕业，即便自己可以试着托托琴姐的关系，明年努力毕业后的择业也未必乐观。

现在开始验证上周五的两个实验假设。

先开始第一个，使用一个聚类+簇类标记的方法选择出一个用户的亲密关系群簇，然后比照该群簇计算场景二30个用户的。

2018年10月19日星期五

上午9时54分

今天上午，周五，继续昨天的工作。

自己重新考虑了一次用户邮件通讯特征中反映的可以表现亲密关系的变量，可以基于以下基本假设：

* 核心假设：朋友关系的用户间存在大量邮件记录
* 推理假设：朋友间的信息流应当尽量均衡，反映为通信的信息量以及邮件次数尽量均衡

在上述假设下，我们提出一个可以标记KMeans聚类结果群簇的定性公式：

上述公式中有几点必要说明：

1. Email\_Ratio是在初始阶段提取特征时计算的，因此依据的邮件发送/接收数量为原始数量；
2. 在进行KMeans聚类前，对每个用户的七元特征进行了MinMax归一化，因此后续在表示通信信息量时，直接将数据量与附件个数进行相加；
3. RelationLevel公式中每一个变量均假设与亲密度高低成正比例关系
4. 需要重新修改邮件特征提取部分代码，更新通信比的计算
5. 分析提取的邮件通讯七元特征中，邮件size与附件个数默认都是平均值；在计算群簇的RelationLevel时，需要先计算列均值，然后再按照上述公式计算中心代表的整体通信量定性特征；

2018年10月19日星期五

下午4时29分

下午终于完成了上午的程序，继续：

首先以分析一个用户为例，即BYO1846

分析得到其邮件通讯聚类得到6个群簇，然后计算中心比较得到朋友群簇

Cluster3

***Macy\_Patterson,WAS1823,HMC1847,ESM1828,GRB1842,YDM1822,CBW1826,HDA364,WCR1830,NRR1835,ONB1833,MUP1819,JKB1843,HDA1824,ACH1831,MPK1844,Rebekah\_Santos,WLH1827,TRC1838,NJV1818,RMV1820,GFV1837,WOS1834,ACH1840,RSS1825,NZH1839,CCB1836,BGF1845,JBP1832,***

对比下该用户的离职人员关系

BYO1846,2010-12:

Insider\_LaidOff\_0,RMB1821,FDS1841,

Insider\_LaidOff\_1,NWP1609,ZAD1621,TAG1610,MAR1075,JXH1061,LSM1382,KBC1390,IVS1411,MFM1400,CTT0639,MZO1066,

Insider\_LaidOff\_2,WSW1091,JHP1654,JDM0208,HSF1115,UAM1108,CIM1095,ZHB1104,PTM1432,WDT1634,DHS0204,DDR1649,HMK0653,HMS1658,

Insider\_LaidOff\_3,FAM0495,SDL0541,CTH1812,JBG1375,LRF0549,NTV1777,GMM1037,MAF0467,XMG1579,HBH0111,KBC0818,EAL1813,WBP0828,NWH0960,BRM0126,GWH0961,HFF0560,MMB0556,BNS0484,CKP0630,JKM1790,PTV0067,MGM0539,OJC0930,KLB0918,MTD0971,TTR1792,MCP0611,VVG0624,LLW0179,RAT0514,NAH1366,PLF1030,TPO1049,RRS0056,MJA1784,EJV0094,GCB0118,

Insider\_LaidOff\_4,JAT1218,ADL1898,KSW0708,WMH1300,LAS0256,GWO1660,MIB1265,TCP0380,CBC1504,JSB0860,CDO0684,KDP1706,CDG0770,FKH0864,CLL0306,JRC1963,QSG1150,QAP0266,BAR1328,OCW1127,PKS1187,GER0350,BSS0847,OCD1985,USM0703,RKW1936,RFP1918,RDP1751,FKS1696,BMR0865,AWW0718,EJO0236,HKK0881,ESP1198,MMR1458,JIB1258,SCO1719,ZJN1492,ZIE0741,DTB0722,CEW1960,ILG0879,DMP0344,DEO1964,CNM0787,NBL1190,ALT1465,WHG1669,SMS0432,WFV0687,STH0353,RPJ1159,JKB0287,ELM1123,DXF1662,NAO1281,SCI0778,AYG1697,LMW0837,ICB1890,NTG1667,PCK0271,DHR1157,ZVW1475,BRG0728,HPM0360,KJG1121,JOE1672,UKM0845,KVF1143,

我们进行了LaidOff与Friends群簇的交叉验证，得到了下述结果：

LaidOff Friends is

0 ['RMB1821', 'FDS1841']

1 ['KBC1390']

2 []

3 ['BRM0126', 'OJC0930']

4 ['TCP0380', 'DTB0722', 'RPJ1159', 'ZVW1475']

即只有第2号邮件联系群簇在离职人员中没有出现