* 校色背景与目标概要

用户画像需要一张基础事实数据表来为下游任务和应用做支撑，目前采用正则抽取的方法来解决平台大部分短信量的字段抽取，并将抽取结果存入一张画像基础数据表。

根据2019年一整年的数据统计结果显示（见图1），平台80%左右的短信发送量是由top 4k的签名所完成，所以采取正则抽取每个签名下的字段可以覆盖大部分短信的情况，但这需要人工查阅每个签名下的可抽取字段，所以此文档的目的是将正则抽取短信内容到基础表的数据的流程文档化。



图1

* 正则抽取流程

1. 复制粘贴签名到百度百科或直接百度，了解签名，比如签名所属什么样的行业

2.将分好类的签名文件用编辑器打开（建议notepad++）（签名文件见图2）

3.查看打开之后的签名短信，并编写正则。编写格式为{“签名名称”:{“抽取字段名称”:[““，”“]}}

详见附件property.json，字段说明详见此文档最后一页

4.填充comment字段，comment字段是给人看的，它主要是用来说明签名所属行业，如果你定义了一些“新奇“字段或已有字段都要给出相应解释也放这里，它在画像中的意义或使用也可以放这里，总之这个comment包含签名和签名下抽取出的字段所反映的信息

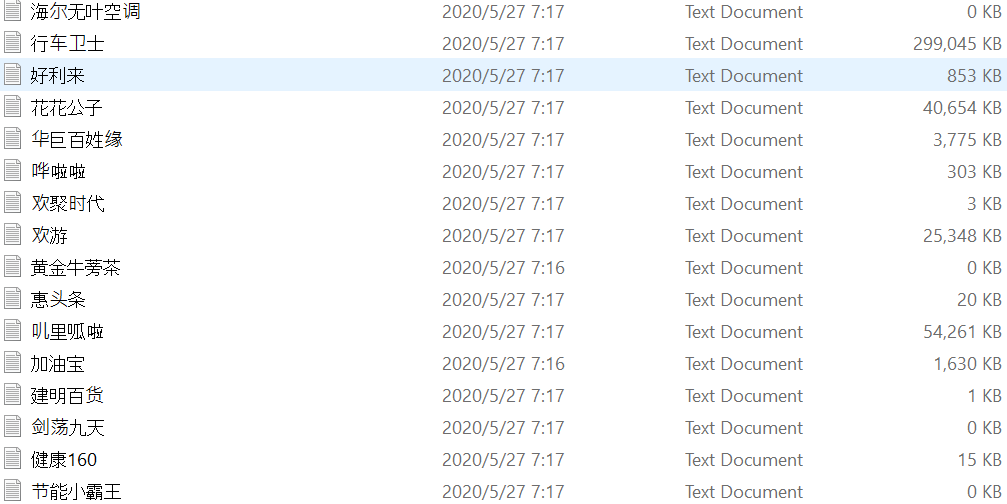


图 2

* 一些特殊情况和应对方法

特殊情况1：某个签名文件为空，由于这些数据是来自2019年11月，出现这种情况是因为那个签名没有在11月份发送任何短信，那么将这些签名一行一个的存到 “样本没有数据.txt” （存储格式见图3）之后会统一抽其他月份来获取相关签名的数据样本

特殊情况2：某个签名文件数据太少，少于5w条的叫太少，那么也把它们按照图2的格式一行一个的将签名存到 “样本过少.txt”

特殊情况3：某个签名文件有一定量数据，但观察下来抽不到什么东西。这种情况按照图2的格式一行一个的将签名放到 “其他.txt“ 下面



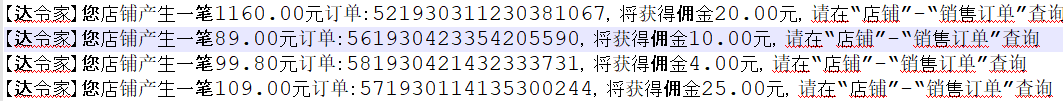
图 3

* 交付的东西

1. 你整理的正则json文件
2. 样本没有数据.txt文件
3. 其他.txt文件

* 关于抽取时的一些注意事项：

所有的抽取的字段必须是关于这个号码主的，非号码主的不能抽取，例如里面包含了非号码主的店铺地址或者快递小哥的号码等这些不能抽，金额不和号码主相关的也不能抽，短信里面出现号码的，需要通过查询二次确认是否和号码主相关。有一些甚至要根据签名查看相关企业商业模式才能知道，比如达令家，达令家的店铺并非号码主的，而是号码主引流而来的客户在达令家购买的商品店铺，所以此签名下的店铺产生的订单金额和号码主是没有关系的，佣金才是奖励号码主引流的报酬



抽取时给字段取名要特别小心，优先考虑property.json表里已有的字段，观察字段描述和comment里面的描述是否和你正在抽取的字段描述一致，如果一致，那么优先使用property.json里面的名字，如果描述不一致或者找不到你想要的字段，那么你需要给抽取的信息重新命名，取名字时要能客观反映你所抽取的内容事实，例如你想要抽取地点，如果所抽取的字段可以细分成家庭住址和常去住址，那么给他们做一个细分，但细分后的结果要能客观反映那个签名下短信的90%以上的绝大部分情况

字段命名-归类问题

Purchased系列

* 关于property.json正则注释

这是property.json里面的字段注释，property.json是关于100多个签名的正则，正则大部分都通过校验，主要用到前向和后向断言。其中有抽取日期的（xx年-xx月-xx日）、车牌、金额（包含千位和小数位，例如9,927,932.23），字段对应的如果时 “”（string类型） 那么就代表字段为人工分类的结果，如果是 []（list类型），那么代表这是一套或多套正则，interest一般是“” 为人工根据行业分类的结果，例如签名是一家游戏装备交易平台，签名下含有大量验证码或游戏虚拟装备交易订单，那么interest为“网游” 或 “网游装备交易”

pNam：姓名

pronoun：人称（称呼）

gender：性别

date: 日期，格式为 xx年-xx月-xx日

houseLoc：家庭住址

freqVisLoc：常去地址

overdue：逾期情况，只出现在借贷平台当中

loanIn：收到的还款、贷到的金额等等，描述钱在当时在短信内容中的归属状态是属于号码主的，即将借出去的钱在短信内容中也是属于号码主拥有的

loanOut：申请的借款、应还金额、待还金额等等，描述钱在当时短信内容中的状态还不属于号码主或已经借出去了的金额

bank：金额所绑定的银行卡

investment：号码主投资的基金信息

deposit：存款金额

hospital：门诊医院

healthBillCovered：被医保报销的医疗账单金额

healthBillPaid：号码主最终支付的医疗账单金额

healthBillTotal：号码主看病所产生的医疗账单总金额

houseBill：房屋装修、物业费等金额

violationRecord：车辆违章记录

carInfo：有无车辆，反映号码主使用过共享交通工具时才会被打上此类标签

carBrand：车型，反映号码主的车品牌和具体型号

licensePlateNum：车牌

purchasedAmount：购买金额，它是指在交易过程中发生的金额

purchasedItem：购买产品

purchasedDrug：购买的药品

insurance：购买的保险类别

position：工作

jobRecDetail：被推荐的工作（来自招聘平台）

grade：小孩读书年级

school：小孩所上学校

child Name：小孩名字

kidInfo：有无小孩

accommodation：出行住宿信息

traffic：出行交通信息

scenicSpot：出行旅游景点信息

membership：会员信息（包括积分）

id：身份证号码

interest：兴趣，目前根据签名所属行业划分，为人工打上，例如某奢侈品牌签名，那么兴趣直接填“奢侈品（xx旗下）”

complexDetail：综合信息，【百度】签名下面有一条短信包含以下所有内容时使用：学历，年龄，姓名，电话，拆迁信息，违章记录，酒驾记录

departure：出发地

arrive：到达地

visLoc：曾拜访地址

shoesSize：穿的鞋码

balance：账户余额

fitnessProgram：健身课

eat：喜欢吃的东西

movieTicketNum：电影票数

movieName：电影名称

movieCombo：电影票套餐名称

houseAcPay：累计已付房款

houseOneTimePay：一次性付清房款

houseDeposit：购房押金

houseSoldPrice：房产成交价格

houseInfo：是否有房

food：点过吃过的食物

restaurant：吃过的餐馆

medicalInsurance：医疗险

criticalIllnessInsurance：重疾险

clInsurance：交强险金额

cvInsurance：车辆商业险金额

vehicleTax：车船税金额

carInsuranceTotal：车险总金额

carInsuranceCoverageDetail：车辆保险明细信息

carInsuranceDue：车险到期日期

loanAndDue：借款还款日期

loanDate：借款日期

loanDue：还款日期

creditDue：信用卡账单到期日期

outpatientDate：面诊日期

travelDate：出行日期

withdrawDate：提现日期

purchaseDate：购买日期

workoutDate：锻炼日期

movieDate：观影日期

salaryWithdraw：正式工作薪资提现金额

investWithdraw：投资提现金额

loanWithdraw：借贷提现金额

transactionWithdraw：交易提现金额

volunteerInfo：是否参加过志愿者活动

volunteerDetail：志愿者项目名称

volunteerWithdraw：志愿者报酬提现金额

fundraiseInfo：是否发起过募捐

fundraiseTopic：募捐项目名称

fundraiseWithdraw：募捐提现金额

customer：号码主的客户

CREATE TABLE `persona` (

`pk` bigint(12) NOT NULL AUTO\_INCREMENT COMMENT '自增主键',

`mobile` bigint(12) DEFAULT NULL COMMENT '电话号码',

`signature` varchar(12) CHARACTER SET utf8 COLLATE utf8\_general\_ci DEFAULT NULL COMMENT '短信签名',

`name` json DEFAULT NULL COMMENT '名字、名称、称呼、姓氏',

`ids` json DEFAULT NULL COMMENT '身份证',

`gender` char(1) CHARACTER SET utf8 COLLATE utf8\_general\_ci DEFAULT NULL COMMENT '性别',

`job` json DEFAULT NULL COMMENT '职业、被招聘平台推荐的职业、是否参加过志愿者活动、志愿者活动项目名称',

`outfit` json DEFAULT NULL COMMENT '鞋码',

`eat` json DEFAULT NULL COMMENT '点过吃过食物、吃过的餐馆',

`watch` json DEFAULT NULL COMMENT '电影名字、电影票数量、电影套餐',

`activity` json DEFAULT NULL COMMENT '是否发起过募捐、募捐项目名称、募捐提现金额',

`fitness` json DEFAULT NULL COMMENT '健身课',

`salary` json DEFAULT NULL COMMENT '正式工作薪资提现金额、志愿者报酬提现金额',

`property` json DEFAULT NULL COMMENT '一次性付清购房款、累计已付购房款、购房押金、房产成交价格、物业和维修支出',

`car` json DEFAULT NULL COMMENT '车牌、车品牌、车辆违章记录、有无车辆、车辆商业险金额、交强险金额、车船税金额、车险总金额、车辆保险明细信息',

`insurance` json DEFAULT NULL COMMENT '医疗险、重疾险',

`date` json DEFAULT NULL COMMENT '车险到期日期、借款还款日期、借款日期、还款日期、信用卡账单到期日期、面诊日期、出行日期、提现日期、购买日期、锻炼日期、观影日期',

`location` json DEFAULT NULL COMMENT '现居住地、常去地址、拜访过地址',

`trip` json DEFAULT NULL COMMENT '出行方式、入住酒店信息、去过的旅游景点、会员信息、出发地、到达地',

`cashflow` json DEFAULT NULL COMMENT '投资理财金额、逾期金额、现金流相关的银行信息、入账金额、贷款金额、还款金额、投资提现金额、贷款提现金额',

`outpatient` json DEFAULT NULL COMMENT '面诊医院、医疗账单总金额、医疗账单待付金额、医保报销金额',

`kid` json DEFAULT NULL COMMENT '姓名、读书年级、所上学校、有无小孩',

`interest` varchar(20) CHARACTER SET utf8 COLLATE utf8\_general\_ci DEFAULT NULL COMMENT '签名所属行业分类',

`transaction` json DEFAULT NULL COMMENT '购买物品、购买金额、购买药品、余额、交易提现金额',

`other` json DEFAULT NULL COMMENT '综合信息、号码主的客户',

`msg` mediumtext CHARACTER SET utf8 COLLATE utf8\_general\_ci COMMENT '短信文本',

`last\_update\_time` datetime(6) NOT NULL DEFAULT CURRENT\_TIMESTAMP(6) COMMENT '当前记录的最后一次更新时间YYYY-MM-DD',

PRIMARY KEY (`pk`),

KEY `last\_update\_time` (`last\_update\_time`) USING BTREE

) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4 COLLATE=utf8mb4\_general\_ci ROW\_FORMAT=COMPRESSED COMMENT='画像基础数据表'