博雅大数据学院 | 数据酷客

北京市大数据人才培训示范基地

数据科学“云实训”项目训练营

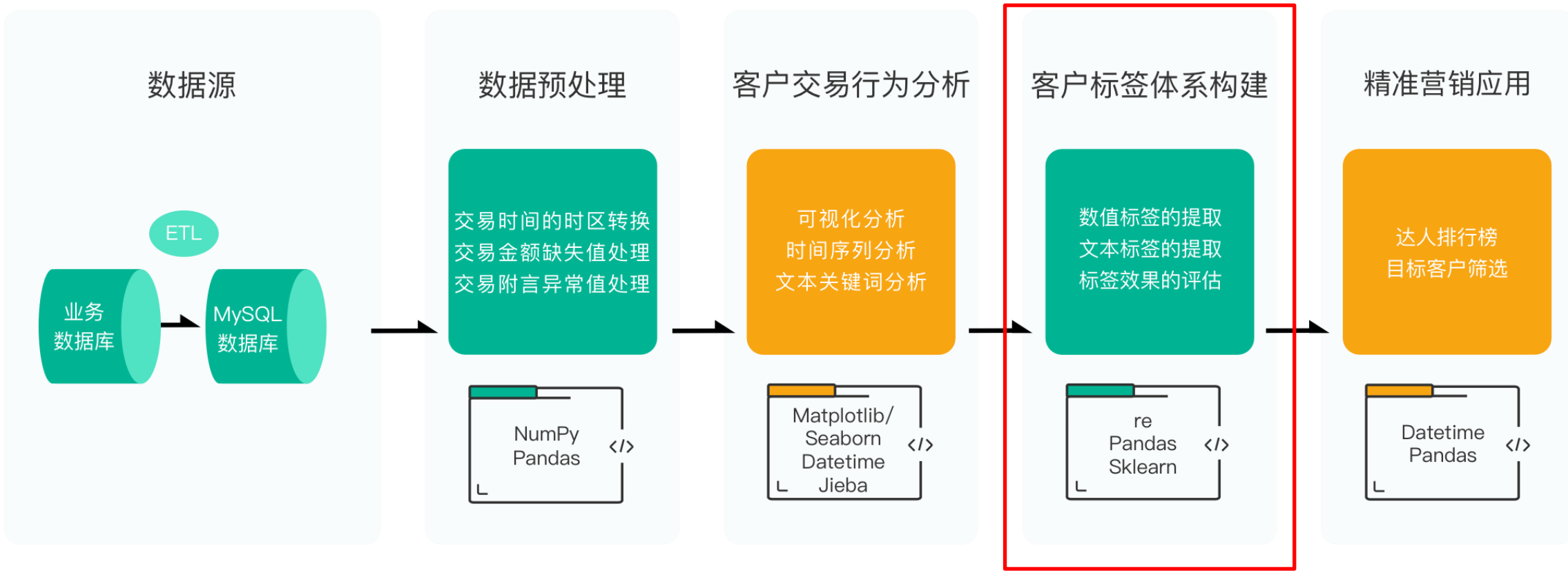
第五课：预测类标签与文本类标签构建

讲师：张嘉田

数据酷客

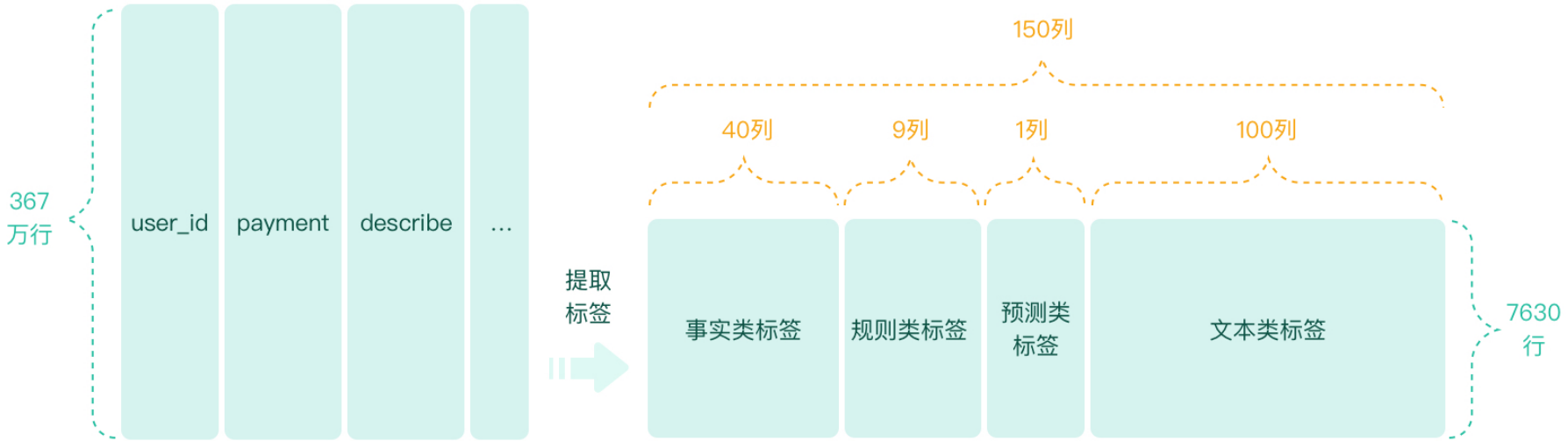
数据科学人工智能

项目流程


数据源
ETL
业务
数据库
MySQL
数据库
数据预处理
交易时间的时区转换
交易金额缺失值处理
交易附言异常值处理
NumPy
Pandas
客户交易行为分析
可视化分析
时间序列分析
文本关键词分析
Matplotlib/
Seaborn
Datetime
Jieba
客户标签体系构建
数值标签的提取
文本标签的提取
标签效果的评估
re
Pandas
Sklearn
精准营销应用
达人排行榜
目标客户筛选
Datetime
Pandas

标签构建流程

- 预测类标签：原始数据中不能直接提取，需要借助模型进行预测的标签，例如客户价值等级
- 文本类标签：从客户交易记录的文本中提取的关键词，也可用于描述客户偏好，将此部分关键词作为文本类标签，例如彩票、儿童、孕妇、基金等



预测类标签构建流程

根据提取出的事实类和规则类标签建立模型，对没有客户价值等级标签的客户进行预测

- 针对字符型特征，需要进行数值编码，以便模型能更好的处理数据
- 处理连续型数据可以将数据进行离散化，使模型对异常数据有更强的鲁棒性，同时也能降低模型运算复杂度，提升模型运算速度
- 将有客户价值等级的数据随机划分为训练集与测试集，训练逻辑回归模型
- 使用训练好的逻辑回归模型，对客户价值等级未知的客户进行预测

特征处理

- One-Hot编码：
  - 将包含K个取值的字符型特征转换成K个取值为0或1的二元特征

样本	国家
0	澳大利亚
1	新西兰
2	加拿大

One-Hot

样本	国家_澳大利亚	国家_新西兰	国家_加拿大
0	1	0	0
1	0	1	0
2	0	0	1

特征处理

- 等距离散化：
  - 根据连续型特征的取值，将其均匀地划分为k个宽度近似相等的区间
  - 根据每个客户该特征的取值，相应地划入对应的区间，每个区间赋予一个标签

- 等频离散化：
  - 特征数据总量为n，将其划分为k个区间段，使得每个区间段包含的数据个数为$\frac{n}{k}$，离散化后区间内的数据量尽可能均衡
  - 根据每个客户该特征的取值，相应地划入对应的区间，每个区间赋予一个标签

训练集和测试集划分

- 特征变量X：预测所需的特征，使用所有事实类和规则类标签(共49个)作为X
- 目标变量y：需要预测的特征，客户价值等级为y

Sklearn中的model_selection模块中的train_test_split()函数，用作训练集和测试集划分

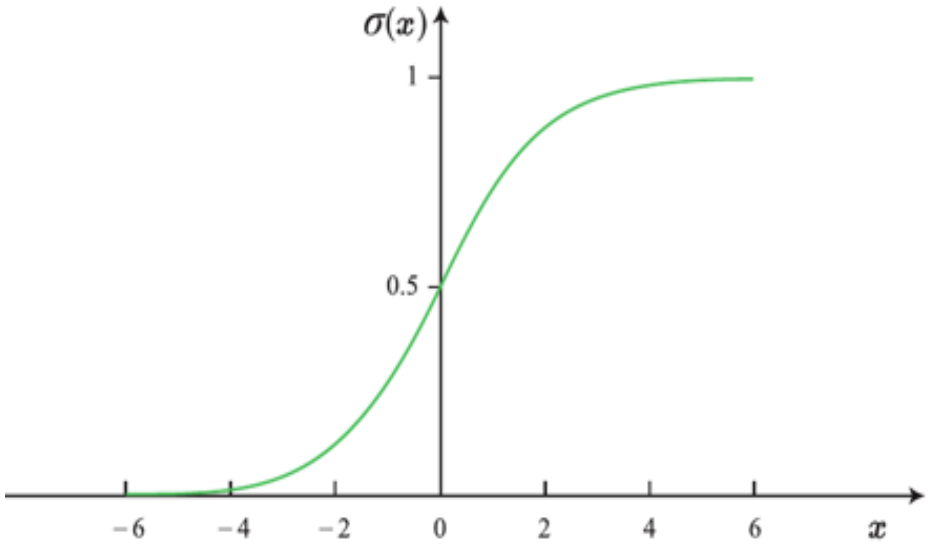
```
train_test_split(x,y,test_size = None,random_state = None)
```

- X,y: 分别为预测所需的所有特征，以及需要预测的特征(即客户价值等级)
- test_size: 测试集比例，例如test_size=0.2则表示划分20%的数据作为测试集
- random_state: 随机种子，因为划分过程是随机的，为了进行可重复的训练，需要固定一个random_state，结果重现
- 函数最终将返回四个变量，分别为X的测试集和训练集，以及y的测试集和训练集

预测模型：逻辑回归

- 逻辑回归引入Sigmoid函数，将连续型的输出映射到区间(0,1)， Sigmoid 函数如下所示

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



- 当输入x很大或很小时，该函数以接近于0或1的值输出
- 在逻辑回归中，若 $y_i \in \{1, -1\}$，则输出可以解释为样本属于正类的概率

$$P(y_i|\boldsymbol{x}_i) = \frac{1}{1 + e^{-\boldsymbol{y}_i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i}}$$

模型评估和预测

$$Accuracy = \frac{\text{所有正确预测的样本数}}{\text{总样本数}}$$

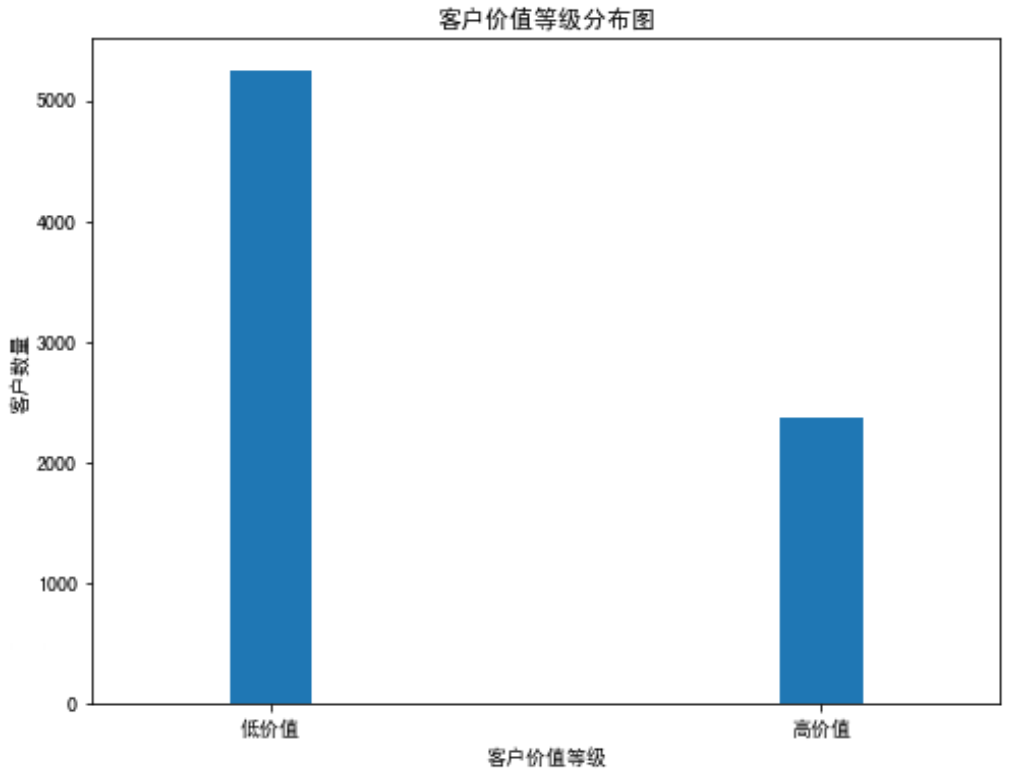
$$Precision = \frac{\text{预测为正类且正确预测的样本数}}{\text{所有预测为正类的样本数}}$$

$$Recall = \frac{\text{预测为正类且正确预测的样本数}}{\text{所有真实情况为正类的样本数}}$$

$$F_1Score = \text{精确率和召回率的调和平均数} = 2\frac{Precision \times Recall}{(Precision+Recall)}$$

将训练好模型对价值等级未知的客户进行预测，

将结果进行拼接，客户价值等级分布如右图：

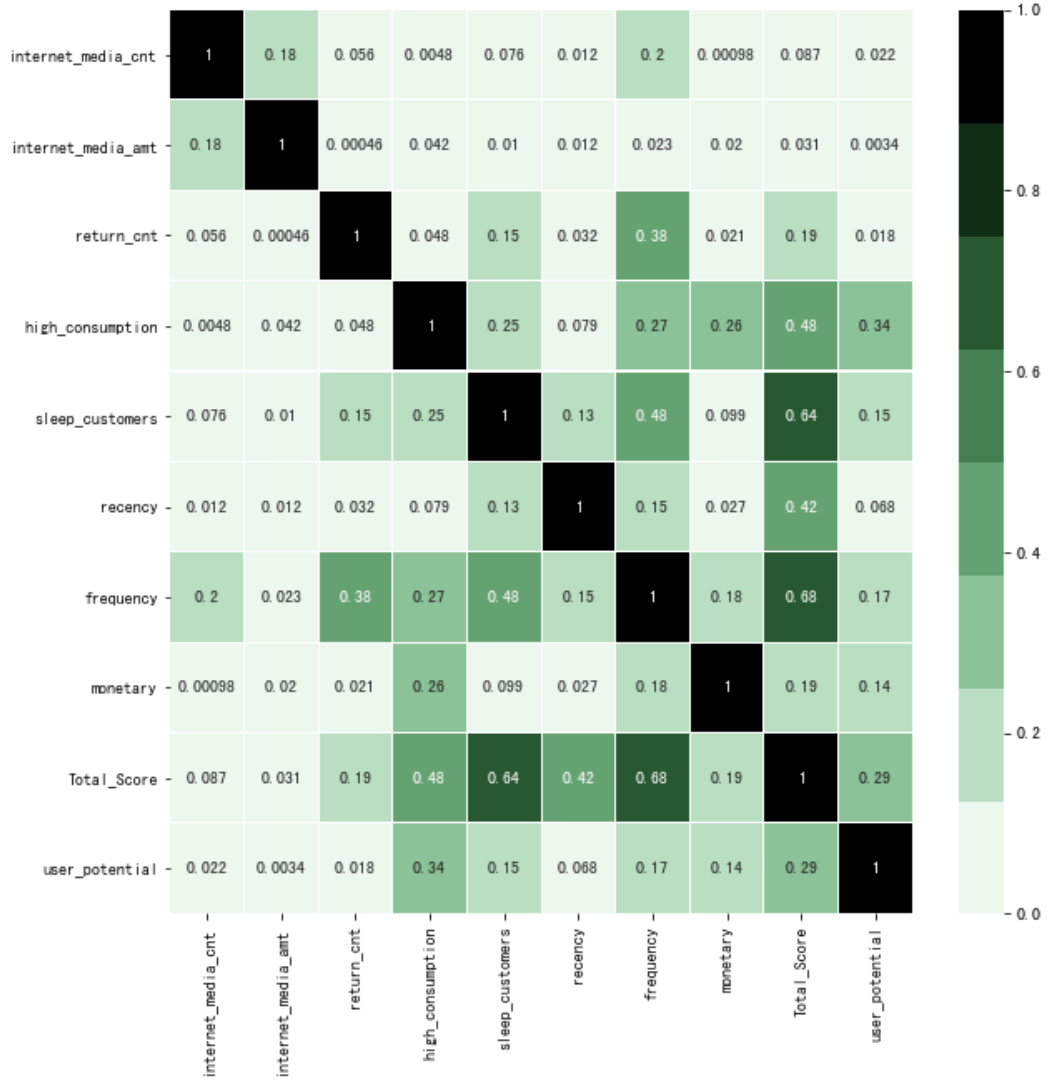


标签间相关系数矩阵可视化：热力图

皮尔森相关系数是用来反映两个变量线性相关程度的统计量，是一种线性相关系数，一般用r表示，r的取值在-1与1之间

- 若$r > 0$，表明两个变量是正相关，即一个变量的值越大，另一个变量的值也会越大
- 若$r < 0$，表明两个变量是负相关，即一个变量的值越大，另一个变量的值反而会越小
- r的绝对值越大表明相关性越强

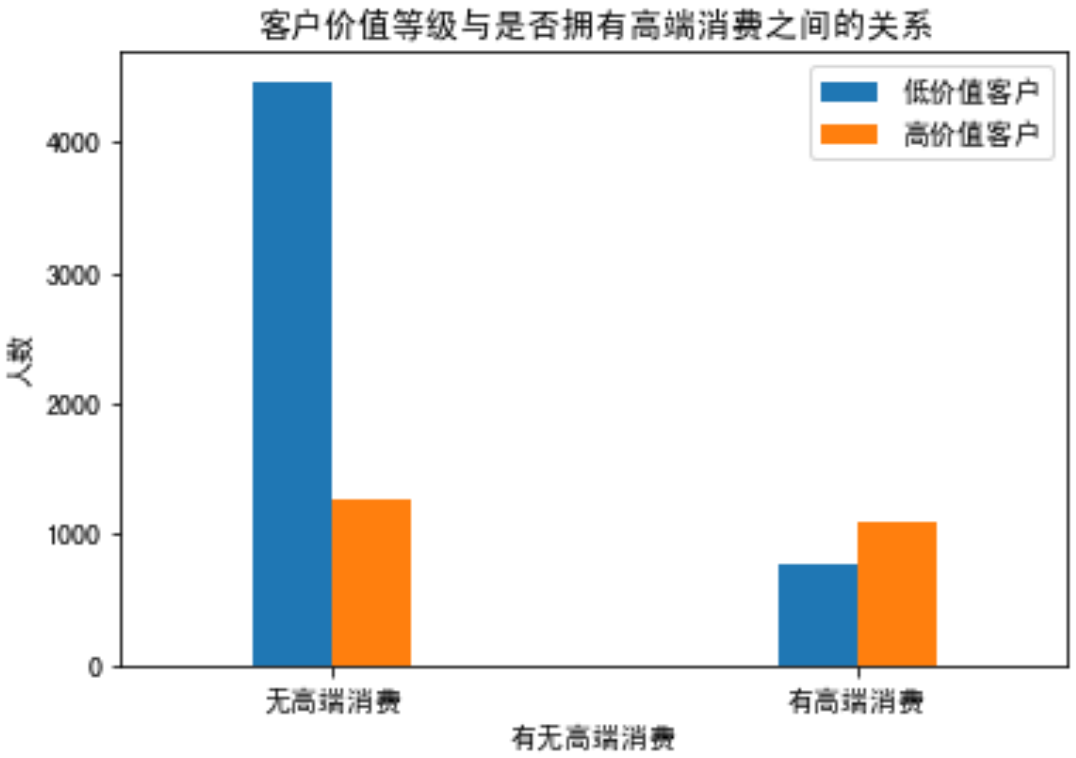
热力图是数据可视化中比较常用的显示方式，颜色的深浅表示了描述量的大小



交叉表

交叉表是一种常用的分类汇总表格，用于频数分布统计对于多重分组计算个数的问题，通常使用交叉表来计算

分组特征	低价值客户	高价值客户
无高端消费	4470	1275
有高端消费	779	1106



文本标签的提取方法

项目中使用了两种文本特征提取方式，分别为CountVectorizer()和TfidfVectorizer()：

- CountVectorizer() 对于每一个训练文本，它只考虑每种词汇在该训练文本中出现的频数
  CountVectorizer会将文本中的词语转换为词频矩阵，它通过fit_transform()函数计算各个词语出现的次数

  不考虑词法和语序，每个词语相互独立

  Mary wants to go to Japan.

  Bill wants to go to Germany.

  建立一个词典用于构建特征向量

  [Mary，wants，to，go，Japan，Bill，Germany]

  向量每个位置表示的单词与上面的数组一致，值为该单词在句子中出现的次数

  第一个句子： [1，1，2，1，1，0，0]　　第二个句子： [0，1，2，1，0，1，1]

文本标签的提取方法

- `TfidfVectorizer()` 评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度，字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降

$$tf(t) = \frac{\text{词}t\text{在文档中出现的次数}}{\text{文档的总词数}}$$

$$idf(t) = \log\left(\frac{\text{语料库中文档的总数}}{\text{包含词}t\text{的文档数}+1}\right)$$

$$tf\text{-}idf(t) = tf(t) \cdot idf(t)$$

文本标签的分类

文本特征	对应人群
停彩、大乐透、双色球、福利彩票、彩票、竞彩、追号	爱好购买彩票的人群
儿童、卡通、可爱、女童、零食	家庭中有孩子的人群
男士、衬衫、透气、真皮、足球、运动、小米、汽车、电子、汽车	广大的男士人群
修身、专柜、保湿、女士、女装、显瘦、打底、蕾丝、欧美、韩版	爱美的女士人群
婴儿、孕妇、宝宝	家庭中有孕妇/婴儿的人群
分期、贷款、金融服务	有借贷分期需求的人群
基金、投资、证券	有投资需求的人群

描绘用户画像

对客户进行分析，对客户的交易附言信息进行整合，画出交易附言的词云分布图

出现很多护肤品、化妆品类的词汇，如洗面奶、控油、波斯顿等，可以大致推断客户经常接触此类商品

数据酷客

数据科学人工智能

加入数据酷客交流群