# 朋友圈分析

## 1 相关接口说明

## 2 朋友圈文本主题分布分析

### 2.1 朋友圈主题分布分析

通过文本聚类的方法，初步了解朋友圈主题分布，进而开展后续工作。

### 2.2 聚类方法

采用K\_MEANS 聚类，分别选取主题类别为100，500，1000进行聚类分析。

2.3 文本特征表示方法

测试并通过聚类结果分析了以下模型：

1. 词袋模型：短文过于稀疏，不适用
2. LDA模型：对于短文本主题分布与长文差异性较大，基本不适用
3. ParagraphVec: 采用关键词抽取、词空间聚合、神经网络训练出句子的向量表示，聚类结果较好。

2.4 文本标注工作

通过大量的文本标注工作，确定类别候选集合。

## 3 朋友圈文本主题分布分析

这一类多位用户转发或者发布的功能性微博，并不能代表用户真实的兴趣，因此在用户兴趣建模中用该予以单独分类。

#### 3.2.2.2 长微博分类研究

长微博分类主要包括二级分类和三级分类长微博选取的是分词去停用词后文本长度为50词以上的文本，对长微博分析发现长微博几乎全部为广告和短评类文本，因此二级分类采用二分类的方法，本文尝试了朴素贝叶斯方法和SVM分类。

1. 朴素贝叶斯方法
2. 样本抽取和标注

为了克服大量的样本标注工作的困难，本文采用了一种半自动标注方法。对于广告和短评这两类特征比较鲜明的类别，首先采用了模式匹配的方法标注出一批可靠性较强的初步训练样本，初步进行抽样评估，结果如下：

表3-1 分类样本初步评估表

|  |  |  |
| --- | --- | --- |
| 类别 | 数量 | 准确率 |
| 广告类 | 70，000 | 95% |
| 短评类 | 100，000， | 85% |

然后通过贝叶斯分类器本身对标注样本进行预测，选取预测结果不变的那一部分，抽样评估如下：

表3-2 分类样本评估表

|  |  |  |
| --- | --- | --- |
| 类别 | 数量 | 准确率 |
| 广告类 | 42，000 | 98% |
| 短评类 | 60，000， | 95% |

由上表可知，在经过分类器分类之后，样本的准确率得到了较大的提升，但数量保持在同一个数量级上，因此选择评估后准确率较高的部分作为训练集合。

1. 贝叶斯方法用于文本分类

短文本类别主要取决于关键组成元素，出现某一词语W的情况下，该文本属于第i 类的概率：

为第 i 类中 W词出现次数， N为第i类的样本数目。也就是用W词的统计值作为条件概率值。假设为属于于该类，为不属于该类，先验概率为;

这里仅仅是针对一个词的结果，但是文本你的类别归属往往是由文本的若干个词所决定的，把上两公式扩展到多个词后：

在短文本中往往几个关键词便可决定类别归属，这里取n=5，那么属于第该类的归一化概率为：

这里概率的取值范围为(0,1)，P越大表明属于该类的概率越大。

1. SVM文本分类
2. 特征选择与文本特征表示

传统的词袋模型，本文选取的特征是词语的one-hot表示形式，即每篇文档都表示成一个V维向量（V是词典的大小），向量的每一个维度上的值便是该维度上词出现的词出现的次数。这样的表示方法存在着两个问题：

1. 特征矩阵过于稀疏，特别是对于短文本而言，往往只有几十个词语组成，词典的常用词大约为10万左右，也就说矩阵的非零率不到千分之一，而且高维度向量在计算上也存在着耗资源过多的问题，计算效率低。
2. 噪声信息干扰较大，微博中往往无用的常用词等所占比重较大，在这样的特征表示下，信息会在常用词维度较为集中，在计算相似度等方面会造成偏差。

为了解决以上两种问题，采用特征选择的方法对语料集词库进行特征选取工作，这里采用卡方检验的方法进行特征选择。具体步骤如下图：

1. 分别统计每个类别下的每个词的以下四个统计量：
2. 在该分类下包含这个词的文档数量。
3. 在该分类下包含这个词的文档数量。
4. 在这个分类下不包含这个词的文档数量。
5. 不在该分类下，且不包含这个词的文档数量。
6. 在每一类别中根据以上四个统计量，利用卡方检验分别计算出相关性最高的K个词。
7. 合并每个类别的K个词，形成最终的大小为N词典。

图3-4 特征选择步骤图

那么每一篇文本都将表示为一个N为向量，向量的每一维度对应着字典中该词在本文中的TF-IDF值。

1. SVM分类

SVM分类算法是一种分类效果极强的分类算法，在文本分类上也有较广的应用

1. 长微博分类结果
2. Bayes 分类结果