第四次实验报告

161278015 李康

一、实验内容：

•主题：上市公司财经新闻情感分析
•实验目标：
使用多种机器学习算法对文本进行情感判别，包括 KNN、决策树、朴素贝叶斯、支持向量机等，学习如何进行模型训练，如何进行分类预测。要求使用至少两种分类方法。
•要求：核心程序在 MapReduce 上运行，要求使用至少两种分类方法

二、实验流程

2.1 总览

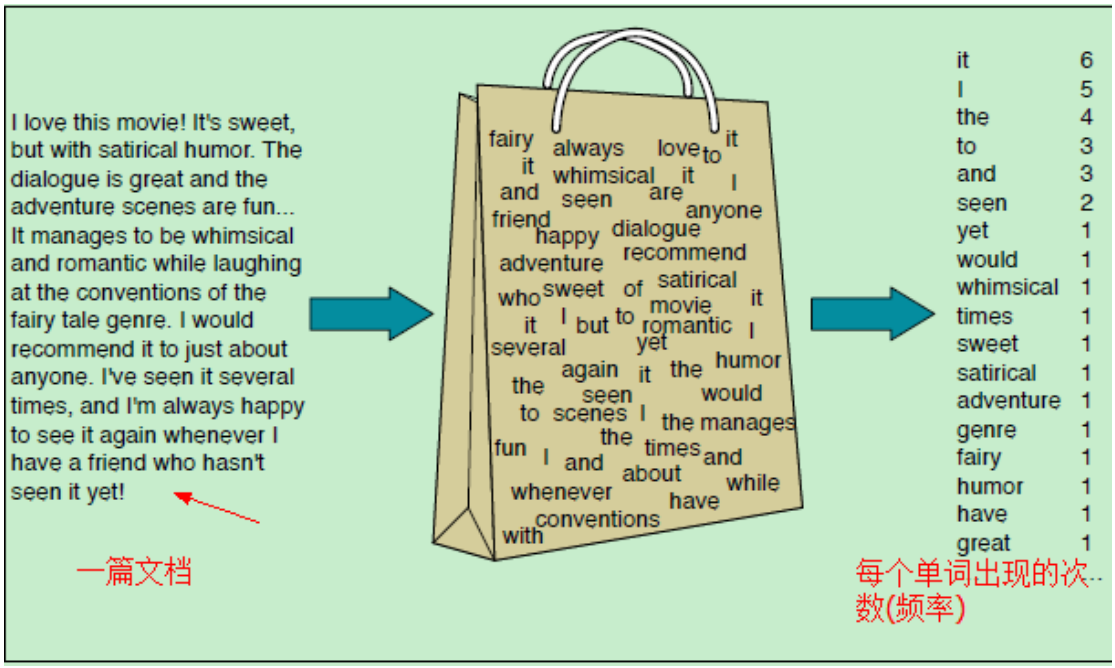
- **首先对给定的数据集和训练集进行数据预处理，生成词向量以及必要的键值对。**
- **接着用两种算法对训练集进行了预测**
  - **KNN**
    - 计算测试数据与各个训练数据之间的距离；
    - 按照距离的递增关系进行排序；
    - 选取距离最小的 K 个点；
    - 确定前 K 个点所在类别的出现频率；
    - 返回前 K 个点中出现频率最高的类别作为测试数据的预测分类。
  - **朴素贝叶斯**
    - 利用 chi_words 先在训练集中计算每个词的 p(a|positive)、p(a|neutral)、p(a|negative)。
    - Pi 为在 i 类别中出现的文件数/总的出现文件数。
    - 在测试集中计算每个词出现的词频，用词频*每个词的 Pi 作为指标，将所有词该指标相乘
    - 取最大值，即为测试集所属分类

2.2 算法介绍

2.2.1KNN

K Nearest Neighbor 算法又叫 KNN 算法，这个算法是机器学习里面一个比较经典的算法， 总体来说 KNN 算法是相对比较容易理解的算法。其中的 K 表示最接近自己的 K 个数据样本。KNN 算法和 K-Means 算法不同的是，K-Means 算法用来聚类，用来判断哪些东西是一个比较相近的类型，而 KNN 算法是用来做归类的，也就是说，有一个样本空间里的样本分成很几个类型，然后，给定一个待分类的数据，通过计算接近自己最近的 K 个样本来判断这个待分类数据属于哪个分类。可以简单的理解为由那离自己最近的 K 个点来投票决定待分类数据归为哪一类。

2.2.2 朴素贝叶斯



在文本分类中，我们面临的问题是给定一个文本 x⃗=[x1,x2,...,xi,...,xn]x→=[x1,x2,...,xi,...,xn]，其中 xixi 从原始文本抽出来的一个特征，可以是单个单词或者是一个 ngram 特征，或者是一个正则表达式特征。我们希望有一个模型可以来预测这个特定文本的标签 y，在邮件垃圾分类中，y 可以是指”垃圾邮件”或”非垃圾邮件”。

而朴素贝叶斯是一种生成模型，也就是它对问题进行联合建模，利用概率的乘法法则，我们可以得到：

$$
\begin{aligned}
p(y,\vec{x}) &= p(y,x_1,x_2,\ldots,x_n) \\
&= p(y)p(x_1|y)p(x_2|x_1,y)p(x_3|x_1,x_2,y)\ldots p(x_n|x_1,\ldots,x_{n-1},y) \\
&= p(y)p(x_1|y)\prod_{i=2} p(x_i|x_1,\ldots,x_{i-1},y)
\end{aligned}
$$

2.3 算法实现

2.3.1 数据处理实现

```
public class DATACREATION{

    public static class DCMapper{

        public void map{
```

```
            分词，并根据指定文件，制作词向量；
            保存的格式为
           <词向量,标签>
        }
    }
    public static class DC_Another_Mapper{
        public void setup{}
        public void map{
            对新闻标题数据预处理
            分词，制作词向量
            保存格式为
            <新闻标题，词向量>
        }
    }
}
```

2.3.2KNN 实现

```
public class KNN{
    public static int getdistance(String a,int[] b){}
    //计算两个点之间的距离

    public static String[] GET_TRIAN_DATA{}
    //获得训练集数据

    public static class KNNMapper{
```

```
        public void setup(){}
        public void map(){
            计算测试词向量到各个训练词向量的距离
            取最近的三个（K=3）向量，统计 label
            投票票数最多的便是预测结果
        }
    }
}
```

2.3.3 朴素贝叶斯实现

```
public class BAYES{
    //首先进行训练
    public static class TrainMapper{
        public void map(){
            //输入训练集
            //统计词向量中每个词的属性（局部）
            //比如 “下滑”: <pos,15%> <neu,15%> <neg,70%>
        }
    }
    public static class TrainReducer{
        public void reduce(){
            //统计词向量中每个词的属性（全局）
        }
    }
```

```
    //然后进行预测
    public static class TestReducer{
        public void reduce(){
            //输入测试样本数据集
            //计算 三个词频*属性，取最大值，贴上标签
            //输出分类结果
        }
    }
}
```

3、实验结果展示

```
A股1年1.6万亿募2400亿:银行股半年融资3千亿 元大最强        positive
上市银行屡吃罚单 中介机构:不构成重大违法违规    positive
专题报告：环杭州湾大湾区与粤港澳湾区的对比分析  positive
估值修复遇上业绩超预期 银行股演绎“王者归来”    positive
健康元隐瞒多次环保处罚 324亿市值蒸发配股股民苦          negtive
南京银行核心一级资本充足率告急 拟定增140亿元补血        positive
历史遗留问题拖累业绩 正虹科技重组预期降温       neutral
同业存单减速 中小行最受伤    negative
国家队增持南京银行 持股4.99%逼近举牌线         positive
探访茅台实体店：最严限价令下的失衡供需          negative
浦发银行同业业务违规被罚款200万 违反审慎经营规则        negative
浦发银行谢伟：银行资管业务未来发展空间巨大       positive
深交所与浦发开展战略合作 推动金融和科技深度融合         positive
茅台成25年A股第一高价股 总市值超半个贵州GDP    positive
跨界大王群兴玩具:我有故事你带钱 九年盈利不抵减持        positive
这些国家队重仓股 还被机构看好          positive
透视银行股大涨之谜：国家队增持南京银行逼近举牌线        positive
郑煤机(00564-HK)认购浦发银行2亿元人民币理财产品        positive
银行股脉冲式上演逆袭行情 工商银行股价创下9年新高        positive
麦当劳今天宣布：2700多家餐厅归中国公司了       neutral
```

4、实验总结

通过本次实验，我了解了数据挖掘的入门知识。学习使用 KNN、朴素贝叶斯等方法，去实现对文本的情感分析。同时也学习了如何进行模型的训练，如何进行调参和代码优化。
但是在本次实验中，也暴露出一些值得注意的问题。实际运行过程中，发现 KNN 的运行效率很低，预测结果也不是很满意。相较于朴素贝叶斯，更加消耗计算机的计算资源。

5、附代码架构

数据处理实现

```java
public class DATACREATION{
    public static class DCMapper{
        public void map{
            分词，并根据指定文件，制作词向量;
            保存的格式为
            <词向量,标签>
        }
    }
    public static class DC_Another_Mapper{
        public void setup{}
        public void map{
            对新闻标题数据预处理
            分词，制作词向量
            保存格式为
            <新闻标题，词向量>
        }
    }
}
```

KNN实现

```java
public class KNN{
    public static int getdistance(String a,int[] b){}
    //计算两个点之间的距离

    public static String[] GET_TRIAN_DATA{}
    //获得训练集数据

    public static class KNNMapper(){
        public void setup(){}
        public void map(){
            计算测试词向量到各个训练词向量的距离
            取最近的三个（K=3）向量，统计label
            投票票数最多的便是预测结果
        }
    }


}

```

Bayes实现

```java
public class BAYES{
    //首先进行训练
    public static class TrainMapper{
        public void map(){
            //输入训练集
            //统计词向量中每个词的属性（局部）
            //比如“下滑”: <pos,15%> <neu,15%> <neg,70%>
        }
    }
    public static class TrainReducer{
        public void reduce(){
            //统计词向量中每个词的属性（全局）
        }
    }

    //然后进行预测
    public static class TestReducer{
        public void reduce(){
            //输入测试样本数据集
            //计算 三个词频*属性，取最大值，贴上标签
            //输出分类结果
        }
    }
}
```