实验四

12.26

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。

- 实现：
- KNN
- Naïve Bayes

逻辑结构

- 数据处理
- KNN
  - 训练
  - 预测
- Bayes
  - 训练
  - 预测

遇到的问题：

- Double变量精确度问题
- 导致最后统计出来的概率都是0
- 采用统计前500个高频词的方法

数据处理实现

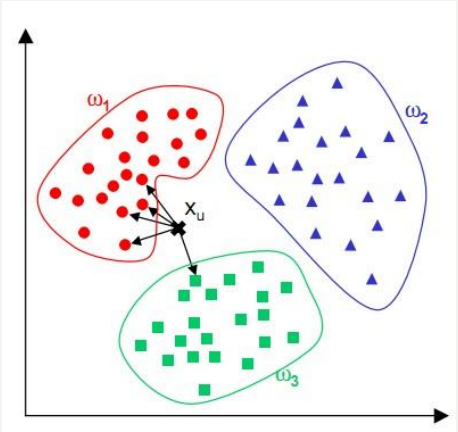
```java
public class DATACREATION{
    public static class DCMapper{
        public void map{
            分词，并根据指定文件，制作词向量;
            保存的格式为
            <词向量,标签>
        }
    }
    public static class DC_Another_Mapper{
        public void setup{}
        public void map{
            对新闻标题数据预处理
 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

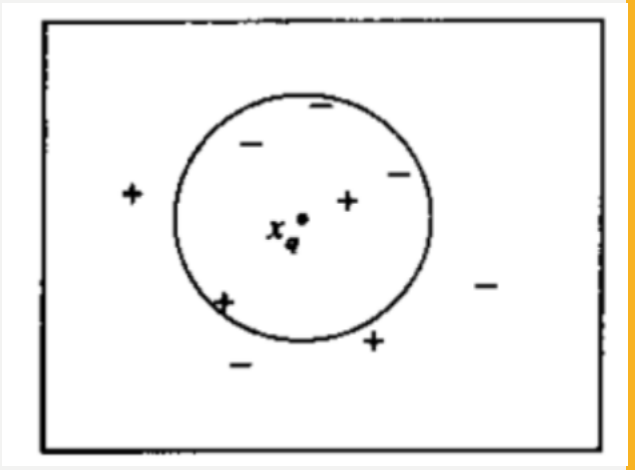
这里没必要用到reduce
特征选择：TF-IDF
稀疏矩阵：统计词频，
选前500个

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
        }
    }


}

```

KNN介绍





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
        }
    }
}
```

Bayes介绍

$$p(x|c_j) = p([w_1, w_2, ..., w_{|x|}]|c_j) = \prod_{h=1}^{|x|} p(w_h|c_j)$$

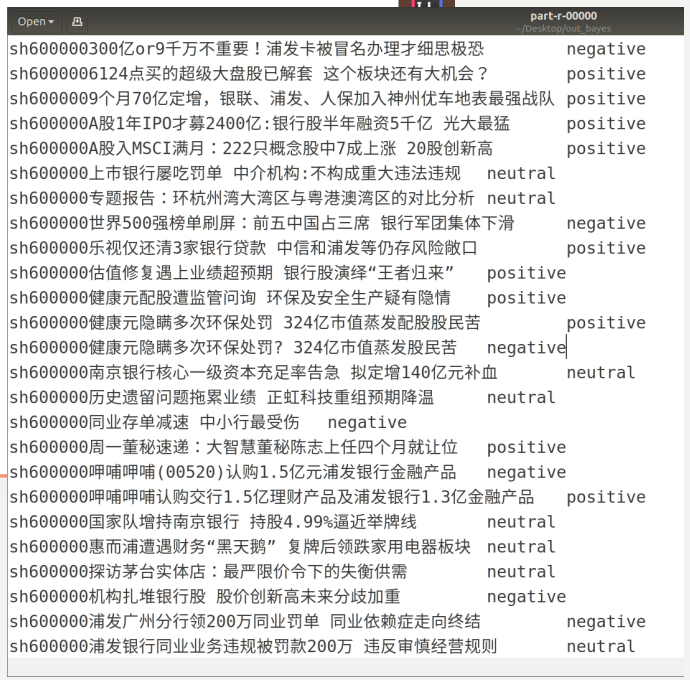
$$= \prod_{i=1}^{V} p(t_i|c_j)^{N(t_i,x)} = \prod_{i=1}^{V} \theta_{i|j}{}^{N(t_i,x)}$$

$$p(y = c_j) = \pi_j$$

$$P(C \mid F_1) = \frac{P(CF_1)}{P(F_1)} = \frac{P(C) \cdot P(F_1 \mid C)}{P(F_1)}$$

A股1年1个IPO募2400亿:银行股半年融资3千亿 无人最猛 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
春光荣今天宣布：2700多家餐厅归中国公司了 neutral

朴素贝叶斯全数据集结果



正确率验证

KNN	32/50	68/100
贝叶斯	28/50	71/100

在验证正确率的时候，不考虑中性词，因为不好做出判断