第四次实验报告

161278049 张昕

**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。

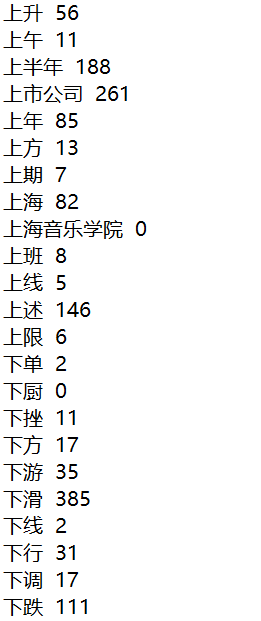
**程序流程设计：**

**KNN:**

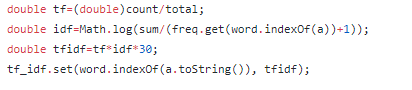
**预处理：（MapReduce下进行）**

1. 以chi\_words.txt中的词作为特征，统计其中每个词在pos、neu、neg中出现的次数。

2. 统计完成后，删去在三个属性集中出现次数均为0的词。以neg集为例，结果如下：

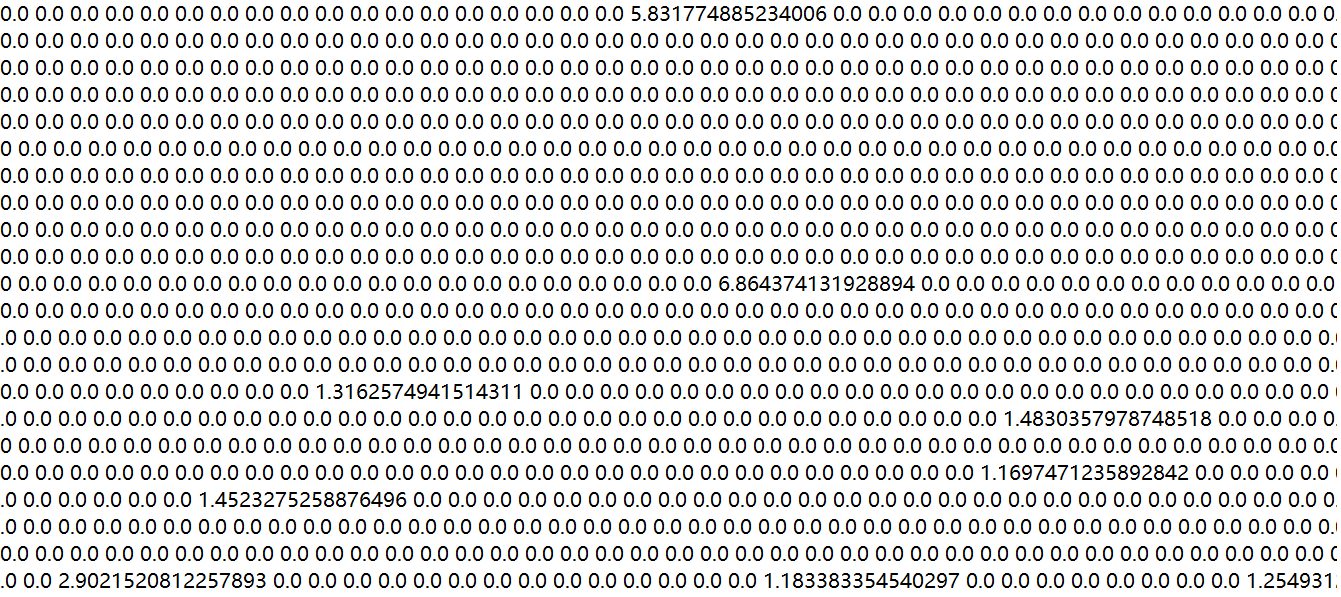


3. 计算每一个属性集的每一个词的TF-IDF值，公式为：



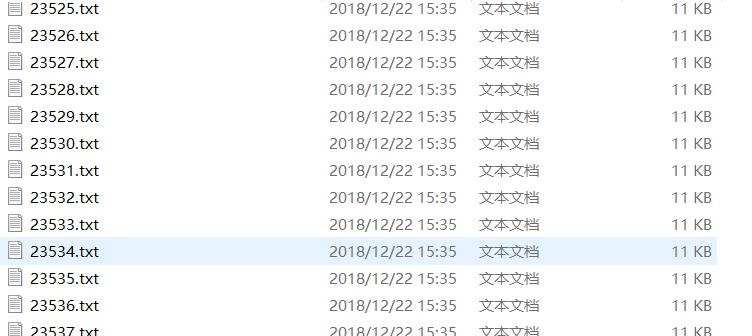
（为了使特征更加明显，这里将每个TF-IDF值扩大30倍处理）

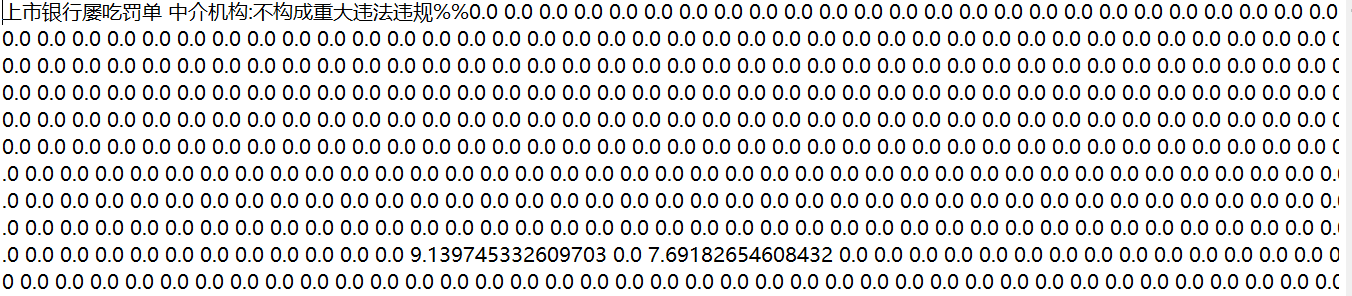
得到的向量如下：



由于特征词较多，所以向量显得过于稀疏

4. 以同样的方式处理待预测数据，将每一条新闻的特征向量写入一个文件中

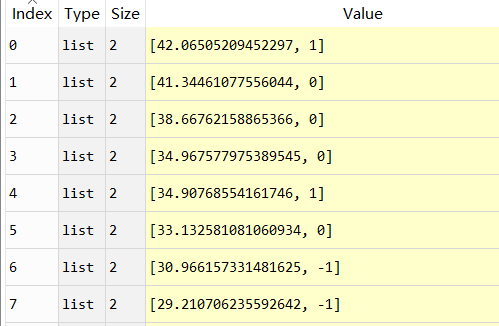




**KNN过程：**

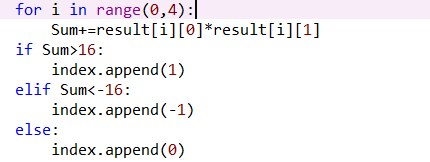
1. 将每条新闻的特征值向量转置和每一条样本特征向量相乘

2. 所得值从大到小排序，取前四个结果，如图



其中，1代表pos，0代表neu，-1代表neg，作为属性标记

3. 前四个结果乘以其对应的属性标记，求和，当和大于16时，记为pos属性；小于-16时记为neg属性；否则为neu属性。



**结果如下：**

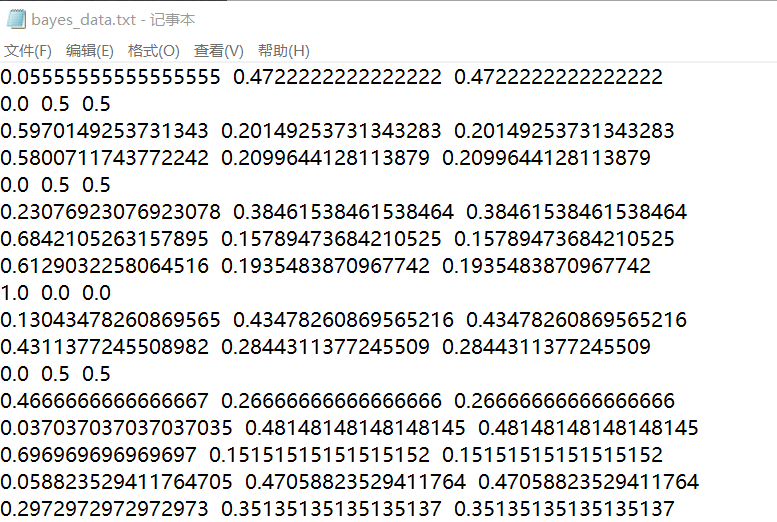


**所表现的结果较好。**

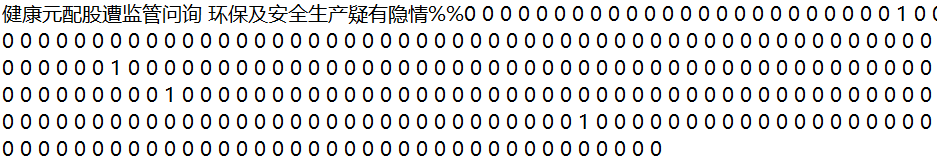
**朴素贝叶斯：**

**预处理：（MapReduce下进行）**

1. 求出样本里每个属于chi\_words的词出现在三个属性里的概率，如“下滑”在属性集[pos, neu, neg]中出现的概率分别为：[8.8%，8.6%，82.6%]



2. 求出chi\_words中每个词在每条新闻中的词频

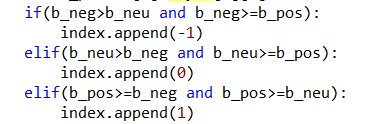


**NaiveBayes：**

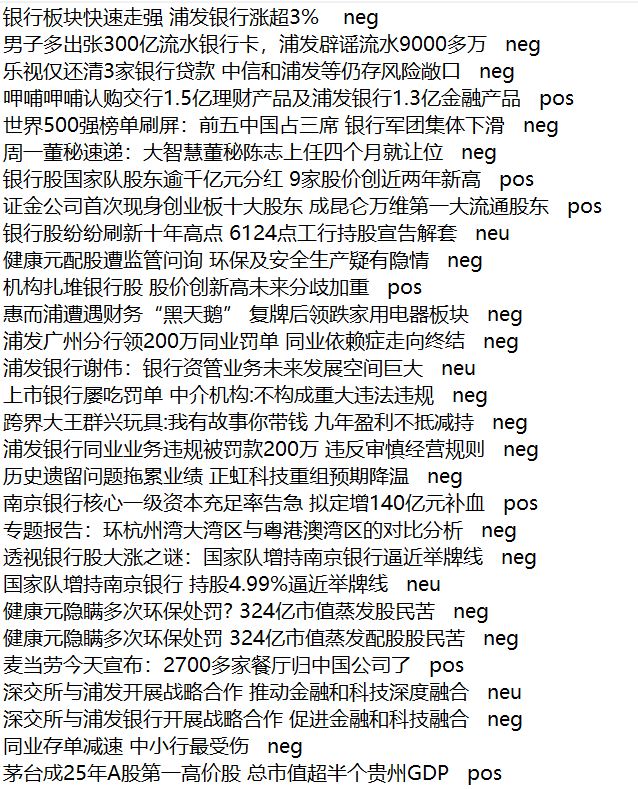
1. 每个词在该新闻中的词频分别乘以该词在三个属性集中对应的概率，作为评估值。

2. 将同一新闻中属于同一属性的评估值相加，排序。

3. 取所得值最大的对应的属性，记为该新闻对应的属性



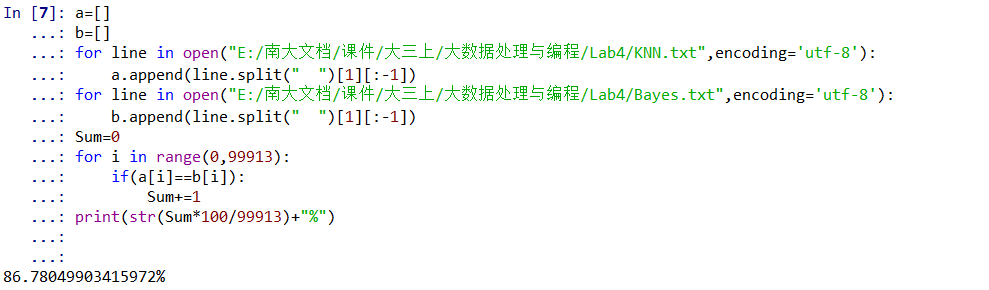
**结果如下：**



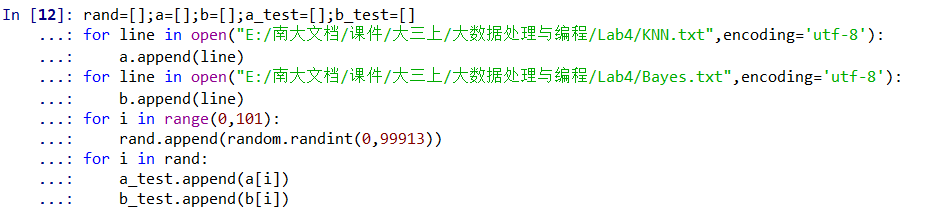
**同样有较好的预测结果**

**结果评估：**

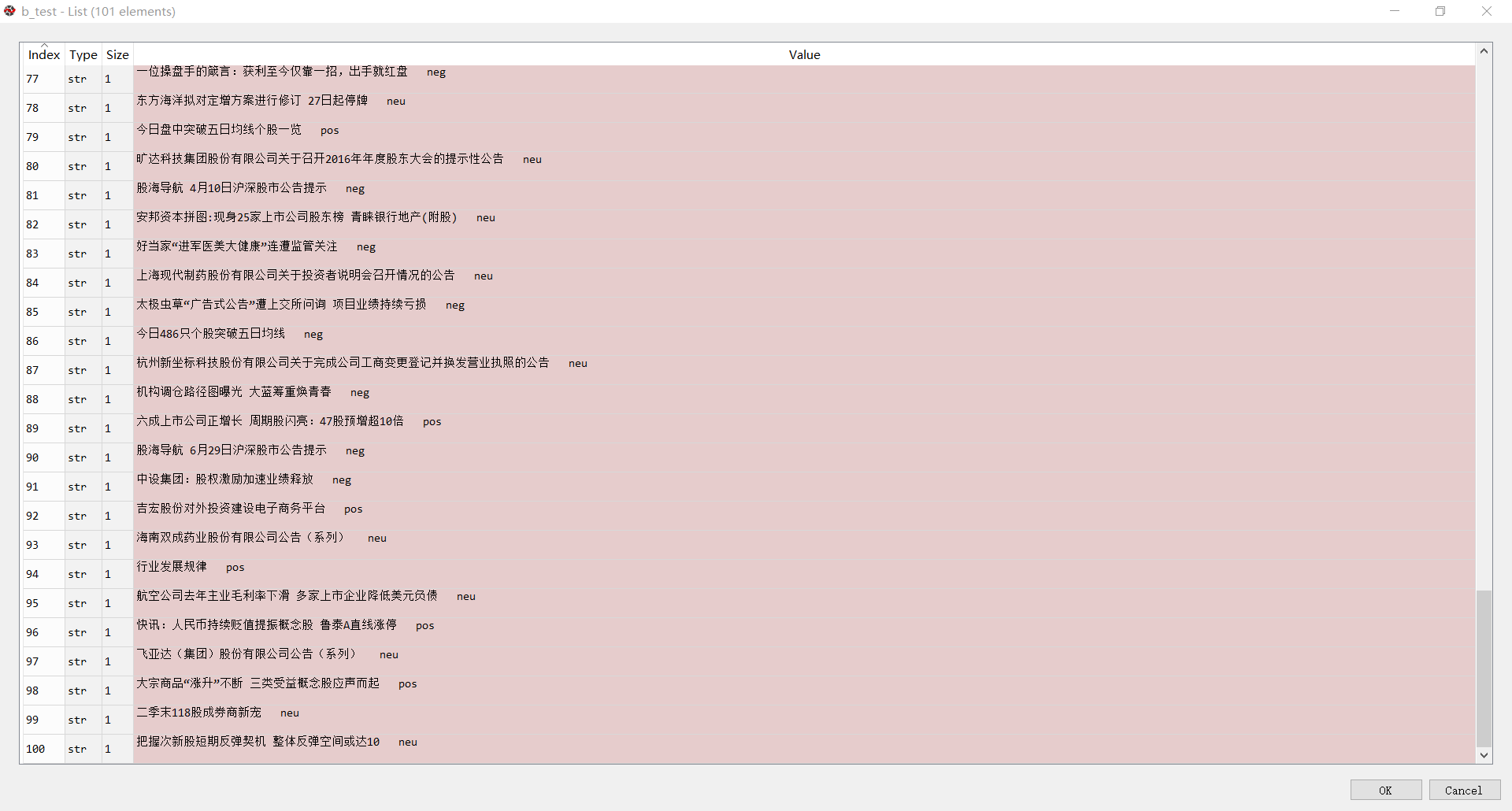
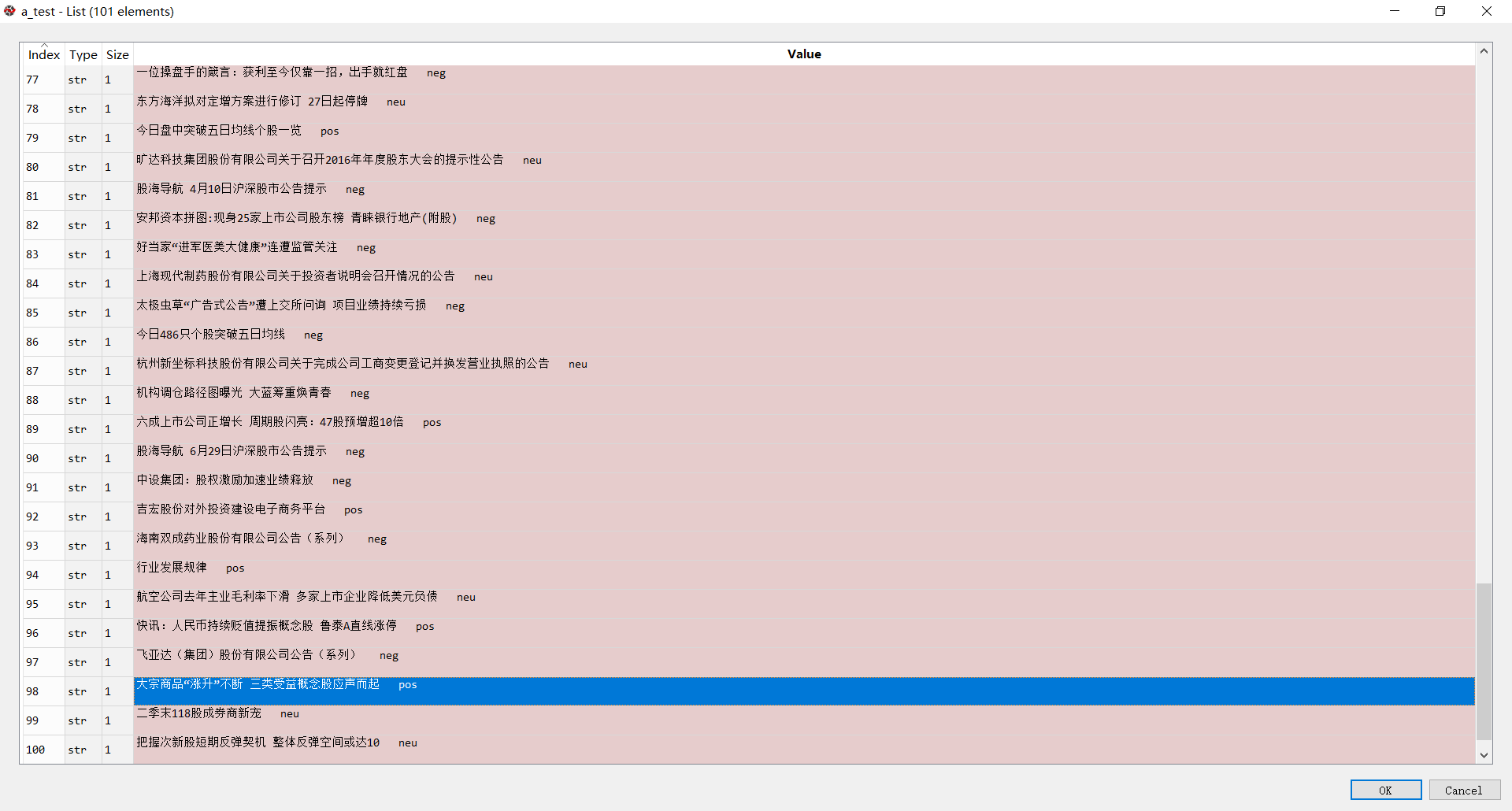
KNN和Bayes得到的结果相似度为86.7%



随机取100个数据检测：



随机检查如下：



**实验总结：**

1. 实验中对特征向量的选取较随意，可以筛选特征词，删除无用的词和在各属性集中均出现较多的词；

2. 训练样本较少，特征不明显

3. Python处理数据远比mapreduce慢，尤其是在进行KNN算法时。