Text 模块大作业

李梓童 2017202121

〇、提交文件说明

1. TEXT_2017202121.ipynb ： 本次实验中用的代码，经整理。
2. 不带标签短信_seg.txt ： 经分词处理后的不带标签短信。
3. 带标签短信_seg.txt ： 经分词处理后的带标签短信。
4. 带标签短信_seg2.txt ： 用于评估分类结果的 demo 训练数据。
5. 带标签短信_seg3.txt ： 用于评估分类结果的 demo 测试数据。
6. 短信分类结果.txt ： 经分类后的不带标签短信，格式与带标签短信相同。
7. 实验报告.doc ： 本次实验实验报告。

一、分词

通过 jieba 分词，首先打开待分词文件，进行分词后写入新的文本文件，得到“带标签短信_seg.txt”和“不带标签短信_seg.txt”。

代码：对照提交代码（TEXT_2017202121.ipynb）中第一部分代码块。

二、文本分类

通过 nltk，用朴素贝叶斯进行分类。选取部分带标签短信进行精度测试。

提交代码说明：

1. 提交代码中的第二部分代码块实现了文本分类 demo 和测试结果评估 demo。其中带标签短信_seg2.txt（约 5000 条短信） 为训练数据，带标签短信_seg3.txt（约 280 条短信）为测试数据。将分类后的数据与实际标签比较，得到评估结果如下：

   ```
   Precision = 0.6327, Recall = 1.0000, F-Score = 0.7750, Accuracy = 0.9268
   ```

   其中 Precision、recall、F-Score、Accuracy 与课程 ppt 中含义一致。Precision=正确预测的正样本数/预测为正例的样本数，recall=预测正确的正样本数/标注的正样本数，F-Score 为 precision 和 recall 的调和平均数，accuracy=正确分类的样本数/总样本数。
2. 提交代码中的第三部分代码块实现全部未标签短信（共 200000 条）文本分类。首先使用 800000 条已标签短信训练，除去停止词后提取特征向量，后用朴素贝叶斯分类器进行分类，输出结果保存在 短信分类结果.txt 中。

三、文本检索

本次文本检索搜索引擎使用 elasticsearch 与其可视化工具 kibana 实现。具体步骤如下：

1. 配置好 kibana、elasticsearch、ik 分词器以及 java 运行环境。
2. 将得到的 短信分类结果.txt 和 带标签短信.txt 转化成 json 格式的文件，便于导入到 kibana 数据库中。由于生成的 json 文件较大（约有 120MB），故不随实验结果一同提交。转化过程使用提交代码中的第四部分代码块，代码效果如下所示：

```
1       .x月xx日推出凭证式国债x年期x.xx.xx%，x年期x.xx%到期一次还本付
0       x强度等级水泥的必要性和可行性进行深入研究
0       Don’tSellaProduct
0       以上比赛规则由江苏科技大学教职工摄影协会负责解释
0       坐12个小时飞机身体已经疲惫不堪
0       为什么不能是你③以多数人的努力程度
0       地址位于天津市滨海新区响罗湾旷世国际大厦A座1801室
```

↓

```
{"index":{"_index":"text","_id":1}}
{"text_entry":" .x月xx日推出凭证式国债x年期x.xx.xx%，x年期x.xx%到期一次
{"index":{"_index":"text","_id":2}}
{"text_entry":"x强度等级水泥的必要性和可行性进行深入研究", "value":"0"
{"index":{"_index":"text","_id":3}}
{"text_entry":"Don’tSellaProduct", "value":"0"}
{"index":{"_index":"text","_id":4}}
{"text_entry":"以上比赛规则由江苏科技大学教职工摄影协会负责解释", "val
{"index":{"_index":"text","_id":5}}
```

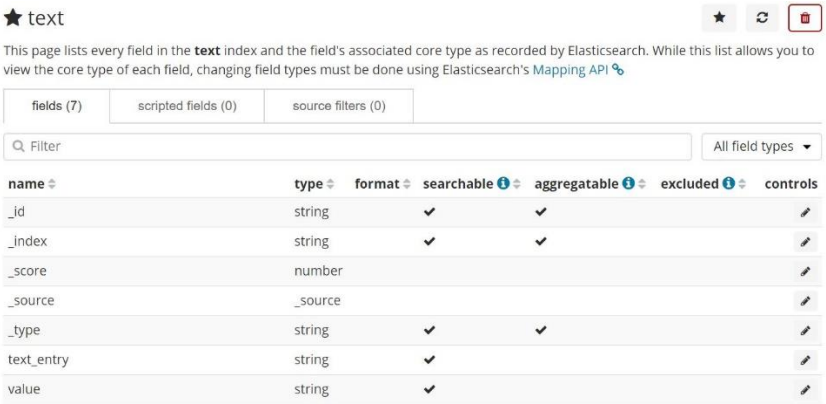
3. 在 kibana 中建立导入文件时需要的映射：

```
PUT /text
{
 "mappings": {
  "doc": {
   "properties": {
    "text_entry": {
      "type": "text"
    },
    "value": {"type": "text"}
   }
  }
 }
}
```

4. 将短信分类结果.txt 和 带标签短信.txt 转化成 json 格式的文件后，在 cmd 命令框中使用 curl 指令将得到的 json 文件（注：out3.json 为带标签短信.txt 转出的 json 文件）：

```
F:\doc>curl -H "Content-Type: application/json" -XPOST "localhost:9200/text/doc/_bulk?pretty" --data-binary "@out3.json
"
```

5. 在 kibana 中建立相应的搜索引擎。下图为搜索条件设置，text 是所有短信存放的索引 index，text_entry 存放短信本文，value 存放短信对应的值（0 或 1）：



6. 结果展示：

（单关键词）

（多关键词，中间用“AND”连接）

7. * 过程中遇到的 bug:

   ① 没有注意映射和 json 文件中键值对的一一对应关系。

   ② bulk API 的接口要求输入的 json 文件必须以换行符结尾，且要求一行基础信息、一行内容，交替输入。

   ③ 生成 json 文件时可以选择生成一个较长字符串后一次性写入，减少读写文件的次数来提高运行效率。

四、结果分析

1. 检查分类器结果的时候发现，含有敏感词的、长度较短的正常短信更有可能被判断为垃圾短信。原因是短信中所有单词数偏少，分母偏小，对概率计算结果产生影响。
2. 我们并不能看到朴素贝叶斯分类器的内部执行流程，训练数据越大也不一定保证评估结果越准确，从这个意义上来说，机器学习的不准确性和黑箱体质得到体现。
3. 通过本次实验操作，我对文本处理（分词、分类和检索）有了更深的理解。实验中使用较多已有的工具（如 jieba、nltk 等），降低了实验难度。中文文本的处理过程较之英文更加复杂，原因在于其本身断句的不确定性和语义在语素单位中分配的不均衡。