

# 项目报告书

@windcode

<https://github.com/windcode/EventExtractByNovel>

系统名称.....	2
系统功能.....	2
基本思路.....	3
实现方法.....	3
前言.....	3
训练文本预处理.....	4
特征项抽取.....	4
归一化处理.....	5
构造分类器.....	5
新文本预处理.....	6
新文本分类.....	6
系统总结.....	7
未来工作.....	7

# 系统名称

《基于小说文本的事件类型识别》

# 系统功能

系统的输入为一部小说的文本，对小说文本中的每句话进行识别和分类，识别出包含主题词的句子，并输出该句子的事件类别。



## 对话事件

## 心理活动事件

## 修炼事件

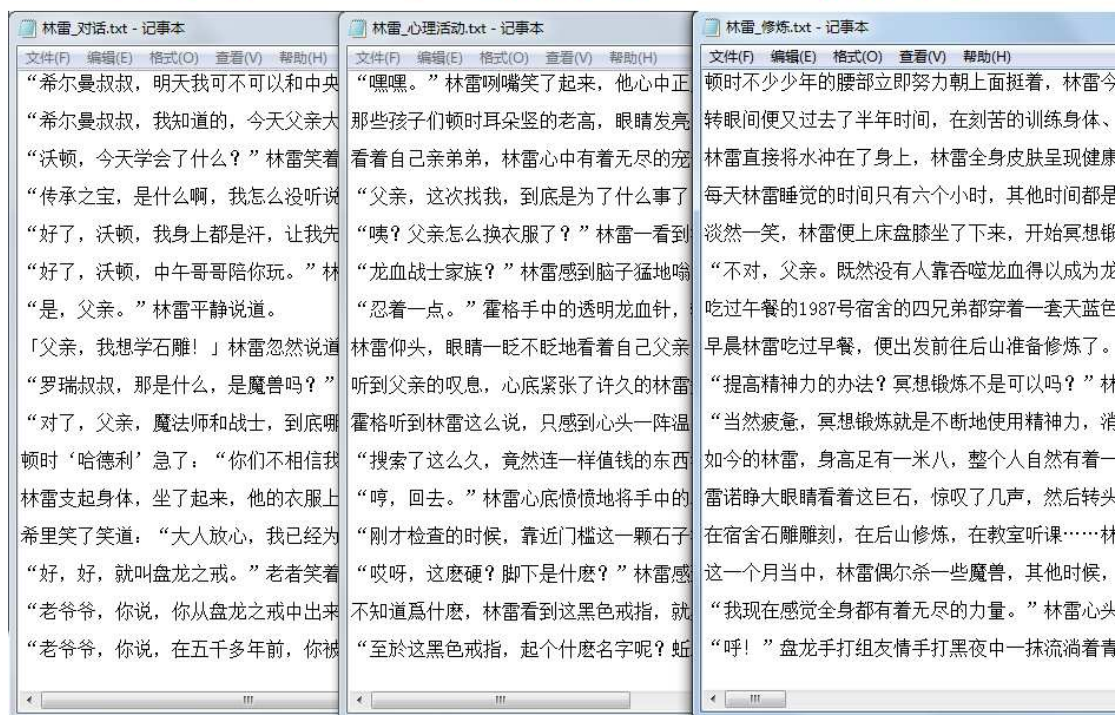


图 1 系统功能效果图

## 基本思路

1. 读取文本中每句话，找到包含主题词的句子（主题词必须作为名词并且后两个词至少一个为动词）；
2. 计算该句子的词频作为特征向量，如果全为 0，则抛弃；
3. 手动标注该特征向量的标签，即该句子属于哪个事件类别；
4. 将所有标注过的特征向量和标签保存在本地；
5. 读取保存在本地的特征向量，进行归一化处理；
6. 使用 SVM 训练归一化处理过的特征向量，生成分类器；
7. 读取文本中每句话，找到包含主题词的句子，对该句子用分类器预测；
8. 循环输出该句子的预测结果；

## 实现方法

## 前言

系统整体架构图如下：

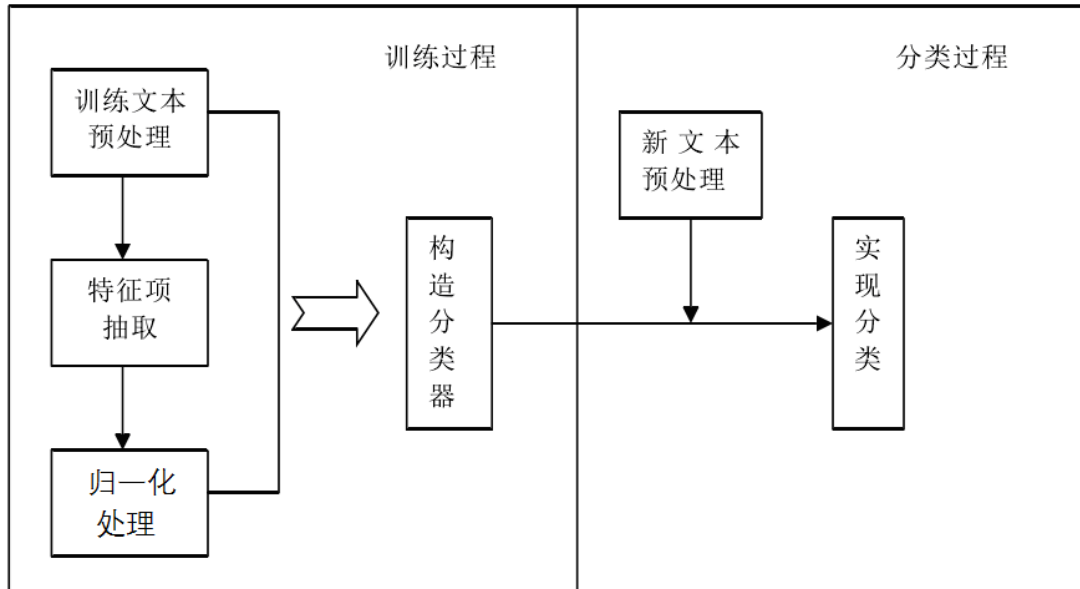


图2 系统整体框架图

整个过程分为两步：训练过程和分类过程。在训练过程中，会对文本进行预处理、提取特征向量、归一化处理以及构造分类器。分类器构造好之后，便可以进行分类，在分类过程中，首先对新文本进行预处理，然后输入到分类器中，分类器会输出预测的分类结果。

## 训练文本预处理

系统会提前设置好“主题词”，以《盘龙》这部小说为例，系统中设置的主题词为小说的主角“林雷”。在预处理步骤中，首先读取《盘龙》中的每一句话，然后对每句话进行预处理，只有符合条件的句子才能进一步处理，在这里筛选的条件是：

1. 这句话中包含主题词
  2. 主题词词性必须为名词
  3. 主题词后的两个词的词性必须至少有一个是动词
- 只有符合以上条件的句子才能进入下一步继续处理。

## 特征项抽取

在本系统中采用词频作为特征向量。首先系统设置有特征向量 tag，比如“锻炼”、“感到”、“心中”；然后统计每个 tag 在这句话中出现的频次，构成一个向量。再然后，手动标注这个特征向量为某一个标签，这个标签对应一个事件类别；最后将提取的特征向量以及该向量对应的标签保存在本地。

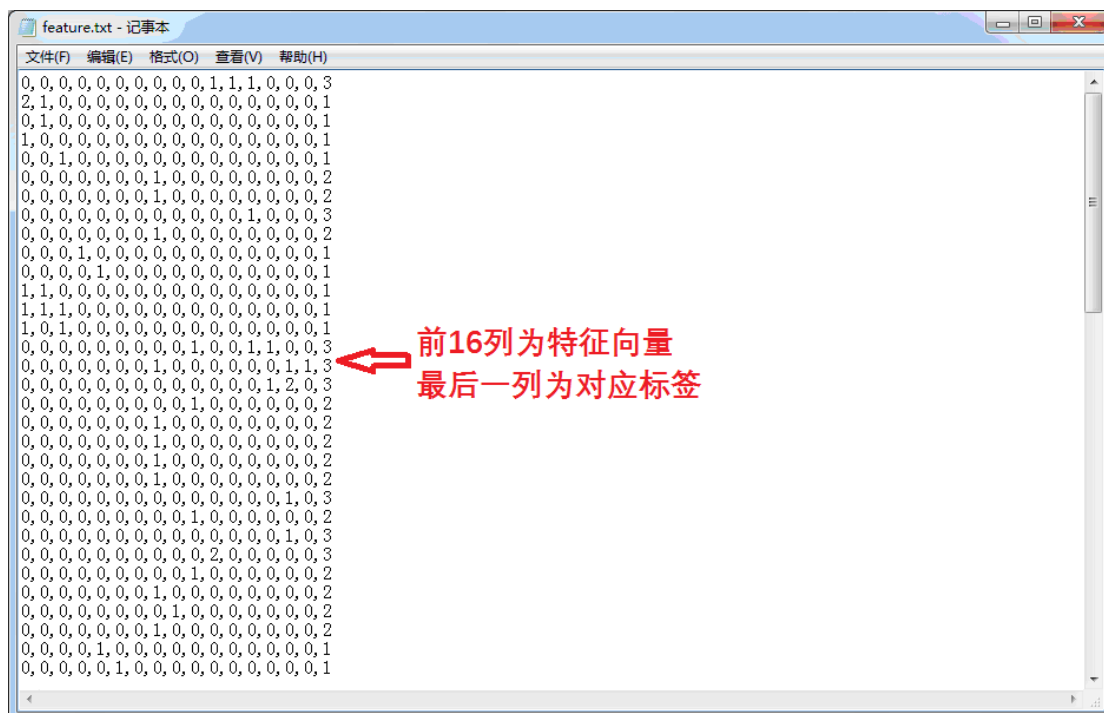


图 3 特征项抽取

## 归一化处理

这一步主要对特征向量进行归一化处理。首先读取保存在本地的特征向量；然后统计全文中出现的最大词频和最小词频；最后，根据如下公式，归一化处理每个特征向量。

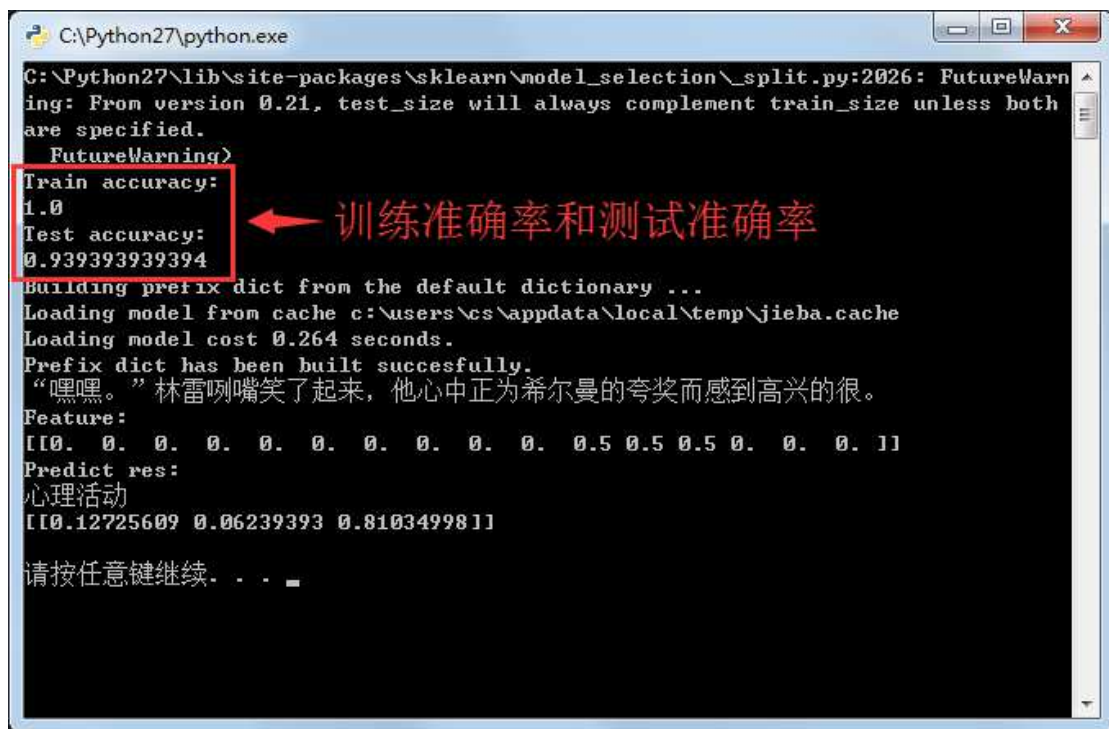
$$\frac{a - \min}{\max - \min} = b$$

图 4 归一化公式

归一化就是要把需要处理的数据经过处理后（通过某种算法）限制在你需要的一定范围内。上图公式中  $a$  为关键词的词频， $\min$  为该词在所有文本中的最小词频， $\max$  为该词在所有文本中的最大词频。这一步就是归一化，当用词频进行比较时，容易发生较大的偏差，归一化能使文本分类更加精确。

## 构造分类器

本系统采用 SVM 训练分类器。首先，读取归一化处理过的特征向量，及其对应的标签；然后，将读取的数据分为训练集和测试集，划分比例为 3:2；最后，用 SVM 训练分类器，并得出训练准确率和测试准确率。



```
C:\Python27\python.exe
C:\Python27\lib\site-packages\sklearn\model_selection\_split.py:2026: FutureWarn
ing: From version 0.21, test_size will always complement train_size unless both
are specified.
  FutureWarning)
Train accuracy:
1.0
Test accuracy:
0.939393939394
Building prefix dict from the default dictionary ...
Loading model from cache c:\users\cs\appdata\local\temp\jieba.cache
Loading model cost 0.264 seconds.
Prefix dict has been built successfully.
“嘿嘿。”林雷咧嘴笑了起来，他心中正为希尔曼的夸奖而感到高兴的很。
Feature:
[[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.5 0.5 0.5 0. 0. 0. 1]
Predict res:
心理活动
[[0.12725609 0.06239393 0.81034998]]

请按任意键继续. . .
```

图 5 构造分类器

## 新文本预处理

这一步在分类器训练完成后进行，重新读取《盘龙》的小说文本，对每句话进行筛选，筛选条件和“训练文本预处理”中相同。只有通过筛选条件的句子才能进入下一步进行分类。

## 新文本分类

首先，统计通过预处理的句子的词频，作为特征向量；然后对特征向量进行归一化处理；最后将处理过的特征向量输入到分类器中，分类器会输出预测的分类结果，并将分类结果保存在本地对应的文件中。

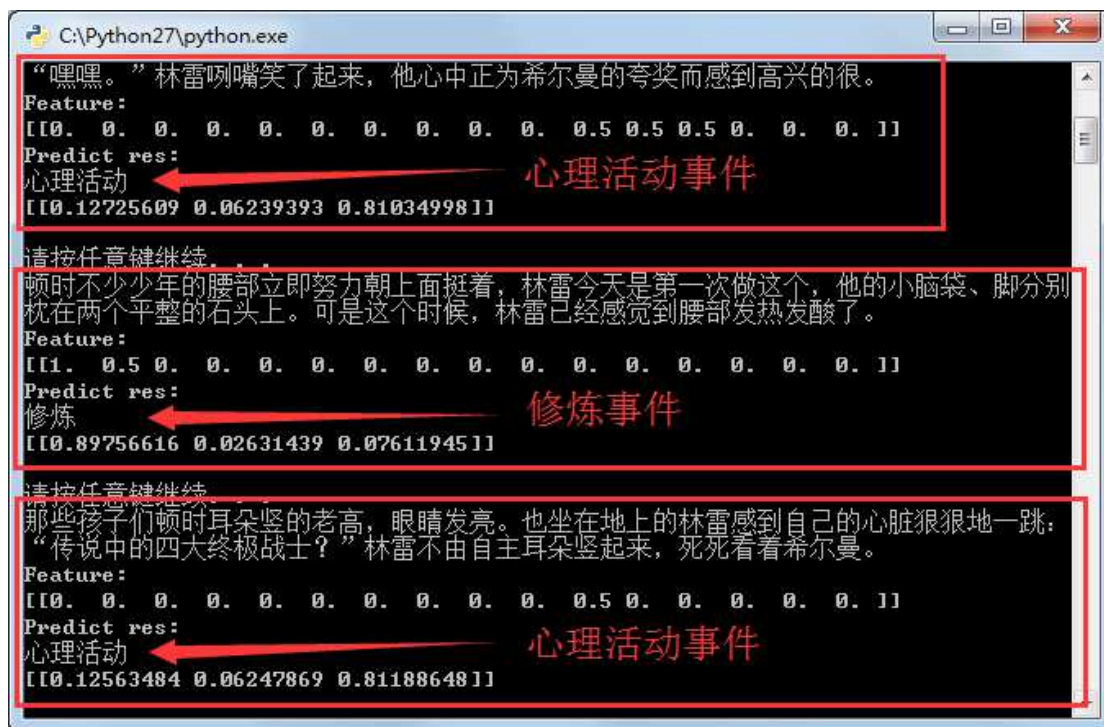


图 6 分类效果图

## 系统总结

本项目对小说文本进行分析，并对其中特定事件类别进行识别，实现方法主要是以归一化的词频作为特征向量，采用支持向量机（SVM）训练分类器，然后对文本中事件类型进行批量识别。

## 未来工作

目前的系统有几个缺点，比如训练集需要手动标注，事件类别较少等。下一步将深入学习一些 NLP 的常用方法，或者采用一些深度学习的方法，改进系统的准确性和易用性。