项目报告书

@windcode

https://github.com/windcode/EventExtractByNovel

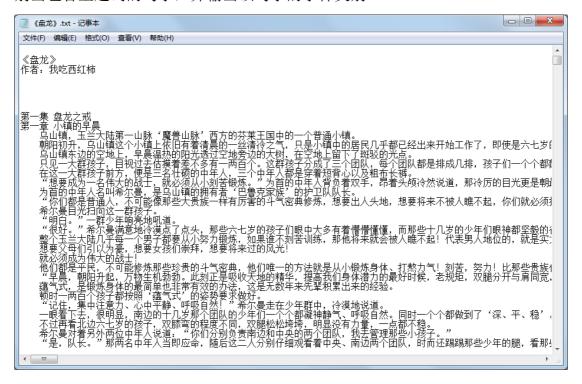
系统名杯	2
系统功能	
基本思路	
实现方法	
前言	3
训练文本预处理	4
特征项抽取	4
归一化处理	5
构造分类器	5
新文本预处理	6
新文本分类	6
系统总结	7
未来工作	7

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

→ 林雷 修炼.txt - 记事本 林雷 对话.txt - 记事本 ◎ 林雷 心理活动.txt - 记事本 文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H) 文件(F) 編輯(E) 格式(O) 查看(V) 帮助(H) 文件(F) 編輯(E) 格式(O) 查看(V) 帮助(H) "希尔曼叔叔,明天我可不可以和中央""嘿嘿。"林雷咧嘴笑了起来,他心中正 顿时不少少年的腰部立即努力朝上面挺着,林雷今 "希尔曼叔叔,我知道的,今天父亲大 那些孩子们顿时耳朵竖的老高,眼睛发亮 转眼间便又过去了半年时间,在刻苦的训练身体、 "沃顿,今天学会了什么?"林雷笑着||看着自己亲弟弟,林雷心中有着无尽的宠||林雷直接将水冲在了身上,林雷全身皮肤呈现健康 "传承之宝,是什么啊,我怎么没听说!"父亲,这次找我,到底是为了什么事了| 每天林雷睡觉的时间只有六个小时,其他时间都是 "好了,沃顿,我身上都是汗,让我先」"咦?父亲怎么换衣服了?"林雷一看到 淡然一笑,林雷便上床盘膝坐了下来,开始冥想够 "好了,沃顿,中午哥哥陪你玩。"林:"龙血战士家族?"林雷感到脑子猛地嗡!"不对,父亲。既然没有人靠吞噬龙血得以成为龙 "忍着一点。"霍格手中的透明龙血针,吃过午餐的1987号宿舍的四兄弟都穿着一套天蓝色 "是,父亲。"林雷平静说道。 「父亲,我想学石雕!」林雷忽然说道 林雷仰头,眼睛一眨不眨地看着自己父亲 早晨林雷吃过早餐,便出发前往后山准备修炼了。 "罗瑞叔叔,那是什么,是魔兽吗?" 听到父亲的叹息,心底紧张了许久的林雷 "提高精神力的办法?冥想锻炼不是可以吗?"林 "对了,父亲,魔法师和战士,到底哪」霍格听到林雷这么说,只感到心头一阵温」"当然疲惫,冥想锻炼就是不断地使用精神力,消 顿时'哈德利'急了,"你们不相信我<mark>"</mark>搜索了这么久,竟然连一样值钱的东西<mark></mark>如今的林雷,身高足有一米八,整个人自然有着一 林雷支起身体,坐了起来,他的衣服上 "哼,回去。"林雷心底愤愤地将手中的 雷诺睁大眼睛看着这巨石,惊叹了几声,然后转头 希里笑了笑道:"大人放心,我已经为<mark>"刚才检查的时候,靠近门槛这一颗石子</mark>在宿舍石雕雕刻,在后山修炼,在教室听课……材 "好,好,就叫盘龙之戒。"老者笑着□"哎呀,这麽硬?脚下是什麽?"林雷感□这一个月当中,林雷偶尔杀一些魔兽,其他时候, "老爷爷,你说,你从盘龙之戒中出来,不知道爲什麽,林雷看到这黑色戒指,就!"我现在感觉全身都有着无尽的力量。"林雷心头 "老爷爷,你说,在五千多年前,你被 "至於这黑色戒指,起个什麽名字呢?蚯 "呼!"盘龙手打组友情手打黑夜中一抹流淌着青 4 III

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:

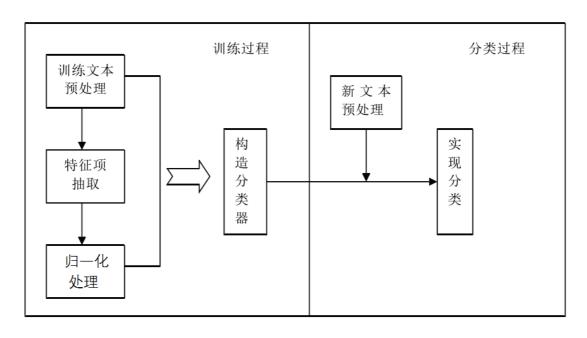


图 2 系统整体框架图

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。

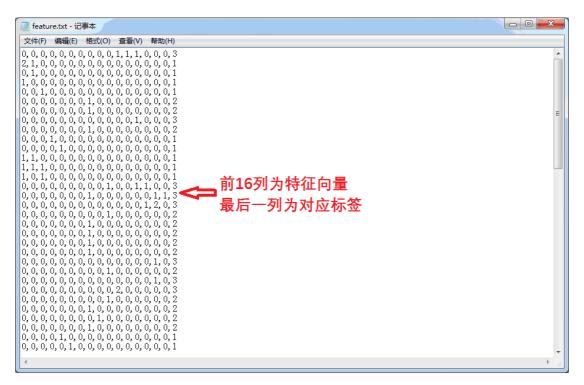


图 3 特征项抽取

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
G: \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
                    训练准确率和测试准确率
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 succesfully.
"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:
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。

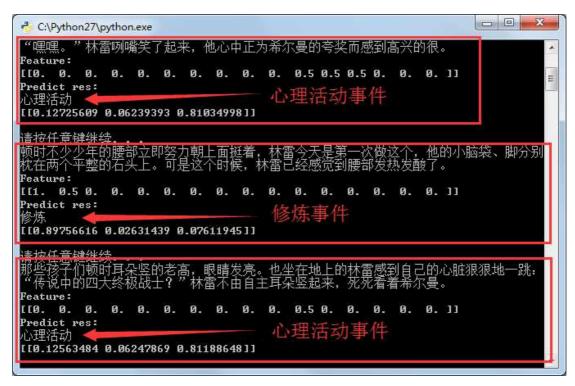


图 6 分类效果图

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。