黑客派

色情文章检测

作者：yudake

原文链接：https://hacpai.com/article/1519289708980

来源网站：黑客派

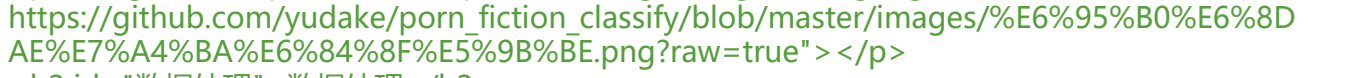
<pre><code class="highlight-chroma">Author:yudake
date:2018/2/22
</code></pre>

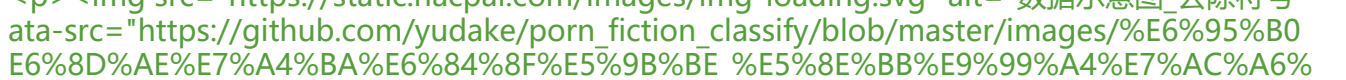
GitHub 地址：https://github.co/yudake/porn_fiction_classify
 转载请注明出处
<h2 id="背景">背景</h2>
信息爆炸是当今互联网的一大特征，人们可以在互联网上搜索到各种各样的信息。互联网因为其播的便利性，也成了很多作者发表文章的地方。现如今很多知名的作者都是通过发表网络小说而依据名，包括我最喜欢的网络小说作家之一——天下霸唱。同时很多经典的名著也被制作成了电子书，方人们阅读。
但是，有些作者发表的文章充斥着色情与暴力，一旦被青少年看到，会产生难以想象的后果。我需要对网络上的文章进行检测，以标记出其是否为色情文章，如果是，那么我们就要将其过滤不予显。而人工检测在信息爆炸的今天几乎不可能实现。所以，我们提出了基于神经网络的色情文章检测。
具体的模型工作原理可以参考我翻译的一个关于 NLP 中 CNN 研究的博文 。
<h2 id="数据">数据</h2>
我们的数据保存在 CSV 文件中，共两列，分别是：

label:

1:色情文章；
0:非色情文章；

fiction:数据。

其中色情文章 11999 篇，非色情文章 11749 篇，共 23748 篇文章。


<h2 id="数据处理">数据处理</h2>
因为色情文章中往往会有较多的标点符号，而且网络小说中也会有各种乱码存在，对我们提取特造成困难。所以，我们先把文章中的特殊字符与标点符号去掉。


然后，我们考虑到一篇小说可能会有上万字甚至更多。而文本卷积神经网络要求所有输入数据有个统一的长度，如此长的数据对计算压力要求太高。而且如此长的数据对于模型来说没有很大意义。
考虑到一篇文章内，往往中间部分的内容比较能够代表本篇文章的主旨，所以我们数据的选取方如下所示：

如果文章大于 3000 个词，则选取中间的 1500 个词；
如果文章小于 3000 个词，但是大于 1500 个词，则选取最后 1500 个词；
如果文章不大于 1500 个词，则利用特殊符号补全到 1500 个词。

我们的数据是中文数据，不像英文单词可以利用空格进行区分单词。我们使用的 jieba/strong> 库对文章进行分词。
将选取好的数据转换成数字后，文章数据如图所示：

<h2 id="模型">模型</h2>

其中 <code>conv1</code> 的卷积核大小为 2，也就是对嵌入矩阵的相邻两行进行卷积计算，<code>conv2</code> 的卷积核大小为 3，<code>conv3</code> 的卷积核大小为 4，<code>conv4</code> 的卷积核大小为 5。每个卷积的维度为 2，也就是有两个大小相同的卷积核进行卷积。过卷积之后生成了两个 <code>1499*1</code> 向量，两个 <code>1498*1</code> 向量，两个 <code>1497*1</code> 向量，两个 <code>1496*1</code> 向量。
在池化层对 8 个向量进行最大池化，分别从每个卷积提取出 1 个特征值。将 8 个特征值拼接成个 <code>8*1</code> 维向量，至此，我们就把文章中的特征提取出来了。
最后，我们将提取出来的特征送入 softmax 层进行分类，获得最终结果。
神经网络具体工作流程见翻译的博文。
<h2 id="训练">训练</h2>
<h3 id="参数设置">参数设置</h3>

batch_size = 16
循环次数 = 3
学习率 = 0.005
嵌入矩阵维度 = 32

<h3 id="交叉验证机与测试集选取">交叉验证机与测试集选取</h3>
我们选取 2000 条数据作为测试集，剩余数据作为训练集。
然后在每次训练循环中随机抽取剩余数据的 10% 作为交叉验证集。
<h3 id="Accurate变化">Accurate 变化</h3>

可以看出，在训练稳定之后，训练集上的准确率保持在 90% 以上，平均准确率在 98% 以上。
在交叉验证集和测试集的平均准确率也在 98% 以上。