IMDB评论情感判断

数据特征分析

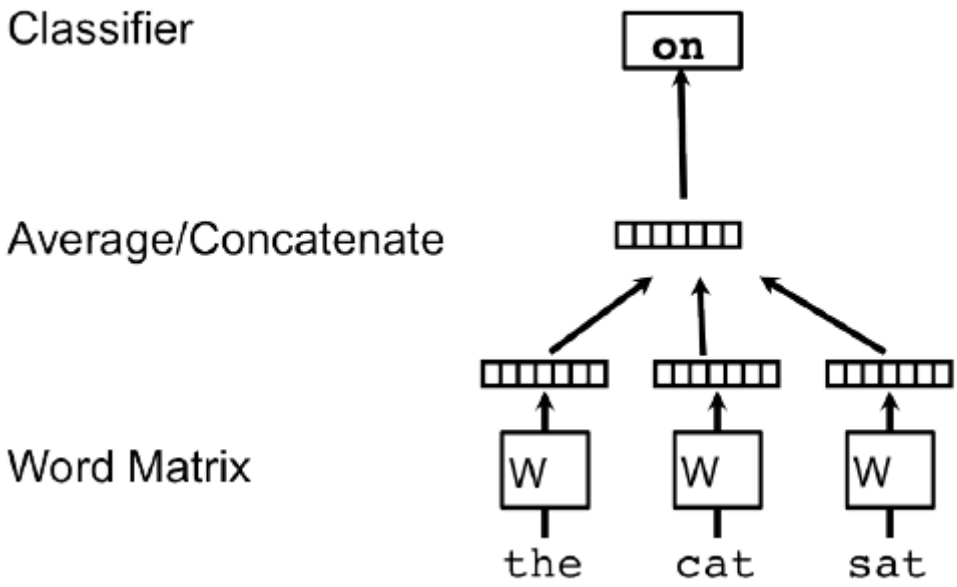
该数据集只有review和sentiment两个维度，其中，review为一段自然语言，且其中有标点符号和br等符号；sentiment只有positive和negative两个值。由此，我们需要先对评论进行清洗，然后转化为便于分类的表示形式，最后再进行分类。

文档向量生成：Doc2Vec

方法介绍

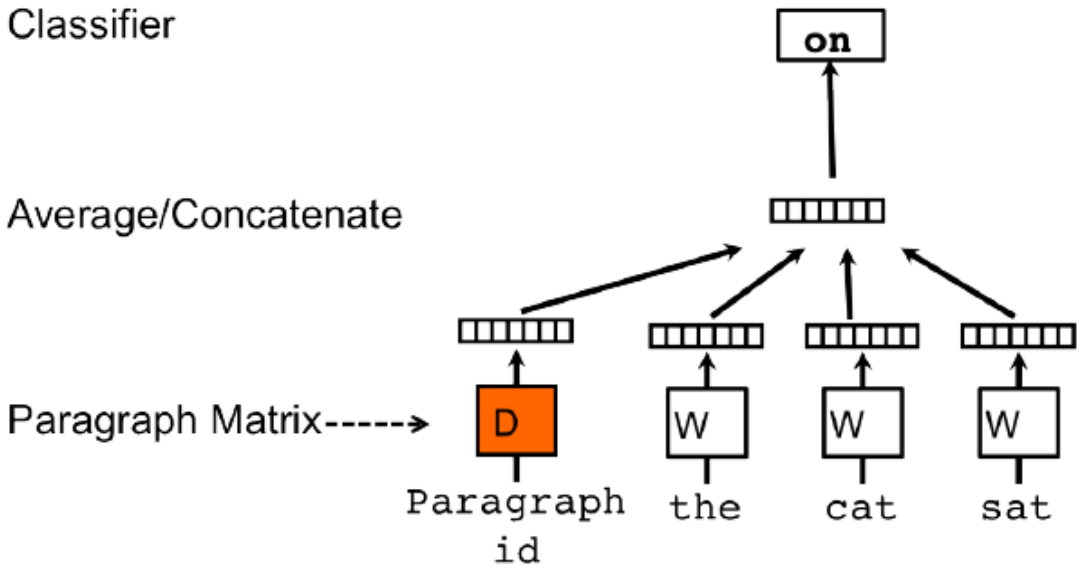
Doc2Vec模型是用于将一段文档转换成一个向量的工具。文档的相似度越大，其向量之间也就越接近。通过对评论应用Doc2Vec，可以将其转换为便于分类器分类的形式。

Doc2Vec模型是基于Word2Vec模型提出的。Word2Vec是一个将单词向量化的工具，单词的相似度越大，其向量之间也就越接近。以其中的连续词袋CBOW模型为例，其可以利用周围的词来预测这个词本身。



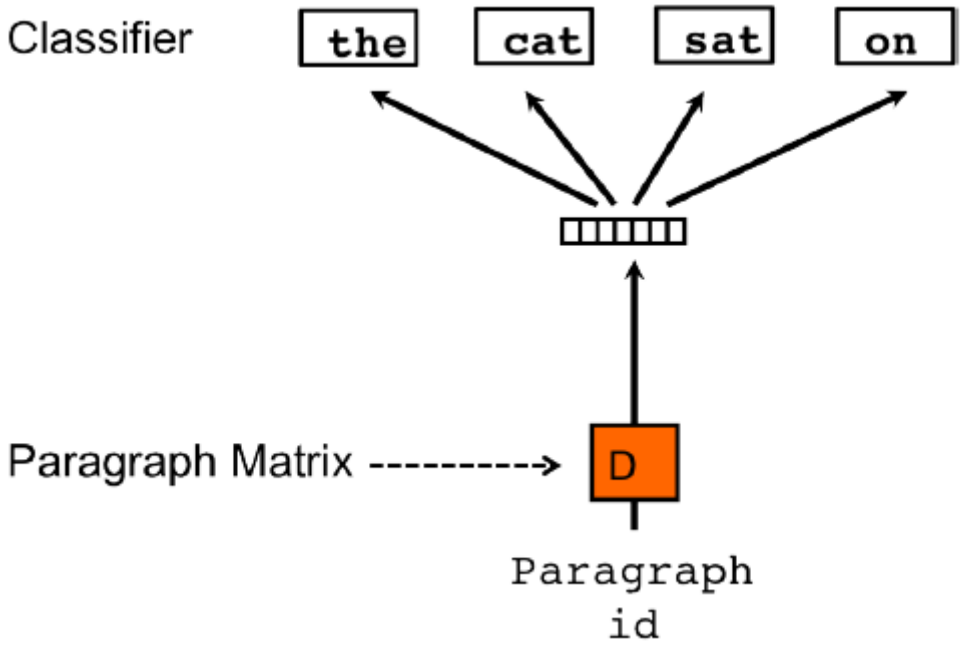
Doc2Vec模型在Word2Vec模型的基础上，进一步考虑了单词间排列顺序的影响。

Doc2Vec模型中的Distributed Memory（DM）模型类似上面的CBOW模型，但是，除了使用单词来预测下一个单词之外，还添加了另一个特征向量，它对于每个文档是唯一的。



当训练单词向量W时，也训练了文档向量D，在训练结束时，它就有了文档的向量表示。

Doc2Vec模型中的Distributed Bag of Words（DBOW）模型是使用文档的特征向量来预测其中一组随机的单词。



这个算法实际上更快，并且消耗内存更少，因为不需要保存词向量。

模型实际上是一个单隐层神经网络。训练过程采用SGD进行，梯度使用BP算法获得。

方法实现

采用Gensim中的Doc2Vec模型。

首先对训练集和测试集中的所有评论进行清洗：将评论全部转为小写并去除br符号。然后对评论进行分词，此时将标点符号作为单独的单词。接着分别建立DM和DBOW模型，并用处理后的评论建立其词汇表后进行训练（每轮训练前打乱评论的顺序）。最后将DM和DBOW预测的向量进行拼接，即可得到所有评论对应的向量。

调参过程

由于该模型参数较多，故参数通过查阅类似的项目直接给定，并未进行调参。其中，min_count设为2可以只统计出现次数大于等于2次的单词，可以显著提高模型性能。

遇到的问题及解决方法

在安装Gensim的过程中，发现其与之前在另一个环境中安装的pytorch有冲突。在一致的python版本下重新安装两者的最新版即可解决。

情感分类：LinearSVC

方法介绍

该方法采用软间隔的线性支持向量机，通过优化问题

$$\begin{aligned} \min_{\mathbf{w},b} \ & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum \xi_i \\ subject\ to \quad & t_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0 \end{aligned}$$

来实现分类。其中，C为超参数，代表惩罚代价（正则化系数）。

由于样本数量（40000）大于样本维数（200），故可以直接对这个二次规划问题进行求解。

对于新样本，可以使用

$$sgn(\mathbf{w}^T\mathbf{x} + b)$$

进行分类。

方法实现

采用SkLearn自带的LinearSVC模型进行训练。

调参过程

惩罚系数C反映了算法对分类错误的样本的惩罚程度，其值影响模型的复杂性和稳定性。C越小，对分类错误的惩罚就小，训练误差变大；C越大，学习精度相应提高，但模型的泛化性能可能变差。另外，C的值影响到对样本中“离群点”（噪声影响下非正常数据点）的处理，选取合适的C就能在一定程度上抗干扰，从而保证模型的稳定性。

由此可见：若能选取到合适的C值，就能得到比较精确、稳定的分类模型。

使用SkLearn自带的GridSearchCV中的5折交叉验证法对C进行调整，并选择交叉验证过程中最优的C。采用4路并行计算。最后得到的最优值为$C = 0.055$。此时交叉验证的精度为0.897。

结果分析

提交精度为0.8967，排名不算靠前。由此，使用其他方法可以进一步提高预测精度。

评测排名

12名（截至2020年12月22日17:29分）。

#	用户名	登录	上次登录日期	R-square ▲
1	18377290_孙亦琦	6	11/29/20	0.8686 (1)
2	18373636_田昶尧	14	12/10/20	0.8669 (2)
3	18373202_刘勇	8	12/07/20	0.8666 (3)
4	18376161李明昕	2	12/09/20	0.8644 (4)
5	18231047_王肇凯	10	12/07/20	0.8637 (5)
6	18231217_吴陶然	24	11/28/20	0.8631 (6)
7	18373805_杨再飞	8	12/11/20	0.8619 (7)
8	18231174_任杰瑞	23	12/22/20	0.8618 (8)
9	18373054_陆晓东	12	12/10/20	0.8611 (9)
10	18373384-欧卓健	3	12/08/20	0.8600 (10)
11	陈天异	2	11/25/20	0.8575 (11)
12	18373109_孔祥浩	5	12/04/20	0.8574 (12)
13	18373488_袁劭涵	3	12/16/20	0.8570 (13)

源代码

https://github.com/refkxh/BUAA_ML_2020Autumn