NLP Lab #3 - Doc2vec 模型

Part #1 实验简介

本次实验要求使用 Doc2vec 架构，利用两种预测模型（CBOW 和 DMPV）以及两种优化方法（Hierarchy-Softmax 和 Negative-Sampling）的组合，即四种学习策略，在 Large Movie Review Dataset 英文语料上进行文本向量学习，并运用文本向量对于文本的情感色彩进行分类，以评估不同学习策略的学习能力。

本次实验的重心在于测试不同训练策略的组合，以获得对于 Doc2vec 架构和训练策略的感性认知。

Part #2 技术细节

a) 文本预处理

在文本预处理中，我们需要对来源于多个文件的语料进行学习，本人采用逐个打开每个文件，将所有语料合并的方法来处理所有文本。值得注意的是，在实际实现中，本人最初使用的是每读取一段新的语料，即将其插入列表的存储方式。但是本人发现这种实现策略在运行时效率很低，于是最后本人使用了 Python 的 yield 语法，利用 generator 迭代器，大大提高了文本收集的效率。

b) Doc2vec 模型的参数选择

在本次实验中，本人选择了 Python 的 gensim.models.doc2vec.Doc2Vec 模型作为训练文本向量的模型实现。

由于本次任务中的训练任务规模较大，本人在 Doc2vec 模型参数的选择上并没有做充分的搜索。本人根据经验，给出了如下的参数配置（省略了仅与模型训练速度有关的变量）：

参数名	含义	取值
dm_concat	合并向量时是否采用向量拼接的策略	False
dm_mean	合并向量时采用取所有向量均值的方案	False
negative	随机采样时所取随机单词的数量	5
ns_exponent	随机采样时使用的词频函数中的指数值	0.75
alpha	模型训练时的初始学习率	0.025
min_alpha	模型训练时的学习率最小值	1e-4
epochs	模型训练的总迭代数	100

c) 无标记文本的处理

在本人的原始设计方案中，本人期待利用 Doc2vec 模型，获得所有文本的特征向量，包括无标记的样本。随后，利用半监督学习的方式，比如半监督 SVM，对于所有无标记样本进行学习，最后，利用所有的样本和标记训练逻辑回归模型，以获得情感分类模型。

本人设计这一套训练方案，以期半监督学习可以标记更多的样本，从而使得逻辑回归模型可以在更多的数据上得到训练，以提高其在测试文本上的分类正确率。

但是，在进行实验时，本人发现这一套训练方案所获得的分类模型的正确率相较于仅在有标记文本上学习的模型的正确率更低。因此，本人最终放弃了所有无标记样本，这意味着在训练 Doc2vec 模型时，无标记文本没有被学习。

d) 逻辑回归模型

本人使用 sklearn.linear_model.LogisticRegression 模型作为本次实验总使用到的逻辑回归模型。在参数选择中，本人没有充分搜索此模型的参数，仅将此模型的惩罚 penalty 选项选为无惩罚。

Part #3 实验结果与分析

基于四种学习策略所获得的文本情感分类模型在测试集上的分类准确率如下表所示：

分类准确率*（%）	DBOW	DMPV
HS	84.795 (0.160)	81.421 (0.084)
NS	87.479 (0.157)	82.619 (0.062)

*每个数据项均重复了 3 次实验，括号内是 3 次实验结果的方差，所有数据保留到小数点后三位数字；

结果分析

1. 从结果来看，DBOW 训练策略在文本情感分类上的表现相比于 DMPV 来说更好，但是方差略大；HS 更新策略的效果不及 NS，这些结论与本人在实验二中对于 word2vec 模型的测试所得结果有所不同。
2. 在实验中，NS 更新方法对应的训练速度比 HS 方案快。