基于预训练词向量与LSTM的古典诗歌风格判定——以唐代七言律诗为例

摘要

在古典诗歌研究领域中，分析某位作家的诗风是一个经久不衰的重要课题。而在历代诗话与各类研究论文中，评判某位诗人的诗风——如“宗唐”或“宗宋”，多凭读者自身的阅读感受，或摘出一些风格明显的句子条分缕析。本文基于以四库全书为训练语料的大规模预训练词向量，使用LSTM搭建了算法模型，提出了一种用来判定不同诗歌风格的模型。并在艳体诗识别与唐代诗风识别等问题上进行应用，取得了较好的效果。该模型对所测试的诗歌作品做出的判定结果与历代评论家给出的定论基本吻合，可与文学史相参照。

关键词

自然语言处理 深度学习 预训练 LSTM 艳体诗 唐诗 明七子 同光体

引言

由于我国诗学传统的悠久与连贯，诗人常有师法前代的行为。宋之江西诗派，取径杜甫，遥尊其为“一祖”；明之前后七子，诗法盛唐；晚清同光体，常被目为宋诗派，诗多宋调。这些诗人的作品中，明确显露出了他们对某些前人诗风的学习与继承。后世人若研究这些诗人，必然无法绕开对他们诗学脉络的探索。判定一位诗人的作品是否有前人（如学杜，学韩）或前代（如学唐，学宋）的风格，有相当重要的意义。

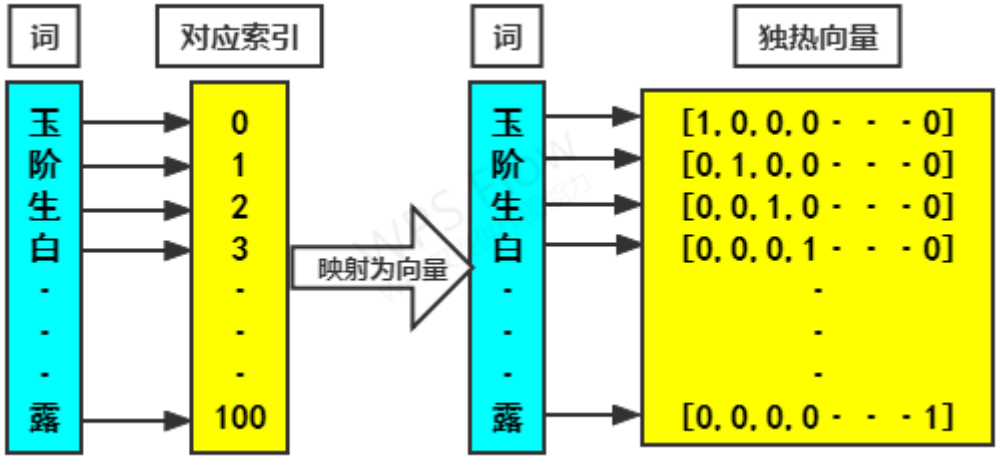
然而如何进行风格的判定？过往的方法是凭借个人的阅读体验，至多摘出一些风格明显的句子进行论证。这种依赖于主观感觉的方法有时会导致不同的结论，而且需要花费大量时间去细读。本文尝试用自然语言处理技术来解决这一问题，首先将该问题转化为一个文本二分类任务，其次将某种风格的诗歌与同等数量的随机诗歌作为训练集，并将训练集向量化。最后将向量化的训练集输入到应用LSTM算法的模型中进行二分类训练，训练集中目标风格诗歌向量映射为1，随机诗歌向量映射为0，训练达到一定精度后停止。此时经过训练的模型就可以进行诗歌风格判定任务——输入需要判定的诗歌，模型会给出一个介于0-1的结果，越接近1说明在此判定模型中该作品越有可能被分类为目标风格诗歌。

为了验证模型的效果，本文先是将此方法应用到了孙原湘《天真阁外集》中，用五百首艳体七言律诗训练出了一个识别艳体风格的模型，以王彦泓，丘逢甲等人的作品为测试集进行测试，取得了预想中的效果。接着扩大了问题规模——以所有现存唐人七言律诗为训练集，训练出了一个唐诗风格识别模型，用以判定一首七言律诗是否有唐人风味。使用该模型对明七子，同光派等代表人物进行判定，取得了与文学史评价基本吻合的结论。

预训练词向量

词向量是用于表示单词意义的向量，也可被认为是单词的特征向量或表示。 将单词映射到实向量的技术称为词嵌入。

最简单的映射方式是使用独热向量来表示词。假设训练文本中有N个不同的词，便给每个词分配一个从0到N−1的不同整数作为索引。接着创建一个长度为N的向量，并将该向量对应索引位置的元素置为1，其余都为0。这样，每个词都被表示为一个长度为N的向量，可以直接由神经网络使用。过程如下图所示。



虽然独热向量原理简明操作简单，但是独热向量不能准确表达不同词语之间的相似度，我们常用余弦相似度来度量两个向量之间的相似度，而任意两个不同词的独热向量之间的余弦相似度都为0，所以独热向量不能表示词语之间的关系与相似性。

为了解决上述问题，Mikolov等人提出了word2vec工具，它将每个词映射到一个固定长度的向量，这些向量能更好地表达不同词之间的相似性和类比关系。

本文选择使用word2vec中的跳元模型（skip-gram）[Mikolov et al., 2013b]，该算法的训练依赖于条件概率，条件概率可以被看作是使用语料库中一些词来预测另一些单词。模型参数是词表中每个词的中心词向量和上下文词向量。在训练中，我们通过最大化似然函数（即极大似然估计）来学习模型参数。这相当于最小化以下损失函数：

$$\sum_{t=1}^{T} \sum_{-m \leq j \leq m,\, j \neq 0} \log P(w^{(t+j)} \mid w^{(t)}).$$

对预定语料训练后，每个词语对应一个定长的稠密实数词向量，维度通常为50~300，每一维均由一个实数表示。

由于Shen Li等人已经在四库全书语料上使用该算法进行了训练并得到了300维的单字词向量（下称字向量），且得到了令人满意的效果，本文在该字向量的基础上进行模型搭建工作。该项目收录了19527个字对应的字向量，前十个字的字向量如下表所示。

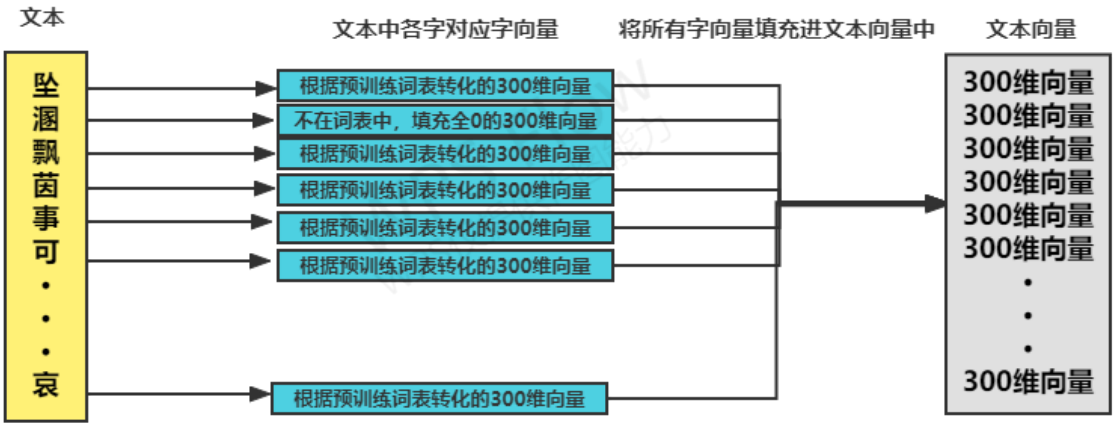
字	字向量
之	[-0.386241,-0.200756,0.106928,0.327273,0.080174 ……-0.365105]
以	[0.042976,-0.105353,0.189561,0.310238,0.402705 ……0.008979]
曰	[-0.253626,-0.486539,0.438117,-0.193879,0.134444……0.182786]
为	[-0.034353,-0.075638,0.247427,-0.560503,-0.003090……0.597949]
其	[-0.197984,-0.268704,0.211067,0.209840,0.415124……-0.235650]
而	[-0.080227,-0.418216,0.417316,0.488205,0.467254 …… -0.357677]
也	[-0.061710,-0.204168,0.370755,0.001222,-0.133613……-0.007368]
人	[-0.295339,-0.088422,-0.435675,-0.117920,-0.176023……-0.071651]
有	[-0.129041,0.190121,-0.133640,0.095273,0.375228……-0.185826]

模型搭建

文本向量化

由Salton 等提出向量空间模型(Vector Space Model, VSM)被广泛地应用于文本分类、检索和相似度计算等任务。

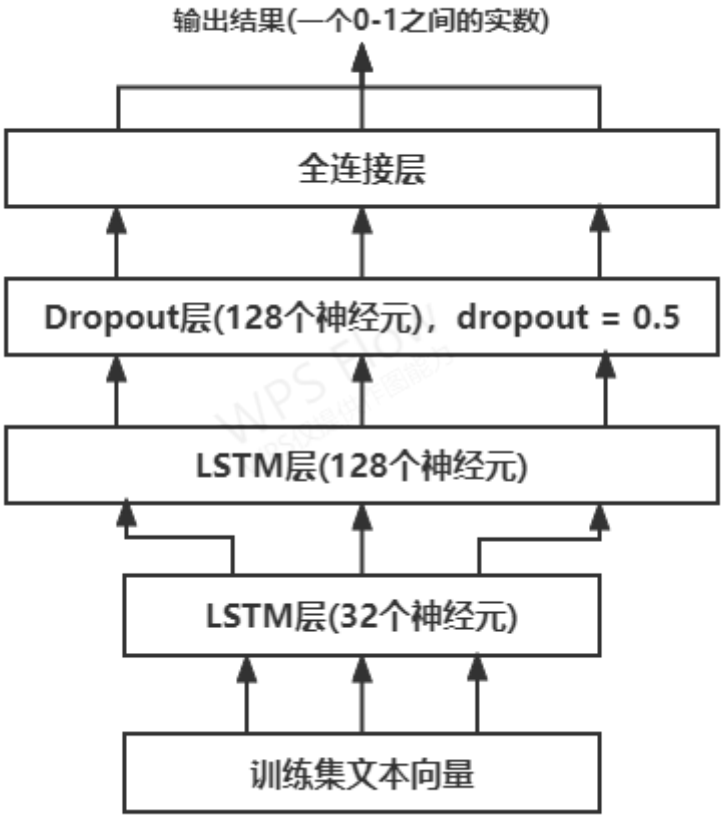
在 VSM 模型中, 一篇文本可视为由词语构成的集合,由此可以将文本转化为一个多维向量，本文在在已有预训练词向量基础上构建文本向量。具体流程如下——



经过以上过程，长度为N的诗歌文本被转化为尺寸为$N \times 300$的文本向量。

模型结构

LSTM(Long Short Term Memory)算法由Hochreiter等人提出，是一种具有记忆长短期信息的能力的神经网络，可以解决序列数据的长期依赖问题，适合处理诗歌文本这种序列数据。本文采用LSTM算法搭建模型，为了提升泛化性，避免过拟合，使用了Dropout策略，每次随机丢失一半的参数，以此提升模型的健壮性。模型结构如下——



损失函数是二值交叉熵，优化算法采用Adam，最后的全连接层激活函数为sigmoid，输出一个0-1的实数。

艳体诗识别

数据预处理

孙原湘的《天真阁外集》多为艳体诗，钱锺书先生认为，这些诗“殆庶《香奁》、《疑雨》 二集”， 又谓能“上配《疑雨》”。

《天真阁外集》作品多为七律，故而本文将七言律诗作为训练和验证的体裁，将孙原湘《天真阁外集》中的七言律诗抽取出来，数量为500篇，标注为1，又随机抽取他人的七言律诗500篇七言律诗作为对比，标注为0。将两者混合，随机抽取10%作为验证集，剩余90%作为训练集。

训练

使用本文构造的模型进行训练，训练10轮后，验证集准确率约为0.9，损失值为0.47，为避免过拟合，停止继续训练。

效果验证

为验证模型效果，抽取王彦泓和丘逢甲的全部七言律诗作品作为验证。效果如下

诗人	模型预测结果(相似度)
王彦泓	0.8653275
丘逢甲	0.1216295

王彦泓的七律作品，模型给出的预测结果的平均值约为0.86，由于孙原湘《天真阁外集》七律作品被标注为1，王彦泓0.86的平均作品与1很接近，也就是说，与孙原湘《天真阁外集》的七律作品很接近。而王彦泓本人正以艳体诗出名，模型预测结果与现有论断相符。

而丘逢甲的七律“开满劲弓，吹裂铁笛”，与艳体诗相去甚远，于是模型给出的预测结果只有0.12，这也符合其作品的风格。

唐诗风格识别

唐诗宋诗正如太极之两仪，明清诗人不宗唐则宗宋，余者寥寥。纵有王闿运等宗汉魏六朝者，也无法影响大势。由于唐诗在后世的影响巨大，本文决定训练出一个模型来判断一首诗是否与唐诗相似，或者说有唐诗的风格。由于唐诗体裁多样，歌行，五古，乐府，七绝，七律，五绝，五律，不同的体裁纵使是一个诗人所写，风格也不一样。于是本文选择了风格较为鲜明的七言律诗进行研究。

数据预处理

首先将所有搜集到的唐代七言律诗抽取出来，计7763首，标注为1。又在与唐诗风格差异较大的宋代七律中抽取相同数量的作品作为对照，标注为0。将两者混合，随机抽取10%作为验证集，剩余90%作为训练集。将全部文本数据进行向量化后输入模型。

训练

使用先前的模型进行训练，由于数据较多，训练20轮后，验证集准确率约为0.8，损失值约为0.9，为避免过拟合，停止继续训练。

验证

明七子

宋后学唐影响最大的莫过于明七子，他们提出“文必秦汉，诗必盛唐”的文学主张，在诗歌创作上对唐人进行了过度的模拟。本文选出前后七子中较有代表性的几人进行测试——

诗人	模型预测结果(相似度)
李梦阳	0.6254571
何景明	0.7542985
李攀龙	0.7922501
谢榛	0.7385051
王世贞	0.5985163

从这几位明七子的代表人物的测试结果可以看出，他们的七言律诗有很重的唐人风味，相似度平均在0.7左右。王世贞略低，但也接近0.6，根据钱锺书先生的理论：“弇州於嘉靖七子，实为冠冕；言文必西汉，言诗必盛唐。《四部稿》中，莫非实大声弘之体. 然弇州《续稿》一变矜气高腔，几乎剟言之瘢，刮法之痕，平直切至。屡和东坡诗韵……则是弇州早作已染指苏诗矣。虽词气尚负固矜高，不肯遽示相下，而乃心则已悦服。”，可以看出王世贞晚年诗染宋调，这可能是他的七律与唐诗相似度略低的原因。

江西诗派

江西诗派虽然主张学习杜甫，韩愈等人，但却是典型宋代诗风的代表。钱锺书先生指出“唐诗、宋诗，亦非仅朝代之别，乃体格性分之殊，天下有两种人，斯分两种诗。唐诗多以丰神情韵擅长，宋诗多以筋骨思理见胜……故唐之少陵、昌黎、香山、东野，实唐人之开宋调者”，根据这种划分，杜甫韩愈等人虽是唐人，作品却与大多数唐人的作品风格不同，算是宋诗诗风，而学习杜甫韩愈等人的江西诗派，作品也与唐诗风格有很大差异。我们选取了江西诗派“一祖三宗”中的“三宗”进行验证

诗人	模型预测结果(相似度)
黄庭坚	0.3392725
陈与义	0.2314215
陈师道	0.2883462

结果符合现有文学史论断，作为宋诗代表人物，黄庭坚等人的七律作品与唐诗存在较大差异。

同光诗派

同光体是近代诗派之一。主要代表人物有陈三立，郑孝胥，陈宝琛，陈衍等人。陈衍宣称此派诗人特点为“同、光以来诗人不墨守盛唐者”《沈乙庵诗序》，但根据钱仲联先生的说法，同光体“远承宋代江西派而来,以黄庭坚为宗祖”，也就是说，同光体诗人以学宋为主，由于数据不足，只选陈三立，郑孝胥，陈宝琛几位同光巨擘进行验证——

诗人	模型预测结果(相似度)
陈三立	0.33115724
郑孝胥	0.33334845
陈宝琛	0.26346013

结果与江西诗派三宗相似，由此可以看出同光体诗人学宋的主张在他们的七律创作中确实有所体现。

总结

虽然大多数诗人的诗歌作品都是出自己手，有自己的特点与原创性，但不可否认的是，我们如今对某位诗人进行研究总还要找出他的诗学脉络，然后断言这位诗人学杜韩，学苏黄，学江西派，学西昆体，乃至学唐，学宋。本文的工作意义在于寻找到了一种通用的，且消除了主观误差的方法，来判断某篇作品与一类作品是否相似。这可以作为古典诗歌研究者用于寻找诗人取径的一个方便的工具，也可以与现有的研究成果互相参照。像袁枚这种标榜自己诗学独树一帜，与古人无关，乃至扬言道“独来独往一枝藤，上下千年力不胜。若问随园诗学某，三唐两宋有谁应。”，竭力掩饰自己诗学取径的诗人，我们也可以用这个工具来分析他的诗风与谁比较相似，遗憾的是，我们没有得到袁枚诗歌的数据，故而没有做这项研究。

当然，本文提出的工具不仅可以用来判断一位诗人的诗学取径，还可以对一个诗派，甚至一个时期的诗学创作形式进行研究。比如同光体各路名家各有所本，不过都是以宋诗为纲，所以后人也提出过“同光无体”的说法。到底同光体各位诗人之间差异大不大？他们能不能算作同一个诗派？学者使用这个工具可以判断出他们之间的相似性来辅助研究。再如竟陵派承继了公安派的“性灵说”，并对其进行了一定的修正，那么竟陵派与公安派的诗歌创作差别有多大？诗风有多像？借助本文的模型，也可以一窥究竟，希望本文的工作对此类问题的研究有所裨益。