基于机器学习的古诗词情感分析

安芊桦

关键词：自然语言处理 机器学习 文本情感分析

一、 提出问题

诗词，是指以古体诗、近体诗和格律词为代表的中国古代传统诗歌。诗是高度集中地概括反映社会生活的一种文学体裁，而词则是属于诗的一种韵文形式，由五言诗、七言诗或是民间歌谣发展而成。本研究旨在建立诗词文字与思想情感方面的数学模型，并利用机器学习算法识别中国古典诗词的情感。

二、 研究方法及理论背景

本次研究主要使用 Python 编程语言，利用词袋模型和字向量的思想，对古诗词进行正负情感的识别。

词向量技术，简而言之就是将词转化成为稠密向量，并且对于相似的词，其对应的词向量也相近。词向量模型中的 word2vec 算法，是将词表征为实数值向量的一种高效的算法模型，其基本思想为：通过训练将每个词映射成 K 维实数向量（K 一般为模型中的超参数），通过词之间的距离（比如 cosine 相似度、欧氏距离等）来判断它们之间的语义相似度。其采用三层的神经网络，核心技术之一是根据词频利用 Huffman 编码，使得所有词频相似的词隐藏层激活的内容基本一致，出现频率越高的词语，他们激活的隐藏层数目越少，从而有效降低了计算的复杂度。中文字向量与词向量类似，也是通过向量的形式来反映各汉字的特征，有利于进行汉字相关度的比较。

词袋模型(bag-of-words model)是用于描述文本的一种简单的数学模型，也是常用的一种文本特征提取方式。词袋模型将一篇文档看作是一个“装有若干词语的袋子”，只考虑词语在文档中出现的次数，而忽略词语的顺序以及句子的结构。词袋模型对文档进行了很大程度的简化，但一定程度上仍然保留了文档的主体信息。忽略难以建模的词句结构、保留体现主题的词语计数，便是词袋模型的基本思想。本课题提出的模型便是一种词袋模型，忽略了词句结构，但保留了各个汉字的原始内涵。

三、 研究过程

本研究主要分为训练与测试两个过程。

1.训练过程

(1)导入数据

选取古诗文网上的大量古诗词，并通过其已有分类（如闺怨、励志、悼亡等）进行人工情感标注。并将文本随机分为 2/3 的训练集和 1/3 测试集。

(2)分割汉字，并标记各字情感

使用 Python 的循环功能分割训练集中的汉字（除去标点符号）。

抽取诗词中的单个字，在积极情感诗词与消极情感诗词库中分别进行字频统计，并根据汉字的出现频率对每个字的情感进行打分（分数值域为 0~1，其中 0 表示极端消极，1 表示极端积极）。

具体流程：

1)计算原始字频。

在积极情感诗词中出现一次：计数器+1;

在消极情感诗词中出现一次：计数器-1.

2)机器校准。

将汉字总频数由高到低排序，得最高频数 x_1，最低频数 x_2($x_1>0$, $x_2<0$).

将 x_2~x_1 映射到 0~1 的值域中，并且确保各字的消极/积极情感属性不变。即若 $x_0<0$，则 $x_0'<0.5$，若 $x_0>0$，则 $x_0'>0.5$。

计算公式如下：

设 x_{max} 为$|x_1|$与$|x_2|$中较大的一个。

则 x 的校准值 $x'=(x-(-x_{max}))/(2x_{max})=(x+x_{max})/(2x_{max})$

3)人工校准。

诗词中有一些直接抒情的字，当出现这些字时，该诗（词）的正负情感便十分明确了，例如出现“喜”、“乐”的诗词多半会表达积极情感，出现“愁”“悲”则表达消极情感。当对于一首古诗（词）进行二分化的情感判别时，可适当调高这些直接抒情字的情感得分，以确保诗词情感可以正确识别。

4)整理字库。

将情感字库进行整合，归入专门文档。

（部分字库实例如下：）

```
In [30]: runfile('D:/02英才计划/课题后期/SentimentAnalysis/
Original/02word_count.py', wdir='D:/02英才计划/课题后期/
SentimentAnalysis/Original')
[('乐', 1.0), ('山', 0.8214285714285714), ('家',
0.8214285714285714), ('笑', 0.8076190476190477), ('开',
0.7738095238095238), ('田', 0.7142857142857143), ('来',
0.7023809523809523), ('村', 0.7023809523809523), ('一',
0.6904761904761905), ('里', 0.6785714285714286), ('爱',
0.6766666666666667), ('好', 0.6766666666666667), ('童',


0.38636363636363635), ('消', 0.38636363636363635), ('肠',
0.38636363636363635), ('泣', 0.37636363636363646), ('妻',
0.37636363636363634), ('柳', 0.375), ('郎', 0.375), ('欲',
0.375), ('离', 0.375), ('桐', 0.375), ('同', 0.375), ('双',
0.36363636363636365), ('残', 0.36363636363636365), ('相',
0.36363636363636365), ('沉', 0.36363636363636365), ('记',
0.36363636363636365), ('怅', 0.35363636363636364), ('寞',
0.35363636363636364), ('思', 0.3522727272727273), ('语',
0.3522727272727273), ('情', 0.3522727272727273), ('倚',
0.3522727272727273), ('阑', 0.3522727272727273), ('处',
0.3409090909090909), ('叹', 0.33090909090909093), ('旧',
0.32954545454545453), ('燕', 0.32954545454545453), ('啼',
0.32954545454545453), ('心', 0.32954545454545453), ('别',
0.32954545454545453), ('哀', 0.3195454545454544), ('何',
0.3181818181818182), ('月', 0.3181818181818182), ('红',
0.3181818181818182), ('香', 0.3068181818181818), ('空',
0.3068181818181818), ('楼', 0.3068181818181818), ('花',
0.29545454545454547), ('谁', 0.29545454545454547), ('伤',
0.2854545454545454), ('以', 0.2727272727272727), ('魂',
0.26136363636363635), ('之', 0.25), ('夜', 0.25), ('断',
0.23863636363636365), ('怨', 0.20590909090909087), ('梦',
0.11363636363636363), ('泪', 0.05818181818181822), ('兮',
0.011363636363636364), ('悲', 0.0), ('恨', 0.0), ('愁', 0.0),
('不', 0.0)]
```

2.测试过程

(1)输入诗词

(2)分割汉字

(3)计算汉字相关度

各汉字间的相关度可以通过字向量的比较获得。中文词向量语料库(Chinese Word Vectors)，包含数十种用不同语料训练的词向量，涵盖各领域，且包含多种训练设置。其中，在《四库全书》的语料库中，提供了古汉语的字向量。本研究将利用该字向量库，并使用 gensim word2vec 接口，来完成汉字相关度的计算。

(4)计算诗词中各字的情感得分

将诗词中的每个字，如果该字在已知情感字库中存在，则使用该字已知的情感得分；若该字不存在于已知情感字库，则利用字向量与已知情感字库中的汉字进行比较，找出与其相关度最高的 5 个，通过加权平均的方式计算该字的情感得分。

计算公式为：$\mathrm{X} = \frac{\sum_{n=0}^{4} a_n b_n}{\sum_{n=0}^{4} a_n}$ （其中 a 表示各相似字与诗中所选字的相关度，b 表示各相似字的情感得分）

（部分“未知字”的情感得分计算结果如下：）

眸:0.5
湛:0.5
丸:0.5
奕:0.47619047619047616
奕:0.47619047619047616
蟠:0.4880952380952381
注:0.5
汴:0.5238095238095238
赴:0.4949528692598268
板:0.5
荧:0.4880952380952381
蝥:0.47619047619047616
推:0.5
咿:0.5
阡:0.4642857142857143
惠:0.5
劝:0.5
劝:0.5
项:0.5238095238095238
掌:0.5357142857142857
拨:0.5

(5)计算该诗的最终情感得分

对诗词的各个情感字得分加和平均，得到该诗词最终情感得分。

例如，一首有 n 个情感字的诗，其最终情感得分将为$\bar{x} = \frac{\sum_{i=0}^{n-1} x_i}{n}$

如果最终得分小于 0.5，则标记为消极诗词；如果最终得分大于 0.5，则标记为积极诗词；如果标记为 0.5，则标记为中性诗词。

以下是部分诗歌得分实例，可见计算机得分基本符合人的情感判别：

鹅鹅鹅，曲项向天歌。白毛浮绿水，红掌拨清波。
0.5207843137254901

缺月挂疏桐，漏断人初静。谁见幽人独往来，缥缈孤鸿影。惊起却回头，有恨无人省。拣尽寒枝不肯栖，寂寞沙洲冷。
0.46908571428571405

日照香炉生紫烟，遥看瀑布挂前川。飞流直下三千尺，疑是银河落九天。
0.5220458553791887

独在异乡为异客，每逢佳节倍思亲。遥知兄弟登高处，遍插茱萸少一人。
0.4885204081632653

(6)计算准确率，优化结果

将机器识别出的诗情感正负结果与测试集原有的人工标注对比，不断调整参数以实现较高的准确率。

四、 研究结果及分析

经过反复调试，机器识别古诗情感的准确率大致在 76.7%左右，还需进一步改进。相比于本人的模型，较为类似的朴素贝叶斯算法多项式模型在进行初步的平滑处理后，准确率大约为 78.6%。仔细分析研究结果，可得出其主要原因在于本课题所设计的模型中，字的情感得分仅和原始词义及其感情色彩有关，而与诗词中具体语境关系不大。想要进一步提高准确率，还需继续改进模型，将字与诗词语境之间的关系引入模型。

	A	B	C	D	E	F	G	H
1							**平均准确率**	
2	**Original**	0.808511	0.765957	0.751773	0.716312	0.794326	**0.76738**	
3	**Bayes**	0.794326	0.803395	0.787037	0.716667	0.83179	**0.78664**	

五、 结论

本课题研究提出了一种基于机器学习的古诗词情感分析方法，有助于运用科学方法佐证人对诗词情感的主观判断，具有较强的针对性。本课题创新点在于：(1)用字向量代替词向量，避免了古诗分词困难、词语重复率低的问题，使识别更加准确。(2)使用加权平均的方法，对古诗词的正负情感进行定量计算，使结果更加可信。

六、 存在的问题及后续思路

对于识别准确率不高、不够稳定等问题，解决方式主要有：(1)继续扩大语料库、调整字库及汉字的分类。(2)引入例如 n-gram 的语言模型，同时处理好古诗中词数据稀疏的问题，提高机器对诗词中汉字排序顺序的识别能力，以使字和语境的关系得以体现。

七、 致谢

非常感谢英才计划给予我这个进入计算机科研领域的机会。在参与英才计划的这一年中，我收获的不仅是知识，更是自我的历练与成长。感谢我的导师吴楠教授，他在本课题的选定、修改及具体实施过程中都提出了宝贵的意见，并最终促成了课题的完成。他与我在计算机方面的多次交流给了我许多灵感和启发，使我受益匪浅，在此我表示深深的感激。此外，感谢学校对我参加英才计划的大力支持，也感谢我的父母，他们的鼓励是使我不断前进的动力。

参考文献

[1] 古诗文网[DB/OL]
[2] 诗词-百度百科[DB/OL]
[3] 汤晓鸥，陈玉琨.人工智能基础（高中版）[M].上海：华东师范大学出版社，2018:124
[4] Chinese Word Vectors：目前最全的中文预训练词向量集合[DB/OL]
[5] Li S, Zhao Z, Hu R, et al. Analogical Reasoning on Chinese Morphological and Semantic Relations[J]. 2018.
[6] 苏劲松，周昌乐，李翼鸿．基于统计抽词和格律的全宋词切分语料库建立[J]．中文信息学报, 2007, 21(2):52-57.
[7] 毛伟，徐蔚然，郭军．基于 n-gram 语言模型和链状朴素贝叶斯分类器的中文文本分类系统[J]．中文信息学报, 2006, 20(3):31-37.
[8] CSDN-词向量及语言模型[DB/OL]