

**数据挖掘期末报告**

|  |  |
| --- | --- |
| 题目： | **古诗词数据挖掘** |
|  |  |

成员： 钟辰丽、王韵、徐嘉辰、柳俊志

2019年 06 月

目 录

# 序论

## 研究内容

钱钟书在《宋词选注》序言部分曾指出，宋人承续唐统，在诗歌方面继续向深入发展，在词的方面则进行了新的探索。这很大程度上出于唐人已经在诗歌方面取得了令后人难以超越的成就——唐人将诗歌中能写的几乎都写了。

这或许能在文本层面找到证据。如果唐人果真把“能写的都写了”，那么显然，后人在进行诗歌创作时会多有“借鉴”。“借鉴”之一便是“意象”——说的更简略些，就是特定的词汇。假若我们掌握了历朝历代的所有诗歌，把它们全部放在一起进行对比分析，尤其对比它们的用词特征，那么此类问题就有可能获得解答。

本研究使用的是量化方法，希望借助**文本挖掘**技术，深入诗歌内部，从用词角度来分析诗歌的历时性特征。

希望回答以下几个问题：

1.不同朝代的诗歌风格之间是否有明显的区别（操作化为，用词特征是否有明显区别）

2.诗歌用词特征是否存在时间维度上的特征——相近时间段的相似度更大（所谓的时代特征），而较大时间差距的诗歌的风格明显不同（所谓的时代差异）

3.情感分析。不同朝代表达的情感是否有差异。

## 1.3 评价指标

是否将个朝代按时间分开。或者能够体现诗词的时代差异。

# 数据分析

## 数据源

本次采用的古诗词数据来源于互联网搜集。数据内容为诗词句三种类型。主要有以下三个数据源，全部为从github上搜集得到：

|  |  |  |
| --- | --- | --- |
| 数据集 | 描述 | 链接 |
| chinese-poetry | 唐宋两朝近一万四千古诗人,接近5.5万首唐诗加26万宋诗.两宋时期1564位词人，21050首词 | https://github.com/chinese-poetry/chinese-poetry |
| poetry | 2017年从古诗文网爬取, 73281首古诗词和3156个诗人的详细数据 | https://github.com/hujiaweibujidao/poetry |
| Poetry | 非常全的古诗词数据，收录了从先秦到现代的共计85万余首古诗词。 | https://github.com/Werneror/Poetry |

## 数据解读

诗词数量总体分布

|  |  |  |  |
| --- | --- | --- | --- |
| 朝代 | 诗词数 | 诗人数 | 高产诗人 |
| 先秦 | 1129 | 31 | 诗经306 |
| 汉 | 557 | 111 | 刘向24 |
| 魏晋 | 3313 | 263 | 曹植172 |
| 隋 | 1652 | 126 | 李世民100 |
| 唐 | 144763 | 5689 | 白居易8582 |
| 五代 | 120 | 26 | 李煜22 |
| 宋 | 585159 | 15745 | 陆游20087 |
| 辽 | 22 | 7 | 萧观音 12 |
| 金 | 5815 | 269 | 李俊民 964 |
| 元 | 54461 | 1455 | 刘基 1731 |
| 明 | 254867 | 4648 | 王世贞8013 |
| 清 | 90587 | 8920 | 丘逢甲 2147 |
| 近代 | 44314 | 812 | 黄浚 1277 |
| 民国 | 17313 | 108 | 曹家达 1874 |
| 当代 | 28208 | 177 | 卢青山 3133 |
| 未知 | 34 | --- | --- |
| 总计 |  |  |  |

# 分析过程

## 全局高频字分析

去除停用词后的全局高频字top-100及各个朝代的高频字top-10。

|  |  |  |  |
| --- | --- | --- | --- |
| 朝代 | 字 | 字频 | 词云 |
| 先秦 |  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| 朝代 | 字 | 字频 | 占比（字频/诗词数） |
| 唐朝 | 月 | 1000 | 1000/144763=0.006 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## 常用双字词发现

词汇探测：唐诗中的常用双字词发现

刚才分析的是单字，而汉语的语素大都是由单音节（字）表示，即所谓的“一音一义”。当这些单音节语素，能够独立应用的话，就是词。古汉语中存在着许多单音节词，这也就是文言文翻译中要经常把一个字翻译成现代汉语中的双音节词的原因。

然而，有些单音节语素，不能够独立使用，就不是词，只能够是语素，如“第-“、”踌-“、”- 们“。

鉴于此，笔者想发现一些唐诗中的常用双字词，看看其中的成词规律是怎样的。这里选取共现次数超过10次的词汇，并列出TOP200的共现双字词和各个朝代的TOP10共现双字词。关键操作步骤如下所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 朝代 | 词 | 词频 | 词云 |
| 唐朝 |  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## TF-IDF诗词关键词分析

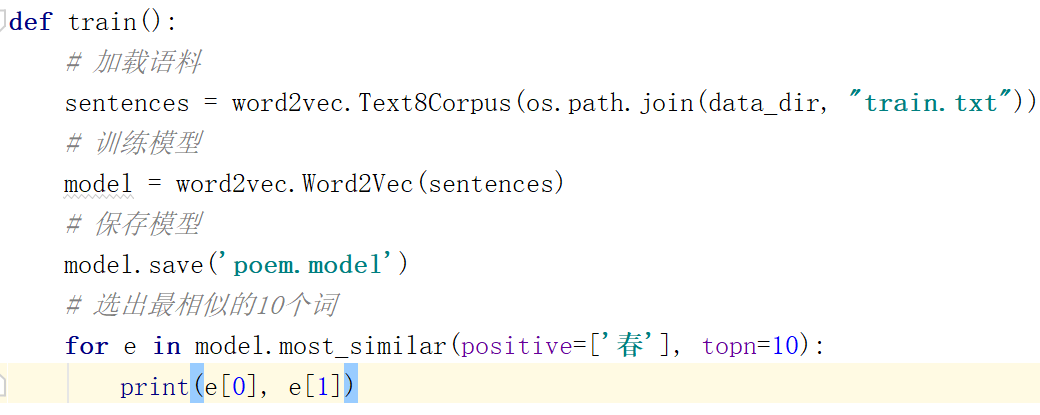
|  |  |  |  |
| --- | --- | --- | --- |
| 朝代 | 词 | 词频 | 词云 |
|  |  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Word2Vec词向量分析

基于的Word2vec的关联字分析

基于Word2vec的词向量能从大量未标注的普通文本数据中无监督地学习到词向量，而且这些字向量包含了字与字之间的语义关系，正如现实世界中的“物以类聚，类以群分”一样，一个词可以由它们身边的词来定义。

从原理上讲，基于字嵌入的Word2vec是指把一个维数为所有字的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单字被映射为实数域上的向量。把每个单字变成一个向量，目的还是为了方便计算，比如“求单字A的同义字”，就可以通过“求与单字A在cos距离下最相似的向量”来做到。

词向量能从大量未标注的普通文本数据中无监督地学习到词向量，而且这些词向量包含了词之间的语义关系，正如现实世界中的“物以类聚，类以群分”一样，字词可以由它们身边的字来定义。我们用jieba分词对文本进行切分，然后使用genism Word2vec训练得到词向量。

下面，选取一些单字进行字向量关联分析，展示如下：

|  |  |
| --- | --- |
| 字 | 关联字 |
| 春 | 清明，时候，春事，芳菲，寒食，花，春色，春光，风光，东风 |
| 思乡 | 望乡，邮亭，乡思，消魂，海边，凄然，别情，迢遥，上楼，落花流水 |
| 梅 | 春思，春晚，早梅，西园，春景，桃，咏柳，赏春，如梦令，闺思 |
| 冬 | 九月，四月，早春，大作，闰，十月，节过，寄王，晦日，西亭 |
| 中秋 | 南楼，重阳，九日，登高，月夜，秋夜，暮春，江楼，元宵，湖上 |

可以看出还是有很高的相关度，我们手动挑选其中的一下词作为关键词给诗词打标签。

## 探究诗词中的七情

我们进一步进行了多维情绪分析，为了丰富分析维度，不采用简单的二元分析，即“积极”和“消极”2种情绪，而是7种细颗粒的情绪分类，即悲、惧、乐、怒、思、喜、忧。

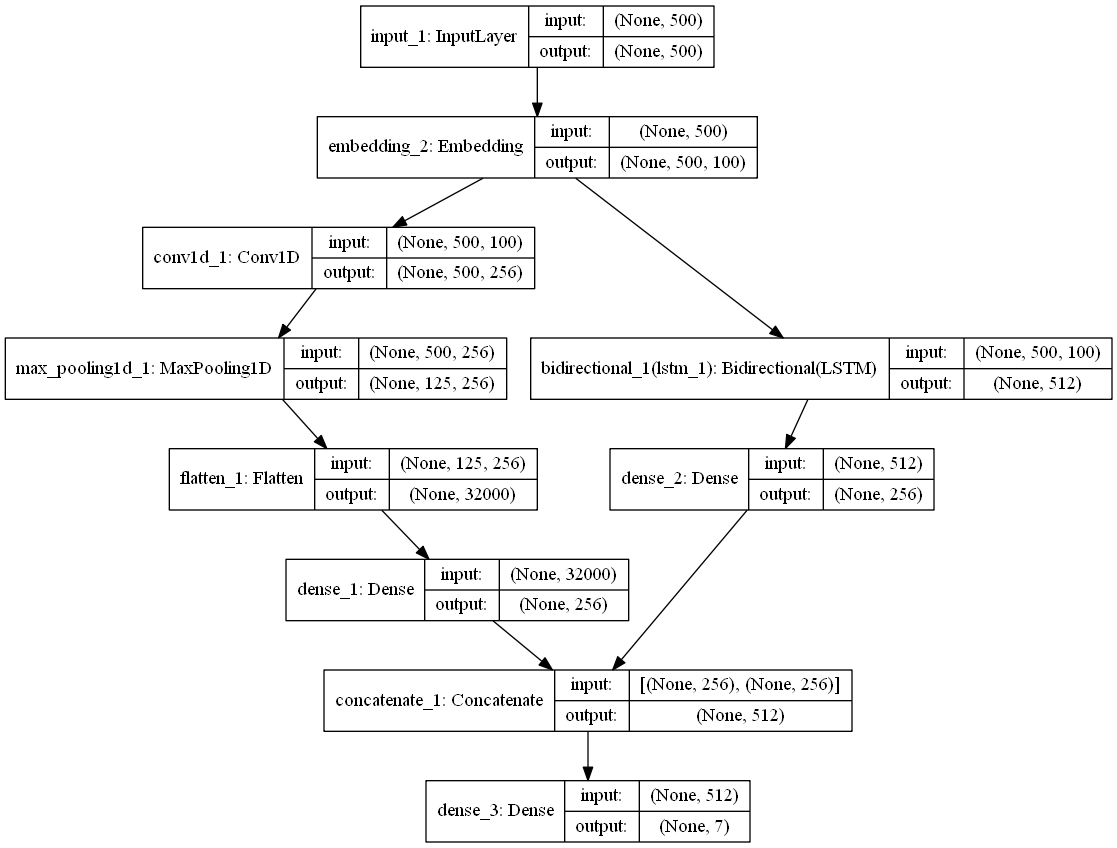
根据上面获取到的字向量，经过人工遴选后，得到可以用于训练的“情绪字典”，根据诗歌中常见的主题类别，情绪类别分为：

|  |  |  |
| --- | --- | --- |
| 标签类型 | 标签 | 关键词 |
| 情绪 | 悲 | 愁、恸、痛、寡、哀、伤、嗟… |
| 惧 | 谗、谤、患、罪、诈、惧、诬… |
| 乐 | 悦、欣、乐、怡、洽、畅、愉… |
| 怒 | 怒、雷、吼、霆、霹、猛、轰… |
| 思 | 思、忆、怀、恨、吟、逢、期… |
| 喜 | 喜、健、倩、贺、好、良、善… |
| 忧 | 恤、忧、痾、虑、艰、遑、厄… |

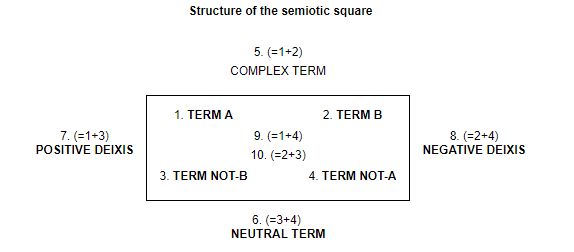
根据以上的得到的标签关键字，在诗词数据的背景、赏析、内容、翻译中匹配这些关键字，给诗词加上对应标签，使用神经网络模型对诗词进行情感分类。

### CNN pair LSTM

1. 网络结构图



## 文本对比，用Semiotic Squares比较各朝代诗词用词特点

“Semiotic Squares"，是一种提炼式的对比分析（Oppositional Analyses）方法，通过将给定的两个相反的概念/事例（如 “生命（Life）”和“死亡（Death）”）的分析类型拓展到4类（如“生命（Life）”、“死亡（Death）”、“生死相间（也就是活死人，The Living Dead）”、“非生非死（天使，Angels）”，有时还可以拓展到8个或10个分析维度。以下是符号方块的结构示意图：

说明：“+”符号将2个词项组合成一个“元词项（Metaterm）”，也称“复合词（Compound Term）”，例如，5是1和2的复合结果。

（1）词项（Terms）

Semiotic Square 由4个词项组成：:

位置1 （Term 1）：词项A（Term A）

位置 2（Term 2）：词项B（Term B）

位置 3 (Term Not-2)：非B词项（Term Not-B）

位置 4 (Term Not-1)：非A词项（Term Not-A）

Term A和TermB是相反的两个概念，二者是对立关系，这是“符号方块”的基础，另外两项是通过对Term A和Term B取反而获得。

（2）元词项 (Metaterms)

Semiotic Square囊括6个元词项。这些元词项由上面的4个基础词项组合而成，其中的绝大部分元词项已被命名。

位置5 (Term 1 + Term 2):复合词项（Complex Term）

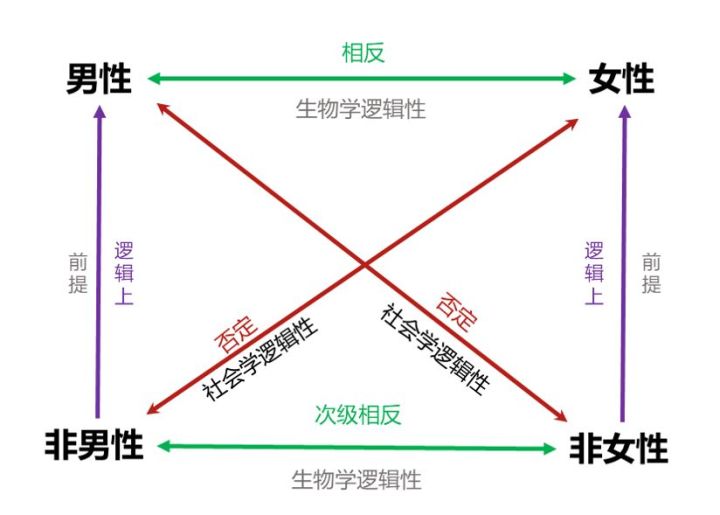
位置6 (Term 3 + Term 4):中立词项（Neutral Term）

位置7 (Term 1 + Term 3):正向系（Positive Deixis）

位置8 (Term 2 + Term 4):负向系（Negative Deixis）

位置9（Term 1 + Term 4):未命名（Unnamed）

位置10（Term 2 + Term 3):未命名（Unnamed）

下面以“男性”和“女性”这两个相对的概念来举个例子，注意其中错综复杂的逻辑关系/类型。

用Semiotic Squares进行分析的结果如下图所示：

|  |  |  |  |
| --- | --- | --- | --- |
| 朝代 | 独有高频字 | 字频 | 词云 |
| 唐朝 |  |  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## 分析结论

通过对历代诗词，情感，用词特点的对比分析。

通过全局双字词、TF-IDF、Word2Vec、神经网络、Semiotic Squares等方法对诗词的用词特点和情感变化做了分析。可以看出…

# 参考文献

[1] 陈运文. 用文本挖掘分析了5万首《全唐诗》，竟然发现这些秘密 - 陈运文的文章 - 知乎<https://zhuanlan.zhihu.com/p/45415824> . 2018.10.9/2019.05.04