

注：教师姓名后留有一个空格，后面填写教师职称。下面加下划线。

阅后删除此文本框。

**本科毕业论文**

**安史之乱对唐诗的影响分析**

**--基于情感分析方法**

|  |  |  |
| --- | --- | --- |
| **学生姓名** | **：** | 黄永静 |
| **学号** | **：** | 151543220 |
| **学院** | **：** | 互联网金融与信息工程学院 |
| **专业** | **：** | 计算机科学与技术 |
| **指导教师** | **：** | 韩冬 **职称：** 讲师 |
| **提交日期** | **：** | 2017 年 03 月 27 日 |

**本科毕业论文（设计）诚信声明**

本人郑重声明：所呈交的本科毕业论文（设计），是本人在指导老师的指导下，独立进行研究工作所取得的成果，成果不存在知识产权争议，除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

 学生签名：

时间： 年 月 日

**关于论文（设计）使用授权的说明**

本人完全了解广东金融学院关于收集、保存、使用学位论文的规定，即：

1.按照学校要求提交学位论文的印刷本和电子版本；

2.学校有权保存学位论文的印刷本和电子版本，并提供目录检索与阅览服务，在校园网上提供服务；

3.学校可以采用影印、缩印、数字化或其它复制手段保存论文。

本人同意上述规定。

学生签名：

时间： 年 月 日

摘 要

唐诗是中华民族珍贵的文化遗产之一。唐诗取材甚广，反应了当时社会的各个方面，触及到当时社会生活的每一个角落。755年末安史之乱爆发，导致人民流离失所，社会动荡，唐诗诗风也随之改变由浪漫豪爽到深沉悲凉，文人心态有积极自信到失落怀疑。

本文主要以安史之乱前后的唐诗作为研究对象，采用Word2vec计算诗歌字与情绪字之间的相似度，再引入TF-IDF模型，以全唐诗单字切割后作为数据集，评估诗中的字对全唐诗集的重要程度，确定诗歌中字的权重。Word2vec与TF-IDF结合分析古诗的二元情感，达到判断古诗词的情感目的。经过对古诗的情感分析以及词频分析，安史之乱前唐诗主要描写山川，白云，宫廷生活为主，诗歌多带有积极的情绪，而战乱爆发后唐诗诗风开始走向惆怅，消极，怀古伤今。

**[关键词]**：唐诗；安史之乱；文本情感分析

Abstract

Poetry of the Tang Dynasty is one of the precious cultural heritage of the China. Poetry of the Tang Dynasty has a wide range of materials, reflecting all aspects of society at that time, touching every corner of social life at that time. At the end of 755, the Anshi Rebellion broke out, resulting in the displacement of people and social unrest. The style of poetry of the Tang Dynasty changed from romantic to deep sadness, and the mentality of literati was from positive self-confidence to loss of suspicion.

This paper mainly takes poetry of the Tang Dynasty before and after the Anshi Rebellion as the research object, calculates the similarity between poetry words and emotional words by Word2vec, then introduces the TF-IDF model, takes the whole poetry of the Tang Dynasty word cut as the data set, evaluates the importance of the words in the poetry to the whole poetry of the Tang Dynasty collection, and determines the weight of the words in the poetry. Word2vec and TF-IDF combine to analyze the binary emotions of poetry, so as to judge the emotional purpose of poetry. Through the emotional analysis and frequency analysis of the poetry of the Tang Dynasty, the poetry of the Tang Dynasty before the Anshi Rebellion mainly described mountains and rivers, clouds and court life, and the poetry were mostly with positive emotions. After the outbreak of war, the poetry style of the Tang Dynasty began to move towards melancholy, passivity and nostalgia for the past and the present.

**[Key Words]:** Tang Poetry; Anshi Rebellion; Sentiment Analysis

**目 录**

[摘 要 I](#_Toc4671632)

[Abstract II](#_Toc4671633)

[1 绪论 1](#_Toc4671634)

[1.1 研究现状 1](#_Toc4671635)

[1.2研究意义 1](#_Toc4671636)

[1.3论文主要内容 1](#_Toc4671637)

[2 文本情感分析技术 2](#_Toc4671638)

[2.1文本情感分析技术 2](#_Toc4671639)

[2.1.1概述 2](#_Toc4671640)

[2.2文本情感分析技术 3](#_Toc4671641)

[3 古诗情感分析模型 4](#_Toc4671642)

[3.1诗词分词 4](#_Toc4671643)

[3.2构造情感字典 5](#_Toc4671644)

[3.3情感倾向判断方法 6](#_Toc4671645)

[3.3.1 权重计算 6](#_Toc4671646)

[3.3.2 Word2vec与TF-IDF结合 6](#_Toc4671647)

[4 诗歌分析及展示 7](#_Toc4671648)

[4.1模型评价 7](#_Toc4671649)

[4.2古诗分析 8](#_Toc4671650)

[4.2.1 诗歌来源 8](#_Toc4671651)

[4.2.2全唐诗分析 10](#_Toc4671652)

[4.2.3 代表性诗人分析 12](#_Toc4671653)

[4.2.4 派别分析 14](#_Toc4671654)

[4.2.5 安史之乱前后古诗分析 15](#_Toc4671655)

[5 总结与展望 16](#_Toc4671656)

[5.1总结 16](#_Toc4671657)

[5.2不足 17](#_Toc4671658)

[5.3展望 17](#_Toc4671659)

[参考文献 18](#_Toc4671660)

[致 谢 19](#_Toc4671661)

**安史之乱对唐诗的影响分析--基于情感分析方法**

1 绪论

1.1 研究现状

国内已经有不少学者进行了安史之乱对唐诗的影响的研究。文献[[1](#_ENREF_1)]总结出由于安史之乱导致国家岌岌可危，这个时代的诗歌多是写现实生活，写社会的动荡、人民疾苦和无限忧虑的国家之叹，战乱爆发后创作的诗歌多为沉重，忧虑的。文献[[2](#_ENREF_2)]从曲江文人的心态，曲江诗的诗风以及曲江诗的文化内涵，阐述了曲江诗经历了安史之乱由浪漫豪爽转变为深沉悲凉。文献[[3](#_ENREF_3)]指出安史之乱的爆发，摧垮的一代人的精神，对政治失去了兴趣，诗歌染上了哀伤之情多于济世救国之志。文献[[4](#_ENREF_4)]以山水为研究对象，指出战争的爆发使积极人世的文人被排挤到社会的底层，往日的雄心壮志被消磨一空，内心苦闷空虚，安史之乱爆发后山水诗歌由初唐的明丽轻快、自然清新转变为凄清萧瑟，寂寞孤苦。综述，就目前来说对于安史之乱对唐诗的影响仅仅停留在传统的研究方法上，研究对象往往是一个派别的诗歌或具有代表性的诗人而不是全唐诗。

1.2研究意义

古诗如同当下的“朋友圈”、“微博”一样作为当时人们表达情感的方式，安史之乱这一重大的事件，必然会引起当时人们的“牢骚”，就如同当下发生重大时间，人纷纷通过社交平台发表自己的看法。单首古诗人们可以容易理解，但对于大量的总体的古诗通过传统的理解方法进行研究就相当的费时费力，因此有必要引入现有的文本情感分析技术，借助机器分析古诗的情感，对古诗的分析研究对理解当时的社会文化对研究唐代的历史有着积极的意义。

1.3论文主要内容

本论文的内容主要包括一下三个方面：

（1）分析当前文本情感分析技术

对文本情感分析技术进行分析与总结，并介绍本文采用的文本情感分析技术。

（2）构建古诗情感分析模型

主要介绍诗句的分词方法，情感字典构造，诗字权重判断，古诗的情感倾向判断，并用Python实现。

（3）实验结果展示与分析

根据构建的古诗情感分析模型，对古诗进行情感分析，对得到的结果进行展示与分析。

2 文本情感分析技术

2.1文本情感分析技术

2.1.1概述

文本情感分析，又称为意见挖掘、倾向挖掘，情感分类等，是指用自然语言处理、文本挖掘以及计算机语言学等方法来识别和提取原素材中的主观信息[5]。通常来说，情感分析的目的是为了找一段文本的两极态度，这个态度或许是他或她的个人判断或是评估，也许是他当时的情感状态，或是作者有意向的情感交流。文本情感分析的一个基本步骤是对某段已知文字进行两极分类，这个分类可能是在句子级、功能级或篇章级。分类的作用就是判断出文本中表述的观点是积极的、消极的、还是中性的情绪。更高维度的情感分析还会寻找更复杂的情绪状态，比如“喜”、“悲”、“怒”、“思”等等。早在2000年之前也有一些对情感分析的工作，但是由于互联网没有广泛普及，可以用到的数据并不多，情感分析并没有得到大家的重视和大规模研究。这几十年来，互联网发展越来越快，社交网络有着大量用户的积极参与，产生了大量的对电影，商品，事件有价值的的评论信息。这些评价信息充斥着用户各种各样的情感倾向与情感色彩，例如喜，怒，哀，乐，赞成，反对等。在国外，Sentiment Analysis这个词第一次出现在Nasukawa and Yi[6]。在国内，2007年开始逐渐引起学者的关注，文献[7]总结了文本情感分类的方法，提出了基于“近似文本分析”的情感分类方法ATA-SC。该方法考虑了实体词汇与情感词汇之间的语义关联，因此对于情感的识别能力要强于基于单对象假设的情感分类方法。文献[8]提出并实现了基于语义的文本情感倾向性算法。该算法通过分析文本中的情感词以及由这些情感词构造的短语模式,来作为文本的情感特征,并利用这些特征实现文本的情感判别。文献[9]采用SO-PMI算法扩大了情感词典的覆盖面，构建了一个面向微博的情感词典。采用微博文本中的情感词作为特征选择，经过实验该方法最高准确率为74.2%。文献[10]提出一种基于情感字典的与朴素贝叶斯理论的本分分类方法。文献[11]使用迁移学习的方式，对唐诗进行了研究。文献[12]提出一种基于词典与机器学习相结合的微博情感分类方法，采用空间模型表示微博文本，提出基于层次的结构的特征将维方法，借助支持向量机机器学习模型将中文微博分为正向、负向和中性三类。

2.2文本情感分析技术

文本情感分析方法主要分为两类

（1）基于情感字典的分本情感分析

基于词典的文本情感分析方法，主要是利用已经建立好的带有情绪极性的情感字典与待判断的文本进行比较，如匹配成功则可以判断待判断文本的情感极性。

（2）基于机器学习的文本情感分析

文本情感分析技术主要是二元分析，即时判断文本是倾向积极还是消极的，因此可以归类到分类问题。所以可以用已经标注好的文本作为训练，用分类方法如朴素贝叶斯、支持向量机、K-邻近算法对待分类的文本进行分类，达到判断情感倾向的目的。

本文采用的是基于情感字典的文本情感分析。绝大多数的古诗中并没有直接的表现出明显的情感，用情感词语作为特征选择显然是不可行的，因此本文采用基于语义的信息的语义相似度情感判断方法，也就是通过诗句中的字与积极或者消极的情感字的语义相似度来判断诗字的情感倾向，达到判断诗歌的情感倾向。获取诗字的语义相关度采用Word2vec。Word2vec是Google在2013年推出的一款用于训练词向量的工具，Word2vec提供了一种使用分布式向量对文本进行表示的方法。与传统文本向量空间模型相比，使用Word2vec模型表示文本，既能解决传统向量空间模型的高维稀疏特征问题，还能引入传统模型不具有的语义特征，有助于短文本分类。Word2vec通过训练，可以把对文本内容的处理简化为K维向量空间中的向量运算，而向量空间上的相似度可以用来表示文本语义上的相似度，利用Word2vec训练学习得到的词向量表示，考虑了词语之间的语义关联关系，利用诗句的中字词的与情感词的相似度，达到判断情感倾向的目的。例如“秋”与“悲”，“欢”与“酒”，“愁”与“酒”的相似度如图2.1所示。

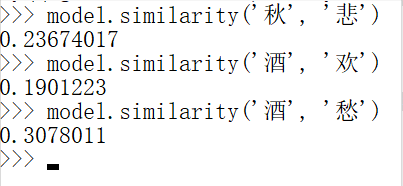


图2.1 相似度示例图

图2.1是采用人民日报语料作训练而成的模型得出的结果[13]。文献[14]中使用 Word2vec与TF-IDF对中文微博进行情感判断平均F值能达到0.9。本文将采用全唐诗单字切割作为语料，窗口设置为5，窗口值指训练过程中当前词前后5个词语范围内的字进行向量计算，除去次数低于5的字，维度为100维，维度一般在100-300之间，维度代表了字词之间的特征，维度越大特征的区分越明显，但过高的维度会淡化字与字之间的关系，全唐诗数据约为10兆，数据比较少，因此选择100维度即可，迭代50次训练成模型用作诗字与情绪字的语义相似度计算并结合TF-IDF计算诗歌中字的权重。

# 3 古诗情感分析模型

3.1诗词分词

汉字是一套书写符号，但这套符号有个比较特殊的地方，“字”不仅是一个音节单位，同时也可以是承载意义的单位，即语素。这是汉语比较有特色地方，即汉语的语素大都是由单音节（字）表示，即所谓的“一音一义”。而当这些单音节语素，能够独立应用的话，就是词。这个在汉语中非常明显，尤其是一些基本词汇中，如“上”、“喜”、“望”、“花”、“走”等。对于古汉语，单字词占古汉语词汇统计信息的80％以上[15]，古诗的每一个字都是经过诗人千敲万琢的，可以说是微言大义，字字千钧。因此现代的分词模型并不适合古诗的情感分析，本文采用单字划分的方法。而对于词频的统计，则采用THULAC进行[16]。

3.2构造情感字典

情感词典在文献[9]中已经进行总结过了。目前在互联网上的两大基础情感词典是知网（HowNet）与台湾大学NTUSD。2007年07月22日知网发布新版本的情感分析用词语集，它包含中文词语（词形）91016，英文词语（词形）85295， 其中包括新词语“长不大的成年人”、“朝核问题”、“去功能化”、“防地雷反伏击车”等。NTUSD的英文全称是National Taiwan University Sentiment Dictionary，它是由台湾大学整理并发布的情感词典，分为繁体中文和简体中文两个版本。两个版本都包括有2810个正面情感词语和8276负面情感词语。本文是采用语义相似度作为情感判断的，因此并不需要囊括所有情感字词的极大情感字典，相反的为了避免噪音字的引入，仅仅需要尽可能少而全的情感词典就可。本文采用的情感词典取自于《知网》情感分析用词语集[17]。本文的文本处理对象是诗歌，因此仅仅选用其中的正面情感词语（中文）以及负面情感词语（中文），作为基础情感词典。由于分词模式采用逐字切割，因此对基础情感词典进行适当的处理，仅仅留下单个表达情绪的字预处理后如图3.1，3.2所示。

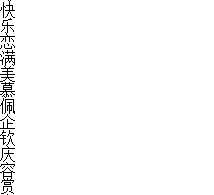


图3.1 部分积极情绪字



图3.2 部分消极情绪字

得到的单字情绪字典需要进一步进行人工筛选，出去其中的噪音字。比如“思”原本是在积极的情绪字典里面的，但是在古诗中多为表达思念，思乡，思家的消极之情，应该归到消极的情绪字典中，积极情绪字典中“望”一字没有明显的情绪倾向，因此应当删除，避免影响判断。

3.3情感倾向判断方法

3.3.1 权重计算

TF-IDF（Term Frequency-inverse Document Frequency,词频与逆文档频率）是一种统计方法，用以评估一个字词对于文件集或语料库中的一份文档或一个类别的重要程度[18]。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TF-IDF实际上是：TF \* IDF，TF词频(Term Frequency)，IDF逆向文件频率(Inverse Document Frequency)。TF表示词条在文档d中出现的频率。IDF的主要思想是：如果包含词条t的文档越少，也就是n越小，IDF越大，则说明词条t具有很好的类别区分能力。如果某一类文档C中包含词条t的文档数为m，而其它类包含t的文档总数为k，显然所有包含t的文档数n=m+k，当m大的时候，n也大，按照IDF公式得到的IDF的值会小，就说明该词条t类别区分能力不强。

3.3.2 Word2vec与TF-IDF结合

Word2vec 是 MIKOLOV 等人提出的一种神经网络概率语言模型，可以用于计算单词的词向量。与传统的高维词向量 one-hot representation 相比，Word2vec 词向量的维度通常在 100~300 维之间，减少了计算的复杂度，也不会造成向量维数灾难。除此之外，Word2vec词向量是根据词汇所在上下文计算出的，充分捕获了上下文的语义信息，很容易通过它计算两个词汇的相似程度。Word2vec包含了两种训练模型，分别是CBOW（Continuous Bag-of-words Model）和 Skip\_gram。本文采用Skip\_gram模型。情感倾向的计算公式如公式3.1所示。

（3.1）

其中e为情绪倾向度，w为权重，p为语义相似度。计算出积极与消极的情感倾向度，再进行比较，得出古诗的情感倾向。

# 4 诗歌分析及展示

4.1模型评价

诗歌的情感判别问题的本质也是二元分类问题，因此可以用二元分类的评价标准对本次实验建立的模型进行评估。目前通常采用的评价指标有准确率（Precision）、错误率（Error rate）、灵敏度（Sensitive）、特效度（Specificity）、精确率（Precision）、召回率（Recall）、F值(F-Measure)等。因为本次的分配是情感分类，因此不采用精确率，召回率以及F值作为评价标准。准确率即是被分对的样本除以所有的样本数，灵敏度表示所有的积极诗歌中被判别为积极的比例，衡量模型对积极类型诗歌的识别能力，特效度表示所有的消极诗歌中被判别为消极的比例，衡量模型对积极类型诗歌的识别能力。准确率、灵敏度和特效的计算公式如公式4.1，4.2，4.3所示。

（4.1）

（4.2）

（4.3）

其中的TP为将正类预测为正类数，TN为将负类预测为负类数，FP为将负类预测为正类数，FN将正类预测为负类数。

实验选择了50首唐诗进行了情感分析，其中表达积极的情绪诗歌25首，表达消极的诗歌25首，部分诗歌如表4.1所示。

表4.1 部分诗歌

|  |  |  |  |
| --- | --- | --- | --- |
| Id | 题目 | 情感 | 情绪 |
| 01 | 望岳 | 充满了诗人青年时代的浪漫与激情 | 积极 |
| 02 | 春夜喜雨 | 看见雨意正浓，就情不自禁地想象天明以后春色满城的美景。其无限喜悦的心情， | 积极 |
| 03 | 闻官军收河南河北 | 诗人听到官军收复失地的消息后，十分的喜悦 | 积极 |
| 04 | 客中作 | 抒写了诗人身虽为客，却乐而不觉身在他乡的乐观情感 | 积极 |
| 05 | 登鹳雀楼 | 盛唐时期人们积极向上的进取精神 | 积极 |
| 26 | 伊州歌 | 女子在秋夜里苦苦思念远征丈夫 | 消极 |
| 27 | 新年作 | 诗的情感哀切深至，颔联意绪剀切，首尾感叹往复。唯颈联写景，淡密而不显焕 | 消极 |
| 28 | 送李判官之润州行营 | 以春天春草“不肯留”“送马蹄”的无情反衬出离别者内心的有情。 | 消极 |
| 29 | 谪仙怨·晴川落日初低 | 对友人的深切思念和被谪的遗恨 | 消极 |
| 30 | 送李中丞之襄州 / 送李中丞归汉阳别业 | 此诗抒发作者对主人公被斥退罢归的惋惜不满与感慨之情 | 消极 |

实验结果如图4.1所示

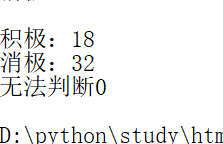


图4.1 部分诗歌实验结果

根据程序输出日志，TP值为16，TN值为23，FP值为2,FN值为9。模型准确率为78%，特效度为64%，灵敏度为92%。从实验结果来看，模型对消极诗歌的识别率较高，准确率达78%，基本能判断古诗的情感倾向。

4.2古诗分析

4.2.1 诗歌来源

诗歌来自全唐诗数据库[19]，其中收录2528位诗人共43030首古诗。如图4.2，4.3，4.4所示。

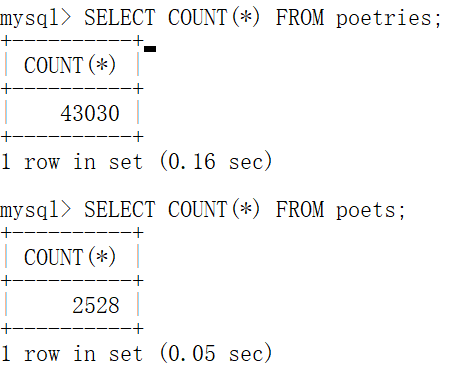


图4.2 全唐诗与诗人数据统计

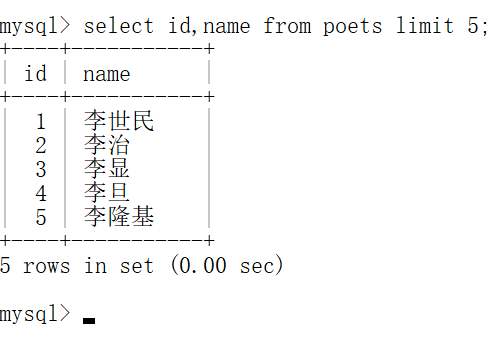


图4.3 部分诗人

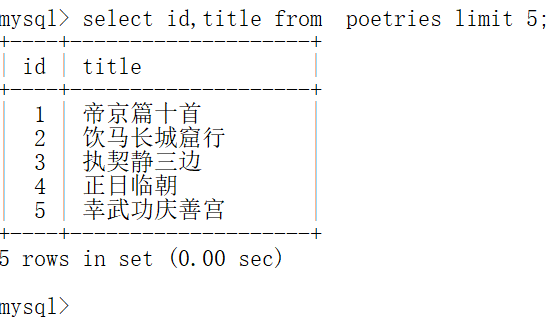


图4.4 部分诗歌

诗歌的创作时间通过爬取百度百科的方式进行获取如图4.5所示。

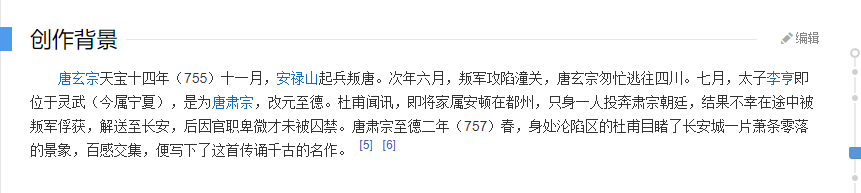


图4.5 诗歌创作时间

通过爬虫方式能够确定具体创作时间的有1237首。

4.2.2全唐诗分析

唐诗按照时期来分，可以分为初唐，盛唐，中唐，晚唐。每个时期诗歌的取向是不一样的。初唐诗坛有两大创作取向，可以把初唐诗人分为二类：一类是围绕在唐太宗和唐高宗、武则天身边的宫廷诗人；另一类是与宫廷关系较为疏远、或者是飘游在山野之中的诗人。诗人的审美理想和审美情趣直接影响着他们诗歌创作的取向和诗歌风格的形成，代表诗人有陈子昂，上官仪，王绩，王勃、杨炯、卢照邻、骆宾王等。热情洋溢、豪迈奔放、具有郁勃浓烈的浪漫气质，是盛唐诗的主要特征代表性诗人岑参，李颀，王昌龄，王维，孟浩然，储光羲，李白等。中唐国家衰落，诗歌有反映民生疾苦，反映现实、批评时政的政治，韦应物、白居易，元稹，韩愈，孟郊，贾岛，柳宗元等。晚唐诗歌多表现为伤时忧国，哀怨深沉，没落之感。代表性诗人有李商隐，杜牧，温庭筠，韦庄等。本文在每个时期选取十个具有代表性诗人的诗歌进行分析。分析结果如图4.6，4.7，4.8，4.9所示。



图4.6 初唐词云图

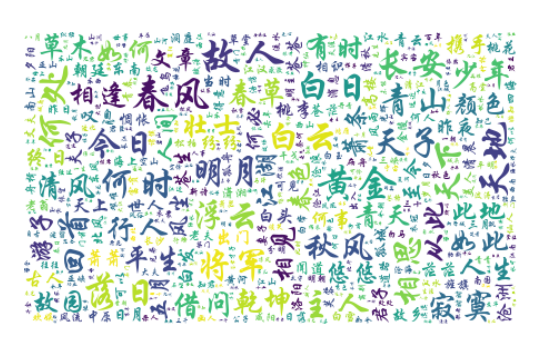


图4.7 盛唐词云图



图4.8 中唐词云图

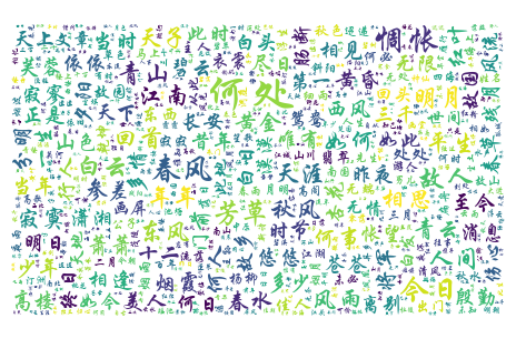


图4.9 晚唐词云图

从每个时期的诗歌词云图来看初唐诗歌主要分为描写“天子”，“歌舞”，描写“白云”，“山川”，也就如同本节开头所说的描写宫廷生活的上官体，以及远离宫廷的山野诗。进入盛唐，诗歌可谓百花齐放，题材广阔，白云青山的美景描写，对首都长安的追求，天地明月的豪情壮阔，寂寞相思的凄凄戚戚。进入中唐“天子”二字已经逐渐淡出诗人的创作视野，值得注意的是词语出现了“三十”这一个数字，带有这个词语诗句有“未年三十生白发”，“三十登宦途”，“蹉跎三十载”，“三十无人知”等如图4.10所示。

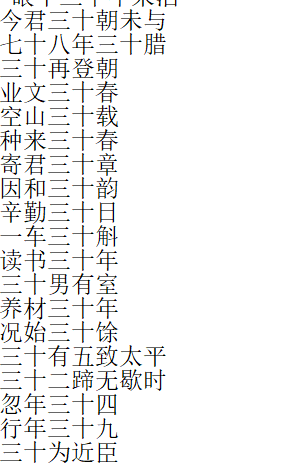


图4.10 晚唐关于“三十”诗句

看来到了中唐时期，大多数诗人已经开始感叹光阴不在，壮志未酬。晚唐，瞩目的是“惆怅”二字，“惆怅异乡人”，“惆怅溪边书细沙”，“昨夜回舟更惆怅”。

4.2.3 代表性诗人分析

杜甫，我国唐代伟大诗人之一，他一生坎坷，应举不第，应试落第，官场不得志，亲眼目睹了唐朝上层社会的奢靡与社会危机天宝十四载（755年），安史之乱爆发，潼关失守，杜甫先后辗转多地。杜甫的一生正处于大唐帝国由盛而衰的关键时期。这样的历史时代，使他的一生颠沛流离，尝尽种种的痛若。从个人生活来说，这实在是大不幸的事情。然而，这种大不幸却成为杜甫艺术生命的大幸。对于杜甫本文从安史之乱前，后分别对他所创作的诗歌进行分析。爬虫得到的结果，能够确认的安史之乱前杜甫创作的诗歌有29首，安史之乱后杜甫创作的诗歌有148首。分析结果如图4.11，4.12，4.13，4.14所示。



图4.11 安史之乱前杜甫诗歌词云图



图4.12 安史之乱后杜甫诗歌词云图

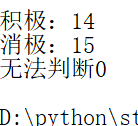


图4.13 安史之乱前杜甫诗歌情感分析

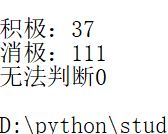


图4.14 安史之乱前杜甫诗歌情感分析

从分析结果可知，安史之乱前杜甫创作的诗歌，消极率约为50%，安史之乱后消极率约为75%，从词频的结果来看，经历安史之乱后杜甫诗歌更多的谈论“可怜”，“叹息”，“寂寞”，“萧瑟”可见战乱使这位诗人愁更愁。

4.2.4 派别分析

唐诗主要分为山水田园派，边塞诗派，浪漫诗派，现实派，本文选取山水田园派进行分析。山水田园派以山水等自然景观为主要描写对象，歌咏田园生活，大多以农村的景物和农民、牧人、渔父等的劳动为题材。诗人们以自然山水或农村自然景物、田园生活为吟咏对象，把细腻的笔触投向静谧的山林，悠闲的田野，创造出一种田园牧歌式的生活，借以表达对现实的不满对宁静平和生活的向往[20]。其中代表性诗人有王维、孟浩然、柳宗元、储光羲等，以下选取代表性诗人的诗歌共1494首古诗进行词频与情感分析，结果如图4.15，4.16所示。

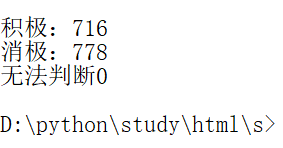


图4.15 情感判断结果



图4.16 词频词云图

根据词频词云图，田园诗歌最多描绘的景物有春风，明月，白云，青山比如“莫问春风动杨柳”，“明月松间照”，“逍遥白云外”，“青山结茅屋”。表达积极情绪的诗与消极情绪的诗歌为一比一，结合词云可以知道，田园派诗歌除了描绘田园生活的闲适之外也表达对故人的思念以及自己的寂寞之情，如“故人不可见”，“为报故人憔悴尽”，“寂寞到城阙”，“可怜寂寞到长亭”。

4.2.5 安史之乱前后古诗分析

通过爬虫采集到的安史之乱前的诗歌共504首，安史之乱后创作的诗歌769首分析如图4.17，4.18，4.19所示。



图4.17 安史之乱前词频词云图

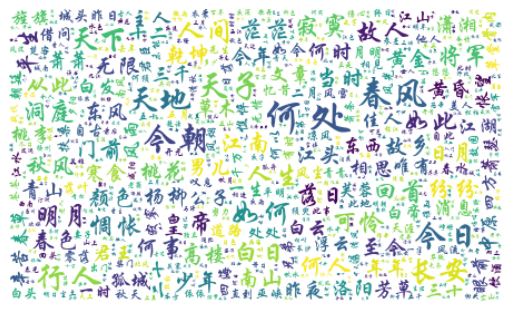


图4.18 安史之乱后词频词云图

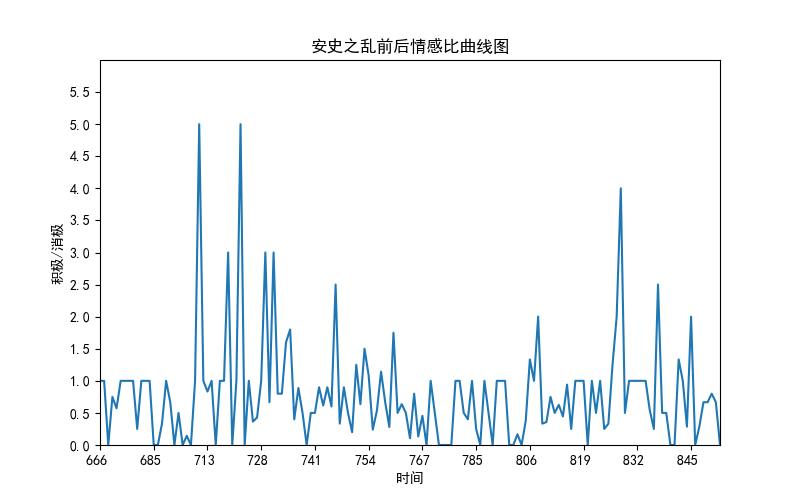


图4.19 情感判断结果

从词频图可以看出以往的“美人”，“黄金”在诗歌中已不服存在取而代之的是“惆怅”，“茫茫”，“相思”。从曲线图看，经历安史之乱后也就是755年后，诗歌流露的感情逐渐偏向消极。

# 5 总结与展望

5.1总结

本文利用TF-IDF模型与Word2vec 搭建了古诗词情感分析模型，基于已有的情感字典，根据语义关联以判别古诗的情感倾向，并用Python实现。根据实验结果，该模型能判别古诗的情感倾向，准确率为78%。对唐诗按照不同的时期进行分析的结果来看，初唐和盛唐诗歌较为积极向上而中唐后唐以后，诗人大多年老力衰或因战争背井离乡，诗歌中多为透露出对故人的思念或对自己壮志未酬的感慨，发出悲凉消极之情。对于山水田园诗，诗人生活或对山水田园的描写并不单是对大自然的热爱，对仕途的随心，也有抒发自己的生活孤独寂寞，思念好友之情。安史之乱前504首和安史之乱后769首古诗的情感分析来看，经历安史之乱后，唐朝诗人的消极率比安史之乱发生前高出约14%。

5.2不足

诗歌分词方法采用逐字分词方法方式，采用TF-IDF作为权重计算，能够适当的规避无意义字对情感判断的影响，但是在诗歌的情感判断中依然存在噪音，未能达到有效的特征选择。全唐诗约四万多首，但是能找到确切的创作时间仅有一千八百多，能用于实验的数据有待扩充。

5.3展望

利用机器对古诗进行性研究仍然值得探究的课题，往后的研究可以从下两点出发：

一、古诗的特征选择

古诗中的每一个字虽然都是经过诗人精心推敲的，但其中也一定有噪音。古诗的特征选择可以有效的降低诗歌中的对于情感判断无用的字，以提升古诗情感判断的准确度。

二、采用分类的算法对古诗进行情感判断

古诗的情感分类，或者说是文本的情感分类一般都是二元分类问题，因此可以引入贝叶斯公式，SVM等分类模型，对古诗进行情感倾向判断。

参考文献

[1] 聂长东. 试谈安史之乱对唐代诗歌的影响[J]. 甘肃农业, 2006, (05): 225.

[2] 欧广玮. 安史之乱前后唐代曲江诗对比研究[D]. 西南大学, 2011.

[3] 王志清. 安史之乱后唐诗的吴风日盛与气骨顿衰[J]. 南通大学学报(社会科学版), 2014, 30(01): 53-58.

[4] 刘佳利. 安史之乱对唐朝山水诗的影响及新题材的产生[J]. 辽宁教育行政学院学报, 2017, 34(06): 90-95.

[5] https://baike.baidu.com/item/%E6%96%87%E6%9C%AC%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90/19431243?fr=aladdin

[6] Nasukawa, T. and Yi, J. (2003) Sentiment Analysis: Capturing Favorability Using Natural Language Processing. Proceedings of the 2nd International Conference on Knowledge Capture, Florida, 23-25 October 2003, 70-77.http://dx.doi.org/10.1145/945645.945658

[7] 刘健. 基于近似文本分析的意见挖掘[D]. 上海大学, 2007.

[8] 宋光鹏. 文本的情感倾向分析研究[D]. 北京邮电大学, 2008.

[9] 陈晓东. 基于情感词典的中文微博情感倾向分析研究[D]. 华中科技大学, 2012.

[10] 杨鼎,阳爱民.一种基于情感词典和朴素贝叶斯的中文文本情感分类方法[J].计算机应用研究,2010,27(10):3737-3739+3743.

[11] 吴斌,吉佳,孟琳,石川,赵惠东,李仪清.基于迁移学习的唐诗宋词情感分析[J].电子学报,2016,44(11):2780-2787.

[12] 孙建旺, 吕学强, 张雷瀚. 基于词典与机器学习的中文微博情感分析研究[J]. 计算机应用与软件, 2014, 31(07): 177-181.

[13] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du, Analogical Reasoning on Chinese Morphological and Semantic Relations, ACL 2018.

[14] 张谦, 高章敏, 刘嘉勇. 基于Word2vec的微博短文本分类研究[J]. 信息网络安全, 2017, (01): 57-62.

[15] 杨志胜. 基于字根法的古汉语分词模型[D].天津大学,2007.

[16] Zhongguo Li, Maosong Sun. Punctuation as Implicit Annotations for Chinese Word Segmentation. Computational Linguistics, vol. 35, no. 4, pp. 505-512, 2009.

[17] http://www.keenage.com/html/c\_bulletin\_2007.htm

[18] https://baike.baidu.com/item/tf-idf/8816134?fr=aladdin

[19] https://github.com/hxgdzyuyi/tang\_poetry

[20] https://baike.baidu.com/item/山水田园诗派/293382?fr=aladdin

致 谢

在本文完成之际，谨向我的导师韩冬老师致以衷心的感谢，本论文是在他的精心指导和关怀下完成的，从论文的选题、方案设计，到论文的撰写和修改，都倾注了无数的心血和汗水，在学习期间，他的言传身教将使我终生受益，他认真严谨的治学态度、豁达宽广的胸怀、平易近人的处事风格是我一生的楷模，值此提交论文之时，在此向韩冬导师表达衷心的感谢！