四 川 大 学

硕士研究生课程考试试卷

姓　　名 王圳川 学　　号 2024223045126

学院（所、中心） 计算机 专　　业 电子信息

|  |  |
| --- | --- |
| 考试课程名称 数据分析与实践 | |
| 考试方式 □笔试 □口试 √撰写论文  任课教师 罗 川 | 考试成绩\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  考试时间 2024.12.18 |

四川大学研究生院制

**基于K-means和谱聚类的古诗情感分类标准挖掘**

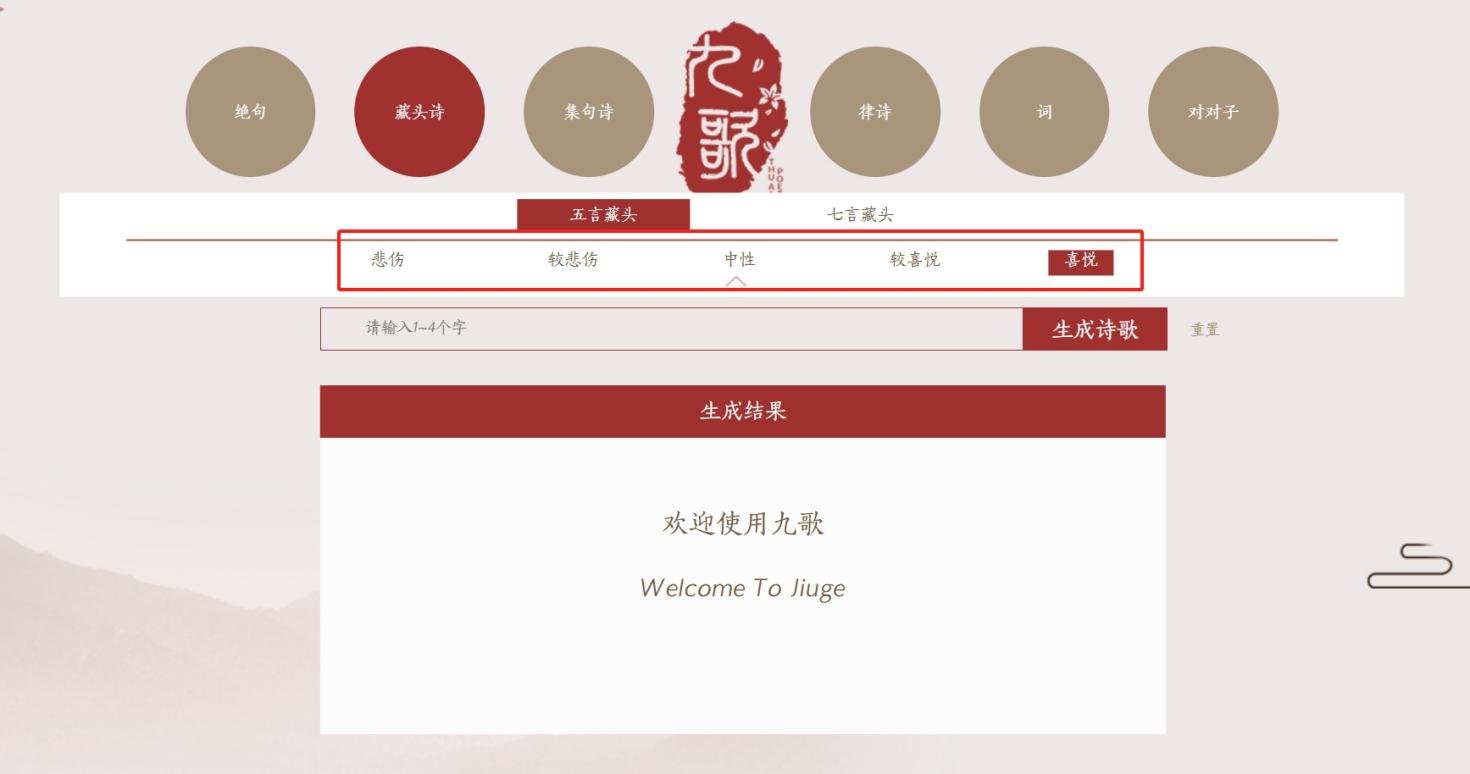
2024223045126 王圳川 计算机学院电子信息专业

**摘 要：**现有的古诗情感分析领域，通常简单的将古诗的情感类别分为悲伤，较悲伤，中性，较喜悦，喜悦等。然而古诗中的情感通常更为丰富，比如羁旅思乡和送别友人同样都是悲伤的情感，但却是截然不同的两种情绪。为了解决这个问题，首先需要确定一个合理的情感分类，因此，我打算使用聚类方法从广泛的古诗数据中挖掘出合理的情感分类。

**关键词：**聚类；诗歌生成；数据分析；

**1.引言**

诗歌是一种高度文学化、艺术化和凝练化的特殊形式。贯穿数千年的人类历史进程，无数诗人以诗歌描写佳景，记载事件，论述哲思，抒发情怀，留下了浩如烟海的诗篇。这些诗作记录了不同时代的民生百态，反映着诗人的心境情怀和伦理思辨，成为不同文化传承的脉络与载体，对人类社会的发展产生了深远影响[1]。特别是古诗的情感表达，研究古诗的情感表达对于理解古代文化、社会风貌以及人们的精神世界具有重要意义。通过分析古诗词中细腻而丰富的情感描写，我们不仅能感受到作者个人的心境变化与生活经历，还能从中窥见特定历史时期的社会背景、思想潮流及审美倾向。然而现在对于古诗中的情感表达的研究，将古诗中的情感简单的分为悲伤，较悲伤，中性，较喜悦，喜悦。这与实际古诗中的情感表达相去甚远，因此我想要借助聚类的方法从广泛的古诗数据中挖掘出合理的情感类别划分。



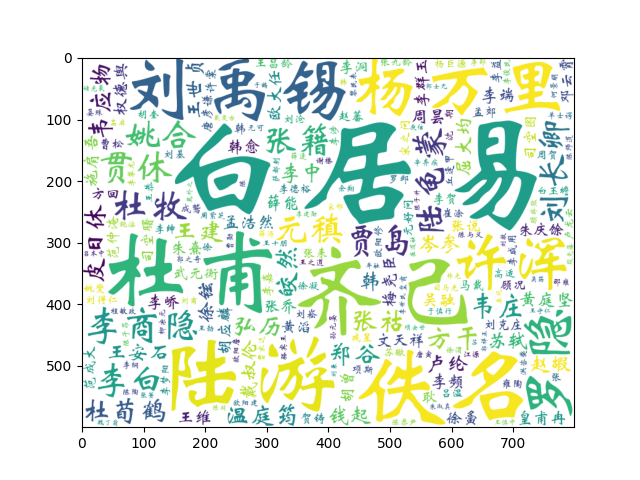
**2.方法**

为了简化对诗歌中情感的讨论，我们把情感限定到“词”的范围内，作出以下合理假设

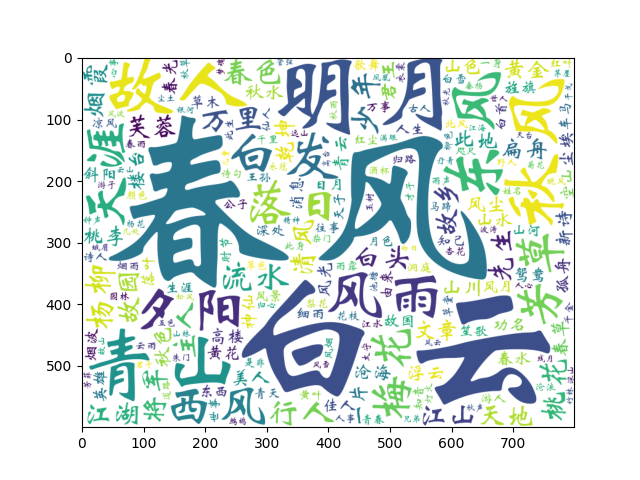
**假设：**古诗中的情感表达主要由意象（古诗中的非命名实体名词）决定

**2.1古诗意象提取**

首先，我使用了网络爬虫，爬取了[古诗文网](https://www.gushiwen.cn/)上的108,197首古诗。为了排除其中的词、曲等其他古诗文体裁，按照五言和七言即为每句诗中包含五个字和七个字，绝句为四句，律诗为八句的格式筛选出了易于分析的近体诗共55,433首。其中诗歌数量占比最多的三个朝代分别是唐代：27,183首，占比25.12%；宋代：11,696首，占比10.81%；明代：8,386首，占比7.75%。下图呈现了筛选出的近体诗中作者的云图。这里要说明的是，其他一些耳熟能详的诗人比如李白、王安石等不甚突出的原因是他们有许多诗并遵循近体诗的格式，所以这里并未统计完全。



其次，对这些近体诗进行分词和词性标注。由于现在大多数常用的开源的分词模型以及词性标注模型是在白话文语料上训练出来的，所以直接将其用在古诗的分词与词性标注上效果并不好，所以使用了jiayan（一个专注于古文上的分词以及词性标注的开源模型）进行上述任务。jiayan采用了863词性标注标准，这里为了简化讨论，按假设将词性为“n”的词语当做该古诗中的意象。下图展示了意象词云。



**2.2基于意象word embedding的k-means聚类**

一个最直接的想法就是直接使用意象的word embedding进行k-means聚类。为了取得最好的效果，所以没有使用Bert这种比较老的特征提取器，而是使用了Qwen的最后一个隐藏层作为的hidden states作为word embedding，然后分别使用了欧氏距离以及余弦相似度作为距离度量，簇的数目取3,6,9,12进行k-means聚类，并使用轮廓系数来评价聚类质量。最后将每个簇中最具有代表性的十个意象提取出来输入大语言模型使用四字词语总结其情感。实验结果放在第三部分。

**2.3基于意象共现频次的谱聚类**

一方面，由于word embedding实际上反映了这个词在大规模语料上的含义，导致实际上word embedding不能很好地表示该“意象”在诗歌中的含义。另外一方面，word embedding的相似度只能够说明这两个词在含义层面接近，并不能说明这两个词在情感表达层面接近，毕竟两个完全不同的意象也能够表达相似的情绪。基于这两点，我决定根据两个意象在同一首古诗中的共现频次作为其情感表达层面的相似度度量。所以我首先建立了意象-意象相似度矩阵，其中每一个具体的值为两个词在同一首诗中的共现频次，然后进行归一化并令对角线为1，代表词与词自己本身是最接近的。然后簇的数目取3,6,9,12进行谱聚类。实验结果放在第三部分。

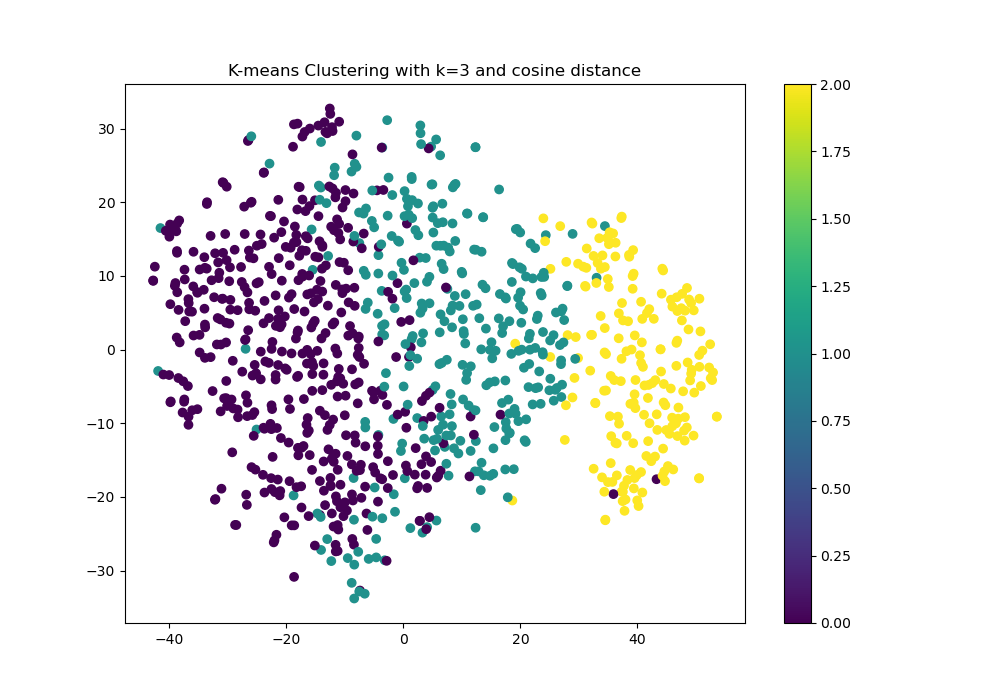
**3.实验**

**3.1 K-means聚类实验结果**

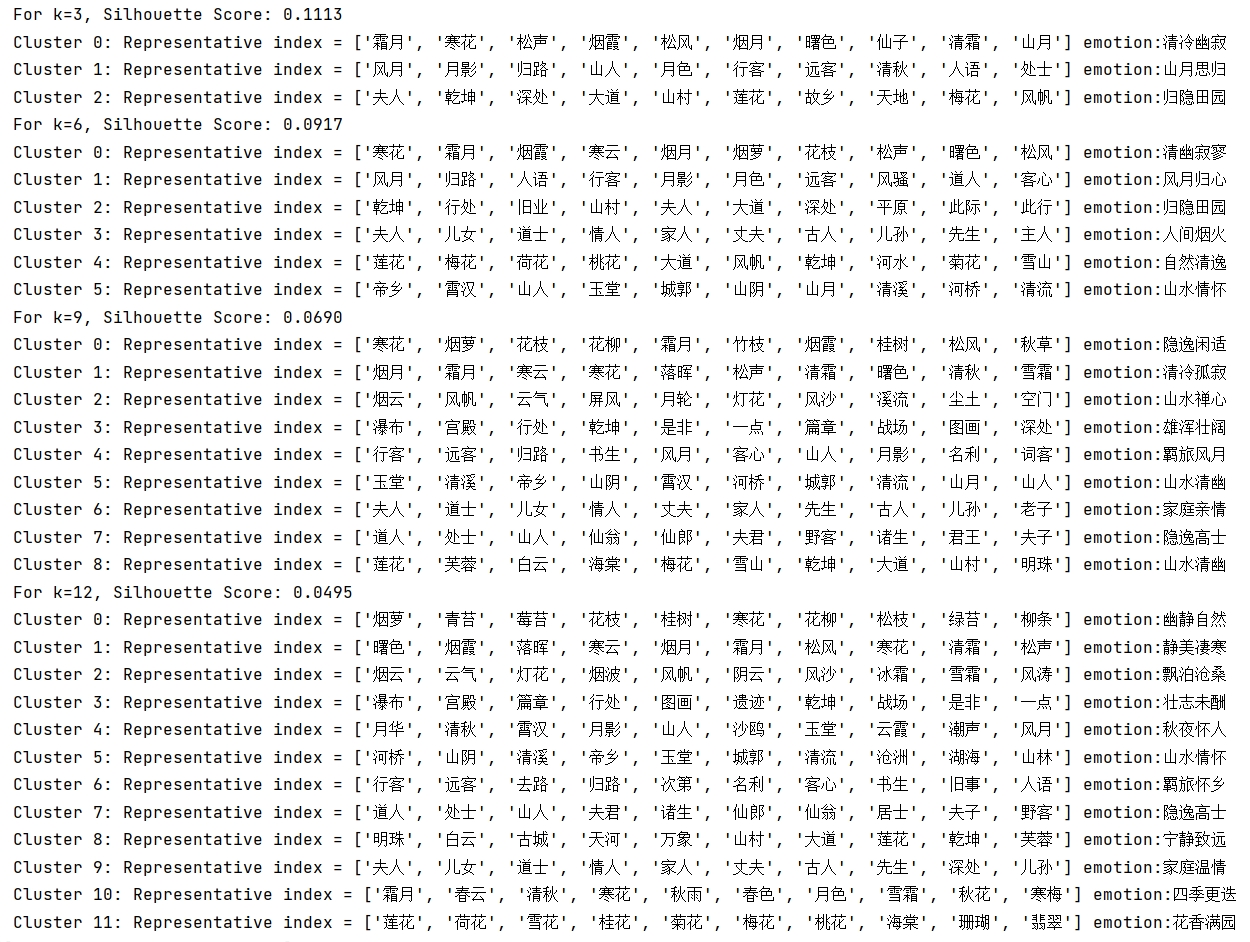
下表展示了在不同距离度量方式和选取不同簇的数目的情况下聚出的簇的轮廓系数（越接近1，效果越好）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K=3 | K=6 | K=9 | K=12 |
| euclidean | 0.0591 | 0.0392 | 0.0451 | 0.0397 |
| cosine | 0.1113 | 0.0917 | 0.0690 | 0.0495 |

相比之下余弦距离比欧氏距离的聚类效果相应要好一些，这可能是由于word embedding是高维向量，在高维空间之中，欧式距离的直观性和有效性会大大降低。但是总体来看效果也并不算太好，下图展示k=3，选取余弦相似度作为距离度量时的聚类效果可视化图。



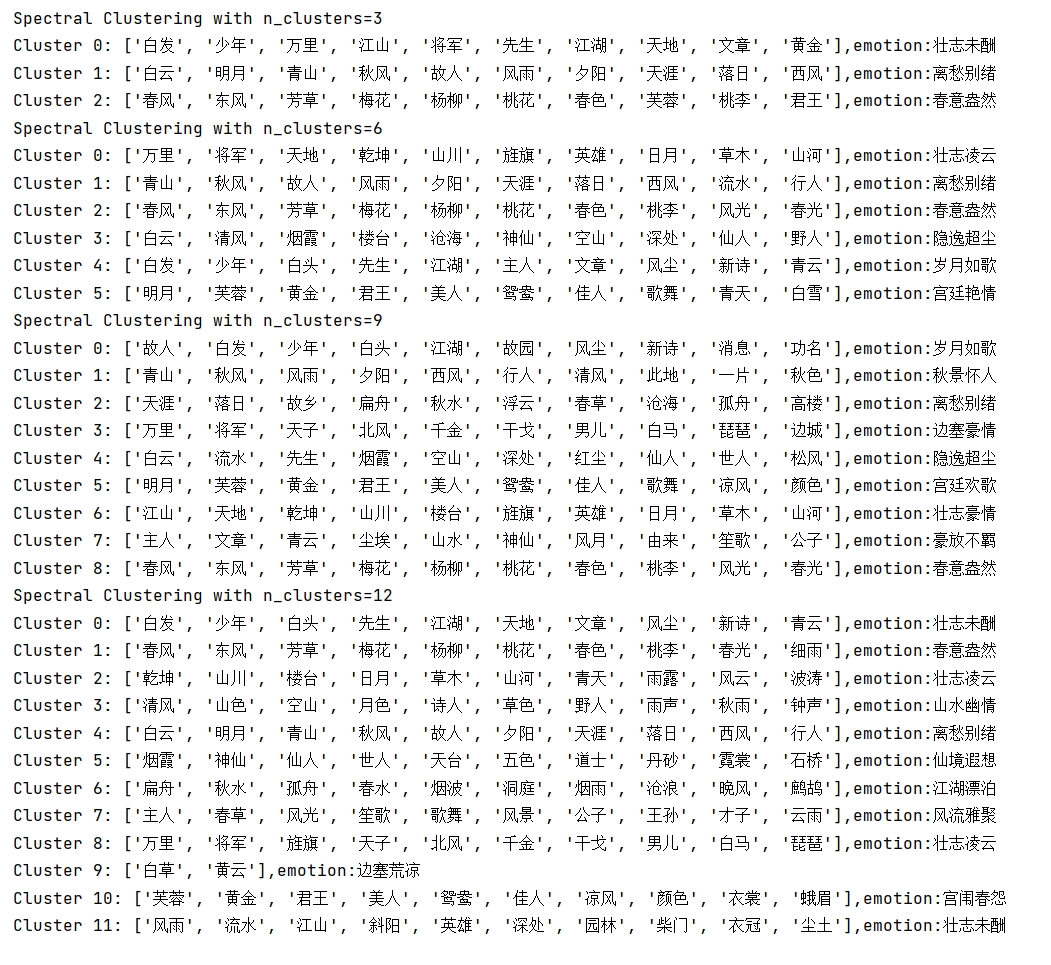
为了总结出这些意象簇所代表的情感，我使用了Qwen-max大语言模型，让其替我使用四字词语总结这些簇中最具有代表性的十个意象所代表的情感作为该簇所代表的情感，下图举出了选取余弦相似度为距离度量的几个实验结果。



可以看出虽然聚出来的簇有一定的可解释性，但是一方面，由于word embedding实际上反映了这个词在大规模语料上的含义，导致实际上word embedding不能很好地表示该“意象”在诗歌中的含义。另外一方面，word embedding的相似度只能够说明这两个词在含义层面接近，并不能说明这两个词在情感表达层面接近，毕竟两个完全不同的意象也能够表达相似的情绪。基于这两点来看，当前基于word embedding的k-means聚类方法还不算特别合理。

**3.2 谱聚类实验结果**

下图展示了谱聚类的聚类结果。



可以看出，相比于k-means聚类，谱聚类的效果要更加接近我想要的效果，所聚出来的簇中的意象并不是字面上接近，而的确是表达类似的情感。

**4.结语**

本文主要采用了数据分析与实践课程中学到的聚类方法，去探索了在广泛古诗数据中蕴含的情感分类信息。我觉得在这门课程中，我确实学到了如何采用各式各样的数据挖掘方法去数据中挖掘有价值的信息，在以后的科研生涯中，我想我也会经常利用到课程上学到的东西。而针对上述工作而言，我认为还可以结合意象本身的word embedding和共现关系两个数据进行分类，可惜的是暂时还没有想到很好的方法，这会是我未来努力的方向。

**参考文献**

矣晓沅. 具有文学表现力的中文古典诗歌自动写作方法研究[D].清华大学,2022.DOI:10.27266/d.cnki.gqhau.2021.000045.