Modern Linguistics 现代语言学, 2020, 8(2), 279-288
Published Online April 2020 in Hans. http://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2020.82038

Hans汉斯

Analysis of Natural Landscape Image of Song Ci Based on Python Text Mining Method

Chuan Chen, Ying Liu, Yu Lu, Yuan Lu, Huqin Yan

Xiamen National Accounting Institute, Xiamen Fujian
Email: 765759311@qq.com, liu18859241099@163.com, 495194130@qq.com, 496811642@qq.com, 2118562528@qq.com

Received: Mar. 24th, 2020; accepted: Apr. 21st, 2020; published: Apr. 28th, 2020

Abstract

As a brilliant pearl in the crown of ancient Chinese literature, Song Ci has a high degree of literary achievement and appreciation. This paper uses Python's BS4, word cloud and Jieba function to select and sort out nearly 10,000 poems of Song Dynasty, the key words with high frequency are picked up by using text mining technology. According to the statistics, 5 natural landscape words with high word frequency are selected for emotional analysis one by one, and the corresponding emotional index is obtained, and the corresponding visual semantic network graph is further constructed. Through the research and analysis of the natural landscape image and the emotion contained in Song Ci, the conclusion shows that the semantic of different natural landscape entries reflects different emotional factors.

Keywords

Song Ci, Text Mining, Natural Landscape, Emotional Analysis

Python文本挖掘方法辅助宋词自然景观意象分析

陈　钏，刘　瑛，卢　玉，路　媛，阎虎勤

厦门国家会计学院，福建　厦门
Email: 765759311@qq.com, liu18859241099@163.com, 495194130@qq.com, 496811642@qq.com, 2118562528@qq.com

收稿日期：2020年3月24日；录用日期：2020年4月21日；发布日期：2020年4月28日

摘　要

宋词作为中国古代文学皇冠上光辉夺目的明珠，其文学成就和欣赏价值极高。本文使用Python的BS4 +

文章引用: 陈钏, 刘瑛, 卢玉, 路媛, 阎虎勤. Python 文本挖掘方法辅助宋词自然景观意象分析[J]. 现代语言学, 2020, 8(2): 279-288. DOI: 10.12677/ml.2020.82038

Wordcloud + Jieba功能方法，对近10,000首宋代诗词进行筛选梳理，利用文本挖掘技术，抓取出现频率较高的关键词。根据统计，选取词频较高的5个自然景观词逐条进行情感分析，得出相应的情感指数，并进一步构建相应的可视化的语义网络图。通过对宋词的自然景观意象和蕴含的情感进行研究分析，结论表明，不同自然景观的词条语义体现出不同程度的情感因素。

关键词

宋词，文本挖掘，自然景观，情感分析

1. 引言

古诗词是中华传统文化的灿烂瑰宝，作为中国文学史上璀璨的明珠之一，宋词代表了宋代文学，并以其别具一格的魅力在中华大地广为流传[1]。宋词拥有韵律和谐、格式错落、情感丰富、形象多样、意境优美等特征。宋词涵盖的题材众多，其中与自然景观相关的占据了不可小觑的数量。古代诗词中多倡导托物寓意，从一景一物之中可以引申出更为丰富饱满的情感和精神寄托，我们常称之为意象。人文情感与自然景观的有机契合，赋予宋词更强的渲染力与表现力。基于对宋词的文本挖掘与意象分析，有利于提升自然景观的审美价值，同时加深我们对古诗词的鉴赏力。

由于宋词数量庞大，词汇丰富，传统的文本分析方法很难遍历所有宋词文本并开展词频统计分析。Python 语言提供了 Beautiful soup (BS4)数据爬虫、Word cloud 文本搜索、Jieba 汉字库分析等大数据文本搜索和分析工具，可以较为容易地解决大量文本下的词汇搜索、词频统计、语义分析等问题。

本文使用 Python 的 BS4 + Wordcloud + Jieba 功能方法，选取了 10,000 首宋词进行文本处理。通过提取其中与自然景观相对应的情感特征高频词汇，然后通过语义网络分析构建可视化视图，通过输出的主要的链接词，来分析景观特征与典型表述，得出对应的正负面情感指数。本文的研究方法迎合时代发展，为诗词的研究开创了新的思路。

2. 研究方法与框架

2.1. 诗词文本挖掘的可行性

吴潇[2]认为，诗词文本具有极高的文学与艺术价值，与其他历史文学文本相比，诗词具有样本量大、整体意象明确、词义高度精准、空间层次清晰、空间要素明确、时间延续性等特点和优势。宋词描述了大量古代的景观环境，更蕴含了许多文学情感意象。本文抓取的诗词涵括整个宋朝，样本量较多且具有代表性，得出的结论也较为可靠。根据提取的高频词加以分析，有助于加深对于诗人思想感情的理解，对于鉴赏古诗词具有重要意义。

2.2. 诗词文本选取方法

通过诗词文本研究某一时段的自然景观格局时，首先应筛选特定历史时期特征类似的诗词文本，选取原则：1) 该诗词文本具备时间与形式的一致性；2) 文本表述中应具备自然景观类型和尺度空间的丰富性；3) 文本表述对特定景观与情感特征认知具有对应性。本文选取的对象为宋代诗词，筛选之后，将所有诗词文本利

用文本挖掘工具逐条进行情感分析得出相应的情感指数，对诗词文本进行分词、词性分析、高频词统计、关联性分析与词义聚类分析，不同词性可用作不同感知层面的自然景观分析。由于诗词属于古汉语，不能仅按照普通的现代汉语进行分词处理，因此采取分词词条与单字词共同参考的方式，以全面真实地展现诗词中描绘自然景观。结合高频词与高频字，即可较全面客观地展现诗词中所描绘的自然景观与情感特征。

2.3. 诗词文本分析的自然景观特征

广义上的自然景观不仅是空间性的，也是时间性的。有时候是静态的，也常有动态的。例如地形景观、地质景观、森林景观、天文景观、气候景观、生物景观等。从“春风又绿江南岸，明月何时照我还”到“明月别枝惊鹊，清风半夜鸣蝉”，从“风流总被雨打风吹去”，到“杨柳岸，晓风残月”，古代诗词中往往有许多描写自然景观的诗句，诗人不仅生动形象地描绘了景物本身，更体现着喜怒哀乐，寄托自己的情感。

以宋词为例，诗词文本对自然景观的描述特征十分鲜明。从微观角度看，宋代诗词往往通过对具体的景观要素的表述来体现情感意象，如“明月”、“梅花”、“芳草”“杨柳”等。同时也有动静结合的景观要素，如“风吹”、“流水”、“风雨”“东风”，以及季节与时间特征，如“春色”、“秋色”。此外，还聚焦于空间场景及地域，如“江南”与“西湖”、“江山”等(如表 1)。

Table 1. Characteristics and expression of ancient poetry text
表 1. 古诗词文本的特征与表述

特征	典型表述
景物与细节	明月、梅花、芳草、杨柳
动态与静态	风吹、流水、风雨、东风
季节与时间	春色、秋风、黄昏、春风
场景与地域	江南、西湖、江山、天地

2.4. 诗词文本分析技术

2.4.1. Jieba 分词技术

在人工智能全球化的信息时代，要想更好地完成自然语言处理，分词是首要任务[3]。本文使用的分词技术是 Python 的 Jieba 分词器。其分词算法总体是：使用基于前缀词典的词图扫描，生成所有可能生成词所构成的有向无环图，再采用动态规划查找最大概率路径，找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的 HMM 模型(使用 Viterbi 算法)来预测分词[4]。

Jieba 分词主要有三种分词模式：1) 精确模式，可以把文本精确的切分，不存在冗余单词，最适合用于文本分析；2) 全模式，可以把文本中所有可能的词语都扫描出来，但存在冗余，不能解决歧义问题；3) 搜索引擎模式，在精确模式基础上对长词语进行再次切分，适合用于搜索引擎分词[5]。表 2 是三种分词模式的实验结果，以“明月几时有，把酒问青天。”(选自苏轼《水调歌头》)为例，三种分词模式的分词结果分别为：

Table 2. Experimental results of three word segmentation modes of Jieba
表 2. jieba 三种分词模式的实验结果

模式		结果
精确模式	代码	jieba1 = jieba.lcut (“明月几时有，把酒问青天。”)
	分词结果	[“明月”，“几时”，“有”，“把酒”，“问青天”]
全模式	代码	jieba2 = jieba.lcut (“明月几时有，把酒问青天。”，cut_all = True)
	分词结果	[“明月”，“几时”，“有”，“把酒”，“问青天”，“青天”]

Continued

搜索引擎模式	代码	jieba3 = jieba.lcut_for_search (“明月几时有，把酒问青天。”)
	分词结果	[“明月”，“几时”，“有”，“把酒”，“青天”，“问青天”]

基于本文的研究目的，本文选取的分词模式为精确模式。

2.4.2. 词云技术

词云是对文本中出现频率较高的“关键词”予以视觉化的展现[6]。在词云图中，不同的单词以不同的颜色表示，不同词频的单词以不同字号表示，使得阅读者只通过词云图，就可以大致的了解长文本中的核心内容，是文本数据可视化的一种常用方式。

本文采用 Python 的第三方库 wordcloud 库来实现词云图的绘制。

2.4.3. 语义网络分析技术与 ROST 软件

语义网络是一种通过使用图形符号来表示概念的知识或底层结构的方法，谭荧[7]认为，在表达知识具有概念化和语境化的优势。概念化是指语义网络可以通过生成一个认知概念的地图，赋予概念意义，并通过关联最终理解每个概念；语境化是指语义网络的图结构可以更好地表示自然语言的情境，从而更好地提取自然语言的语义。

ROST CM6 是武汉大学沈阳教授研发编码的，基于内容挖掘，辅助人文社会科学数字化研究的大型免费社会计算平台，可用于聊天分析、全网分析、浏览分析、微博分析、期刊分析、情感分析、语义网络分析等多项强大功能。

本文运用 ROST CM6 软件及其自带的 Net draw 插件生成语义网络图进行语义网络分析，通过语义网络所展示出来的概念关联与语境关系，分析不同的自然景观在宋词中具有的意象。

3. 以宋代诗词为样本的自然景观分析

3.1. 获取数据及数据预处理

本文以宋朝的诗文作为研究对象，文本内容来源于新派查询网下的诗词大全网址，搜索宋朝即可看到宋代所有的诗文。本文使用 python 软件抓取宋朝的诗文语句进行处理。根据网站显示，宋代诗词一共有 20 万余首，本文选取网站前 10,000 首宋代诗词作为数据样本，再使用 jieba 软件包对诗词进行整理，利用停词表将其中无意义的词和标点符号去除。例如，李清照的诗句“常记溪亭日暮，沉醉不知归路，兴尽晚回舟，误入藕花深处”，经过软件处理分词之后，就剩下“溪亭”、“日暮”、“沉醉”、“归路”、“兴尽”、“晚回”、“舟”、“误入”、“藕花”、“深处”，再经过停词表剔除无意义的词和单个的词，“舟”字就被剔除了。经过分词处理删去部分词之后，总共得到 65,532 个词。本文利用 python 软件绘制宋代诗词的词云图，可以看出宋词中运用较多的词语。从词云图上可以看到，“江南”、“归来”、“东风”、“春风”等词出现的频数较多，如图 1 (图 1 右侧为依照北宋地图板块做的宋代诗文词云图)。

Figure 1. Cloud map of poems in song dynasty
图 1. 宋代诗词词云图

根据诗句分词后的词语和在所有诗句中出现的频数，本文发现宋词提及的自然景观多为“东风”、“春风”、“明月”、“梅花”和“芳草”等，如“东风”一词在整个诗文里出现 299 次，而“春风”一词出现 275 次。同时，我们也发现宋词提及较多的情感词为“相思”、“憔悴”、“寂寞”等负面情感词，如“相思”一词在整个诗文里出现 139 次，而“凄凉”一词出现 117 次，基于词频统计结果，本文绘制部分词及频数表见表 3。

Table 3. Part of words and frequency
表 3. 部分词及频数表

词序	词	频数	词序	词	频数	词序	词	频数	词序	词	频数
1	不知	301	16	无人	181	31	流水	129	46	天下	112
2	东风	299	17	当年	173	32	秋风	127	47	今年	112
3	人间	294	18	归去	172	33	白云	127	48	不可	112
4	春风	275	19	相逢	170	34	斜阳	125	49	春色	112
5	万里	263	20	梅花	169	35	扁舟	120	50	无限	111
6	江南	236	21	故人	157	36	少年	119	51	桃花	111
7	千里	234	22	当时	156	37	今日	119	52	天地	111
8	归来	233	23	青山	154	38	神仙	118	53	江山	111
9	风雨	227	24	如今	142	39	凄凉	117	54	去年	109
10	西风	213	25	黄昏	140	40	悠悠	117	55	百年	109
11	不见	203	26	相思	139	41	人生	116	56	行人	108
12	平生	199	27	天涯	139	42	何事	115	57	何人	108
13	风流	199	28	芳草	134	43	一笑	114	58	深处	107
14	明月	189	29	西湖	132	44	风月	114	59	一片	107
15	回首	181	30	阑干	130	45	十年	113	60	寂寞	107

3.2. 自然景观意象分析

通过对文本的词频统计，本文选取了词频较高的 5 个自然景观词进行意象的研究。首先，使用 EXCEL 对全文本进行筛选，提取出含有自然景观关键词的宋词语句。然后，将文本导入 ROST CM6 软件进行情感分析，利用 Python 提取高词频，并利用 ROST CM6 中加载项工具 Net Draw，构建出可视化的语义网络图。

Table 4. Results of emotional analysis
表 4. 情感分析结果

自然景观	正面	中性	负面
春风	43.62%	14.54%	41.84%
东风	39.30%	12.46%	48.24%
西风	42.73%	11.36%	45.91%
芳草	31.85%	19.26%	48.89%
流水	40.61%	13.33%	46.06%

Table 5. High frequency words related to natural landscape
表 5. 与自然景观相关的高频词

次序	春风		东风		西风		芳草		流水	
	高频词	词频	高频词	词频	高频词	词频	高频词	词频	高频词	词频
1	归来	13	阑干	19	人间	14	东风	15	青山	12
2	桃李	11	江南	17	万里	14	阑干	11	斜阳	10
3	杨柳	11	杨柳	16	昨夜	12	天涯	10	白云	10
4	十里	11	芳草	15	故人	12	春风	9	芳草	9
5	芳草	10	桃花	14	秋色	11	落花	9	千里	9
6	风流	10	垂杨	13	明月	10	斜阳	9	东风	8
7	风雨	9	人间	13	芙蓉	10	绿杨	8	落花	8
8	而今	9	回首	12	斜阳	10	相思	8	回首	8
9	江南	9	桃李	12	梧桐	9	依旧	7	人间	7
10	平生	8	清明	12	回首	9	池塘	7	相逢	7

3.2.1. 春风

根据情感分析的结果(表 4)，与“春风”一词相关的宋词的正面情感指数与负面情感指数相近。根据高频词表(表 5)与语义网络分析的结果(图 2)，“春风”常与“人生”、“平生”、“人间”、“千古”等词同时出现，可见“春风”一词常在一些人生思考主题的宋词中出现。此外，“春风”与“何时”、“相逢”、“归来”、“相思”等词的关联较大，说明“春风”一词常承载着诗人的思乡之情与离别之苦。同时，“春风”作为春天的象征之一，常与“桃花“、”杨柳“、”芳草“此类描绘春色的词语一同出现。

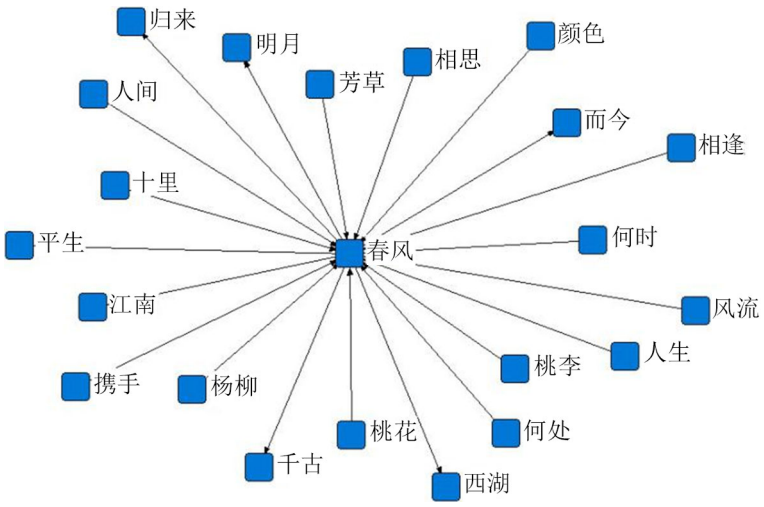

Figure 2. Semantic network analysis of “chunfeng”
图 2. “春风”的语义网络分析图

3.2.2. 东风

根据情感分析的结果，与“东风”一词相关的宋词的负面情感指数较高，说明“东风”在宋词中是比较消极的一个意象。根据高频词表与语义网络分析的结果(见图 3)，“东风”常与“杨柳”、“春色”、

“桃花”、“梨花”、“芳草”、“清明”、“一枝”、“春色”等词同时出现，可见“东风”一词常见于描写春色的宋词之中，“东风”也有春风之意。

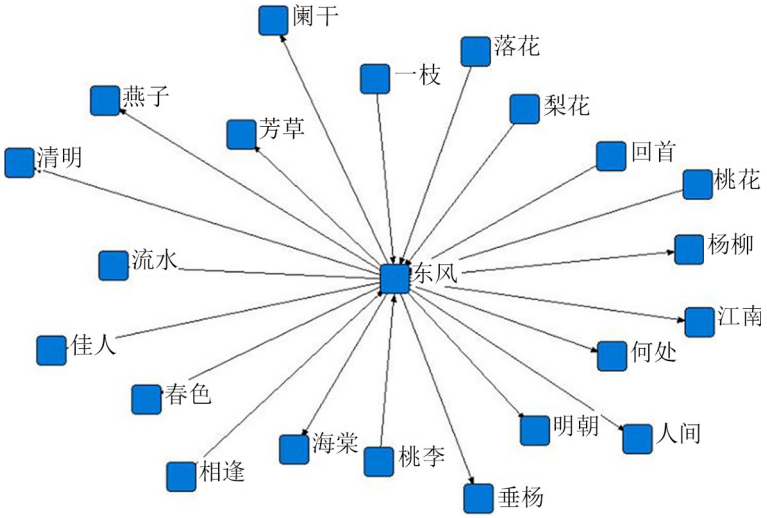

Figure 3. Semantic network analysis of “dongfeng”
图 3. “东风”的语义网络分析图

3.2.3. 西风

根据情感分析的结果，与“西风”相关的宋词的负面情感指数略高于正面情感指数。根据高频词表与语义网络分析结果(见图 4)，“西风”与“重阳”、“黄花”、“秋色”、“芙蓉”、“梧桐”等与秋日景色相关的词语关联性较强，说明“西风”含有秋风之意。同时，“西风”常与“扁舟”、“千里”、“万里”等行程相关的词语，以及“故人”、“明月”等词同时出现，可见“西风”也承载着诗人的离别之情。

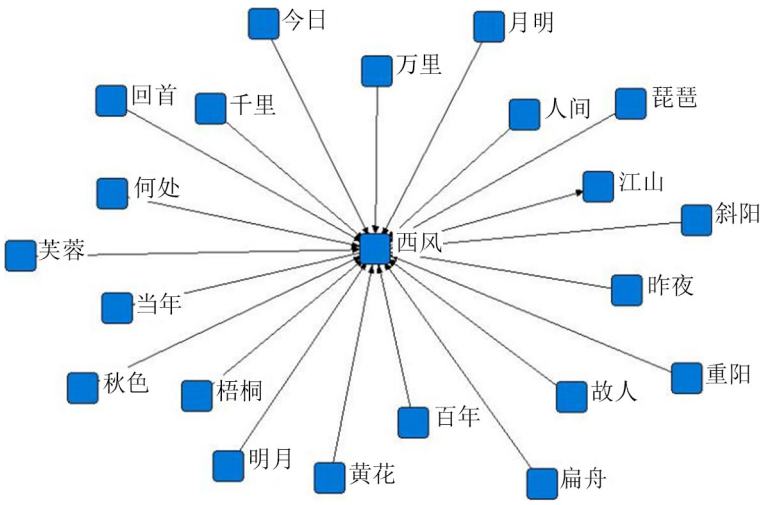

Figure 4. Semantic network analysis of “Xifeng”
图 4. “西风”的语义网络分析图

3.2.4. 芳草

根据情感分析的结果，与“芳草”一词相关的宋词的负面情感指数较高，说明“芳草”在是比较消极的一个意象。根据高频词表与语义网络分析结果(见图 5)，“芳草”与“梨花”、“桃花”、“绿杨”、

“春风”、“东风”等描绘春天的词语关联性较强，可见“芳草”也常用于描绘春色。同时，“芳草”一次常与“落花”、“斜阳”、“黄昏”、“相思”、“无情”等蕴含悲伤情绪的词语同时出现，说明“芳草”常承载诗人的愁思，与情感分析的结果一致。

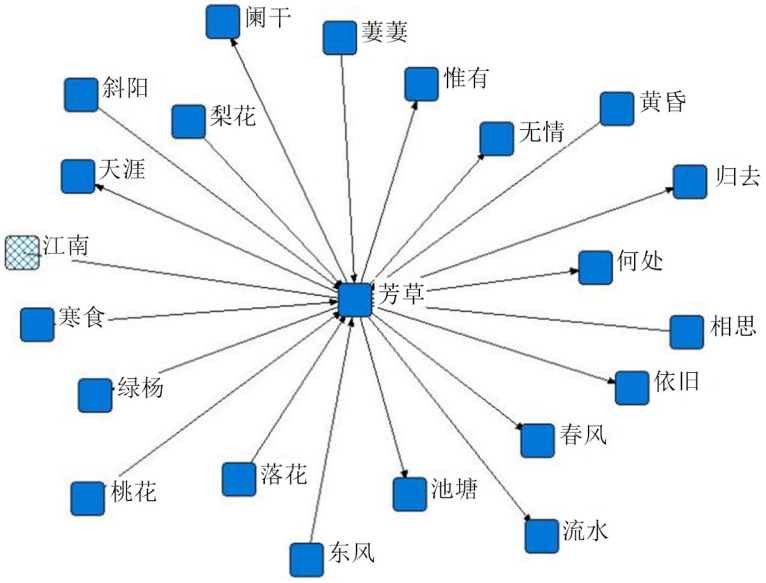

Figure 5. Semantic network analysis of “Fangcao”
图 5. “芳草”的语义网络分析图

3.2.5. 流水

根据情感分析的结果，与“流水”一词相关的宋词的负面情感指数较高，说明“流水”一词在是比较具有负面情绪的一个意象。根据高频词表与语义网络分析的结果(见图 6)，“流水”常与“芳草”、“春风”、“桃花”、“东风”等词同时出现，可见“流水”一词常见于描绘春天景色的宋词之中。同时，“流水”一词于“相逢”、“相思”、“明月”等含有相思之情的词关联性较强，说明“流水”也寄托了诗人的相思之情，与情感分析的结果一致。

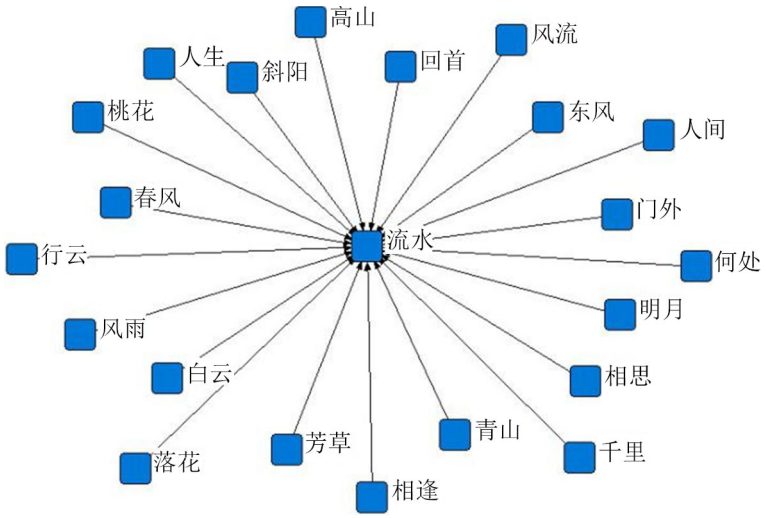

Figure 6. Semantic network analysis of “flowing water”
图 6. “流水”的语义网络分析图

4. 研究结论

通过对宋词中自然景观的描写所体现出来的情感进行研究分析，不仅使得我们更加贴近宋代时期的文化氛围，利于对宋词更清晰和准确的解读，同时在一定程度上也吸引和督促当代人对历史文化瑰宝的珍惜和关注。本文在对宋词进行文本挖掘的基础上，对自然景观的特征以及词人借此而抒发的情感加以分析，成效可观。围绕研究成果，现得出结论如下：

1) “春风”一词，本就给人温暖、融和、舒适的感觉。宋词中的春风多象征着对美好未来的期盼以及词人心中难掩的欣喜之情。在如今的生活中，我们也经常听到“如沐春风”，“春风得意”，“春风满面”等成语的使用。不仅如此，“春风”一词也常出现在事态发展顺利的时候，如“改革春风吹满地”不仅代表了改革开放的顺利推进，借此也体现了人心所向。

2) 根据浙江大学 CAD & CG 国家重点实验室的大数据分析[8]，“东风”一词在宋词中总共出现了1264 次。在过去，“东风”既可是早春的风，象征着希望。多指词人想要乘借东风来实现自己的远大抱负。但与此同时，“东风”也可是晚春那吹落百花的风，暗指词人遇人不淑又或是仕途跌宕的不幸。当代对于“东风”一词的处理，更多是积极向上，假借东风之意。就如我们都知道如今有一著名汽车品牌就叫作“东风”。

3) “西风”一词在古诗词中皆是悲伤，萧瑟的意思。我们最为熟悉的一句词：古道西风瘦马。夕阳西下，断肠人在天涯。“西风”与“瘦马”相对应，并以“断肠人”结尾，无不体现出词人内心的绝望和身心的孤独疲惫。在现代我们对“西风”一词已经很少使用，但近代“西风”曾被毛主席赋予不同的含义。毛主席曾经引用《红楼梦》中的话“不是西风压倒东风，而是东风压倒西风”以此一句来表达对战争胜利的决心和中华民族的尊严与自信。

4) “芳草”一词在宋词中存在两种含义，既是代表生机和春天的希望，但当与落花等结合时又是寂寥，落寞之意。如今我们对“芳草”的使用少之又少，更多的时候它就是简单地指代绿草而已。后代对于“芳草”赋予的新含义在于它可以代表具有高尚品德和美德的个人。

5) “流水”一词在古代多指时光匆匆流逝和借用“流水”的不断来寄托词人的愁绪之多，烦恼无休无止。就如苏轼在“乌台诗案”之后常使用“流水”来表达中年无奈，回归自然本真。现如今我们对“流水”一词的使用可谓是与过去有着较大的区别。“流水”现在的意思可以是单指现金流或者一项工作的某个过程，也可以通过“落花有意，流水无情”来表示感情方面的不顺等。

5. 结束语

众所周知，唐诗、宋词、元曲在中国古代中有着非凡的历史地位。而这当中，宋词与元曲争奇斗艳，与唐诗不相上下，齐头并进。张剑[9]曾说，唐诗宋词是中国古代文学乃至中国传统文化中最具美感和情感性的精品。宋词所代表的宋代文学，是中国文学史上璀璨的明珠之一。吴琼[10]认为，宋词蕴含着古典化的审美、情感和精神，从中焕发出来的诗性、理性与人性历久而弥新。宋词的词句，如今诵读起来，也仍然觉得美妙，不仅陶冶了凡世人们的情操，同时也是一种艺术享受。我们在体会宋词独特魅力的同时，也感受她的千姿百态，娇艳妩媚。宋词通过它的浅唱低吟，给我们带来了那个时代独有的韵味。

不同与以往的鉴赏性研究，本文的创新之处在于采用了 Python 的 BS4 + Wordcloud + Jieba 等先进的大数据处理工具，抓取出现频率较高的关键词进行相关词的寻找和探究，将文字出现的频率构成可视数据，通过对文本挖掘，客观地分析了宋词的自然景观意象及其蕴含的情感，得出相同词语在宋词中所代表的特殊含义和词人独树一帜的写作风格。这不仅对宋词研究具有特殊意义，也对当代人写作的语言、题材和风格上有着不小的指导意义。更加重要的是，提醒世人应爱惜和保护宋词这一历史珍品。

基金项目

本论文得到了厦门国家会计学院2019年“云顶课题：Python财务数据分析”项目的支持。

参考文献

[1] 杨萍. 宋词教学中学生审美能力培养研究[D]: [硕士学位论文]. 西安: 陕西师范大学, 2019.

[2] 吴潇, 李鑫, 赵炜. 基于唐诗文本挖掘的关中地区人文景观格局研究[J]. 风景园林, 2019, 26(12): 52-57.

[3] 李文江. 基于深度学习的商品评价数据分析系统[D]: [硕士学位论文]. 大连: 大连海事大学, 2018: 56.

[4] 严明, 郑昌兴. Python环境下的文本分词与词云制作[J]. 现代计算机(专业版), 2018(34): 86-89.

[5] 望江龙, 王晓红. 基于Python爬虫技术实现[J]. 电脑编程技巧与维护, 2019(9): 18-20+41.

[6] 徐博龙. 应用Jieba和Wordcloud库的词云设计与优化[J]. 福建电脑, 2019, 35(6): 25-28.

[7] 谭荧, 张进, 夏立新. 语义网络发展历程与现状研究[J]. 图书情报知识, 2019(6): 102-110.

[8] 张玮, 谭思危, 刘凯, 石磊, 陈思明, 陈为. 宋词研究的新视角: 文本关联与时空可视分析[J]. 计算机辅助设计与图形学学报, 2019, 31(10): 1687-1697.

[9] 张剑. 唐诗宋词研究专题[J]. 华南师范大学学报(社会科学版), 2018(2): 33.

[10] 吴琼. 近五年来宋词研究的进展及展望[J]. 湖州师范学院学报, 2019, 41(3): 62-70+103.