



北京大学

数据挖掘期末报告

题目： 历代古诗词风格对比分析

成员： 徐嘉辰、钟辰丽、王韵、柳俊志、王胜广

2019 年 06 月

目 录

1 序论.....	1
1.1 研究内容	1
1.2 评价指标	1
2 数据分析.....	2
2.1 数据源	2
2.2 数据解读	2
2.3 数据量	4
3 用词特点分析	5
3.1 全局高频字分析.....	5
3.2 特有高频词分析.....	9
3.3 TF-IDF 诗词关键词分析	13
3.4 常用双字词发现.....	16
4 情感特点分析	21
4.1 Word2Vec 词向量分析	21
4.2 探究诗词中的七情.....	22
4.2.1 CNN pair BiLSTM	23
4.2.2 各朝代诗词的情感倾向.....	25
5 结果分析.....	28
参考文献.....	29
附录 A.....	30

1 序论

1.1 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。不同朝代情感基调和诗词情绪表达倾向是否有差异。

1.2 评价指标

主要采取人工评价，是否将个朝代按时间分开或符合当前的认知，能够体现诗词的时代差异、反映其时代特征即可。

2 数据分析

2.1 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。原始数据格式为 json 和 csv，数据属性包括题目，朝代，作者，内容等。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

2.2 数据解读

原始数据朝代划分较为详细，包含众多唐末宋初，宋末元初，元末明初的划分。朝代更替时期的诗词统一划分到前一个朝代，例：元末明初转划分为元朝处理。进行繁简转换和去重后约 100 万首。以下是对诗词的详细解读。

经处理后诗词格式如下：

	题目	朝代	作者	内容
58259	忆秦娥	唐	冯延巳	风淅淅。夜雨连云黑。滴滴。窗外芭蕉灯下客。除非魂梦到乡国。免被关山隔。忆忆。一句枕前争忘得。
58260	送兄	唐	七岁女子	别路云初起，离亭叶正飞。所嗟人异雁，不作一行归。
58261	再赠	唐	上元夫人	弄玉有夫皆得道，刘纲兼室尽登仙。君能仔细窥朝露，须逐云车拜洞天。
58262	留别	唐	上元夫人	萧郎不顾凤楼人，云涩回车泪脸新。愁想蓬瀛归去路，难窥旧苑碧桃春。
58263	赠封陟	唐	上元夫人	谪居蓬岛别瑶池，春媚烟花有所思。为爱君心能洁白，愿操箕帚奉屏帏。
58264	句	唐	久则	湖上青山今欲买，白云无主问何人。

100 万诗词中，包括“句”这一种较为少见的题材 4968 首，例：“唐·孟浩然。

```
data_df[data_df["题目"]=="句"]
```

题目	朝代	作者	内容
5478	句	元 顺帝	鸟啼红树里，人在翠微中。
7365	句	元 耶律季天	梦蝶岂知真是蝶，骑牛何必更寻牛。
13338	句	元 刘祁	玄猿哭处江天暮，白雁来时泽国秋。
13339	句	元 杨鹏	树古叶黄早，僧閒头白迟。
18319	句	元 吴镇	我亦有亭深竹里，也思归去听秋声。
31719	句	元 商左山	药裹封灾随腊去，酒杯称寿逐年新。
31737	句	元 徐琬	一窍鬼工开混沌，八蚕神茧堕扶桑。

句，微云淡河汉，疏雨滴梧桐。逐逐怀良驭，萧萧顾乐鸣。”可以认为“句”是未能成诗的诗，半成品诗词，诗词的一部分。其数据如下所示：

各朝代诗词数量总体分布

朝代	诗词数	诗人数	高产诗人 诗词数
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	1037791	35227	42805
----	---------	-------	-------

唐宋元明清和近代的诗词数量较多，原因可能是诗歌从唐朝开始兴起。加上年代越近，相关诗词越容易被保留。以李白诗词为例，李白在世时已名满天下，他临终前把毕生作品的手稿，交给了叔叔李冰阳，整理成了 10 卷的《草堂集》，却全部失传。杜甫的诗词大多是四十岁之后写的，所以四十岁之前没有流传下来，且杜甫在世是并不出名，可能散佚者更多，考虑到陆游等人数万首的水平，杜甫所写诗词，应该在相当量级。写下《春江花月夜》，“孤篇压全唐”的张若虚，当时就被人尊称为“吴中四士”，竟只有一首诗词流传。可以预见，朝代越远，流失越多，众多精品诗词散佚，对诗词分析有一定影响，后期处理时，可考虑诗词按朝代远近以一定比例采样。

原始数据所分朝代十分详细，预处理时对其进行了合并。也造成了一些问题，比如李世明，在原始数据属于隋末唐初，合并到隋朝后竟然称为隋朝诗词最为高产的诗人。考虑到时代渐变性，这种合并对诗词的时代特点应该不会造成太大影响。例如隋唐文化本就十分相近。

2.3 数据量

去重后约 100 万首。各朝代具体诗词数据量于 2.2 数据分析中已详细列出。

3 用词特点分析

3.1 全局高频字分析

统计全局高频字，首先需要按朝代取出诗词内容，划分为以字为单位的数据，然后处理为字和频数的键值对。






但是有许多字的价值很低，比如“之”、“乎”、“者”、“也”这样的字。于是我们将类似的语气词和一些虚词划到停用词中。


实现和处理过程见附件的代码。

去除停用词后的全局高频字及各个朝代的高频字如下:

朝代	字	字频	词云
全局	人	631986	
	山	478570	
	风	412634	
	天	401544	
	日	374115	
先秦	子	1682	
	我	1477	
	人	1301	
	曰	1072	
	君	1012	
汉	人	1020	
	子	963	
	王	791	
	天	706	
	下	689	
魏晋	我	1674	

	人	1592	
	天	1312	
	风	1225	
	子	1061	
隋	天	585	
	风	526	
	人	449	
	云	413	
	道	406	
唐	人	67801	
	山	50583	
	日	46971	
	天	39213	
	花	35021	
五代	风	57	
	花	57	
	春	57	
	香	46	
	人	45	
宋	人	277848	
	山	205316	
	天	165638	
	日	153756	
	如	140907	

辽	吟	13	
	君	11	
	待	10	
	当	7	
	深	7	
金	人	3585	
	山	3078	
	风	3012	
	天	1958	
	云	1842	
元	人	50180	
	我	44808	
	云	31856	
	风	31682	
	山	29776	
明	人	123423	
	风	110781	
	山	110557	
	云	94699	
	日	89748	
清	人	52893	
	风	46686	
	山	40951	
	花	36386	
	天	36310	
民国	人	10854	

	风	9209	
	花	8238	
	天	7479	
	山	7244	
近代	人	16821	
	风	16043	
	山	12951	
	天	12917	
	花	12011	
当代	人	15306	
	风	14213	
	天	11386	
	山	10326	
	花	9667	

实验结果比较合乎认知。

从全局的结果来看，“人”出现最多，这体现了《说文解字》里所讲的“人，天地之性最贵者也”，说明唐诗很好的秉承了“以人为本”的中华文化。而后续的“山”“风”“月”“日”“天”“云”“春”等都是在写景的诗句里经常出现的意象。这与古诗词大多借景抒情，寓情于景的方式相符合。

先秦和汉的高频词比较相近，格局都很大，大多都是天下、君王之类的词，这与当时天下的战乱密不可分，很多文章都以赋的形式写出，本身就格局宏大，也有很多文章是为君王上谏，这也是原因之一。

魏晋到隋朝景色相关的字渐渐多了起来，直到唐朝达到顶峰，这一特点持续到宋朝。到了辽，就多了一分悲凉的色彩。之后的朝代又都转向了写人写景。




3.2 特有高频词分析

通过上面高频词的统计，我们可以大体看到各个朝代的用词特点。但是还是有许多朝代的高频词十分相近，虽然确实出现了很多次，但是并不能充分体现朝代的特色。

可以观察到有一些字如“人”，许多朝代出现的次数都很多，但是没有代表性。于是我们利用字的总频数-非本朝的频数且大于阈值的字作为当前朝代的代表字。

实现和处理过程见附件的代码。

分析的结果如下图所示：

朝代	独有高频字	字频	词云
先秦	曰	1072	
	王	979	
	矣	678	
	彼	652	
	既	605	
汉	王	791	
	曰	632	
	故	430	
	夫	392	
	余（余）	380	
魏晋	言	829	
	神	646	
	德	628	
	华	599	
	乐	591	
隋	神	335	

	笑	802	
	醉	721	
	愁	621	
	阳	584	
元	你	28566	
	了	21952	
	儿	20568	
	这	18993	
	他	17055	
明	客	38662	
	海	36812	
	开	33223	
	西	33182	
	石	33073	
清	红	15857	
	愁	14571	
	难	13818	
	小	13342	
	情	13135	
民国	愁	3305	
	旧	2974	
	楼	2561	
	又	2511	
	到	2500	
近代	愁	5245	

	旧	4506	
	楼	4496	
	泪	3828	
	又	3780	
当代	尘	3811	
	影	3522	
	似	3445	
	身	3386	
	旧	3221	

与之前的试验结果相比，这次的结果更加有时代特点。

先秦的高频词有“曰”“王”“矣”“彼”“既”，还有“民”“臣”也比较高频，可以看出这时多为议论天下之文。后面的汉朝也继承了这一特点。

到了魏晋和隋朝有了明显的变化，“神”“言”“德”出现的比较多，由天下议论转向了重礼的文章。

之前统计的唐朝写景比较多，改变方法后有了明显的不同，“朝”“远”“入”“路”“草”成为了高频词。这点我们不难理解，唐代有很多诗人都是贬谪在外，身居其远，心系当朝，与简单的景物相比，这些字更能体现除作者的情感。

到了五代，诗词变得更加婉转，余恨绵绵。宋代的特点，情感不如之前浓烈，更加细腻。到了辽和金，文风变得更加洒脱。

元代的特点相比之前更加明显，十分口语化，这也符合我们的认知。明代郑和下西洋，西方科技逐渐传入中国，“海”和“西”的频率比较高。从清代开始到近代，“愁”字都非常多，这一时期的情感基本是比较悲的，不同的是清代比较委婉，民国和近代比较朦胧。

可以看到，相比于之前简单统计词频的方法，这次的方法有比较大的改进。

3.3 TF-IDF 诗词关键词分析






之前的思路已经可以得到比较好的结果，但是还有一定的不足。因为之前用的是频数而不是频率，当各朝代数据分布不均匀的时候，使用频数会导致结果的不准确。比如唐代的诗歌本身就很多，减去一定的频数之后影响可能不是很大，而金的诗词较少，减去一定的频数影响会很大。






所以我们想到使用频率进行优化。Tf-idf 本身就是使用的频率，而且在自然语言处理方向应用的比较多，多用于找出文本的关键词。我们把一个朝代抽象为一个文本，找出其中的关键字。






实现和处理过程见附件的代码。

结果如下：

朝代	词	词频	词云
先秦	曰	319	
	彼	301	
	我	209	
	矣	201	
	维	155	
汉	曰	241	
	公	183	
	下	127	
	乐	119	
	我	107	
魏晋	我	166	
	玄	162	
	彼	150	
	德	129	
	化	127	

隋	德	149	
	乐	126	
	玄	125	
	肃	116	
	惟	111	
唐	巖（岩）	167	
	羣（群）	158	
	疎（疏）	157	
	山	149	
	劍（剑）	148	
五代	花	186	
	春	186	
	恨	162	
	帘	162	
	轻	148	
宋	隱（隐）	233	
	巖（岩）	215	
	老	148	
	山	146	
	看	143	
辽	瞋	335	
	熱	321	
	恰	295	
	样	256	
	忤	256	

金	山	222	
	老	189	
	閒（闲）	189	
	看	175	
	寒	133	
元	这	438	
	你	291	
	俺	257	
	个	206	
	哥	185	
明	山	183	
	看	173	
	寒	128	
	花	127	
	海	126	
清	山	158	
	看	148	
	花	141	
	影	139	
	寒	134	
民国	花	158	
	看	146	
	山	139	
	寒	137	
	影	132	

	千里	32308	
	何处	32306	
	万里	30415	
	不可	29612	
先秦	君子	430	
	天下	151	
	子曰	109	
	我心	97	
	诸侯	91	
汉	工子	239	
	天下	222	
	将军	96	
	诸侯	78	
	天子	74	
魏晋	日月	151	
	君子	142	
	天下	132	
	悠悠	130	
	逍遥	117	
隋	千里	55	
	万里	45	
	天地	41	
	万国	35	
	变时	32	
唐	千里	3998	

	孩儿	3981	
	甚么	2993	
	哥哥	2625	
	如今	2541	
明	万里	10490	
	白云	8137	
	千里	7670	
	春风	7303	
	青山	7252	
清	万里	3034	
	东风	3001	
	天涯	2975	
	风吹	2721	
	西风	2685	
民国	东风	1028	
	天涯	758	
	相思	616	
	江南	606	
	人间	603	
近代	天涯	1411	
	人间	1346	
	风雨	1218	
	相思	1161	
	东风	1134	
当代	人间	1452	

	天涯	846	
	风雨	807	
	万里	800	
	春风	790	

通过对双字词的挖掘，我们可以更加清楚地看到各时代的用词特点。

先秦的高频词为“君子”“天下”“子曰”“我心”“诸侯”，从上面高频字的分析就可以看出先秦的格局很大，高频词更加印证了这一特点。汉朝也是同样，“天下”“天子”“诸侯”都比较高频，不同点是先秦“文王”比较多，汉朝“秦王”出现比较多，这也很符合时代特点。到了魏晋，“君子”“天下”依然很多，但是多了许多“天地”“四海”“逍遥”“悠悠”，格局依然很大，但是题材不再那么严肃。

到了隋唐，风格有了很大的变化，“千里”“万里”“明月”“故人”之类的词逐渐变多，情感更加细腻丰富，多为思乡思人。到了五代延续了这样的特点，意象的种类更加丰富，情感更加细腻。宋代相比之前的五代又增加了一些厚重感，不仅有温婉的词，“十年”“千里”“平生”这样的抒怀词增多。辽代的词意象更加丰富，描写的更加细致。金又回归了宋朝有点厚重的风格。

元代的风格依然口语化。到了明代描写风的逐渐变多，到了清代达到了顶峰“东风”“风吹”“西风”“春风”“风雨”“秋风”都是常见意象。到了近现代，意象与明清相差不多，“相思”的频率比较高，与古诗词相比感情抒发的更加直白。

通过对字的不同分析方法，到对高频词的分析，我们一步一步的挖掘出了不同朝代的用词特点。随着实验的深入，实验结果越来越与我们的日常积累和主观认知相符合。

4 情感特点分析

4.1 Word2Vec 词向量分析

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。将每一首诗词作为一个样本，切分为单字，然后使用 genism Word2vec 训练得到词向量。

```
def train():
    # 加载语料
    sentences = word2vec.Text8Corpus(os.path.join(data_dir, "train.txt"))
    # 训练模型
    model = word2vec.Word2Vec(sentences)
    # 保存模型
    model.save('poem.model')
    # 选出最相似的10个词
    for e in model.most_similar(positive=['春'], topn=10):
        print(e[0], e[1])
```

下面是一些词向量分析得到的相似词：

字	关联字
春	秋，花，酴，晴，梅，暄，浓，芳，冬，迟
思	怀，念，肠，情，慰，忆，逢，怙，期，望
梅	梨，楝，桃，杏，蓓，枝，蕾，酴，花，榴
悲	哀，悽，伤，嗟，凄，呻，怆，感，悼，哽
秋	春，霜，寒，凉，凄，嫩，摇，砧，颼，晴

月	蟾，雪，影，烛，镜，日，晃，彻，缸，泖
---	---------------------

挑选其中两个进行单字分析：

```
model.wv.most_similar("春", topn=10)
```

```
[('秋', 0.5571610927581787),
 ('花', 0.5310259461402893),
 ('酴', 0.5048748254776001),
 ('晴', 0.5018565654754639),
 ('梅', 0.4905290901660919),
 ('暄', 0.47397080063819885),
 ('浓', 0.45484331250190735),
 ('芳', 0.4447970390319824),
 ('冬', 0.44370919466018677),
 ('迟', 0.44356799125671387)]
```

```
model.wv.most_similar("梅", topn=10)
```

```
[('梨', 0.6264525651931763),
 ('棟', 0.5899635553359985),
 ('桃', 0.5552749633789062),
 ('杏', 0.5295193195343018),
 ('蓓', 0.528668999671936),
 ('枝', 0.5186370015144348),
 ('蕾', 0.5166165828704834),
 ('酴', 0.5039045214653015),
 ('花', 0.49611562490463257),
 ('榴', 0.49319106340408325)]
```

与“春”相关的字大概可分为两类：同属季节，秋、冬；春天的景象：（花）浓、梅（花）；诗词中春和秋大概是出现最多的两个季节，万物复苏和万物凋零。将两个季节对比的诗词例如，刘禹锡.秋词“自古逢秋悲寂寥，我言秋日胜春朝。晴空一鹤排云上，便引诗情到碧霄。”，这首诗中就连续出现了春、秋、晴三个词。

与“梅”相关的字，大致也可分为两类：同属植物，如杏、梨、桃、榴、杨、柳、棟等；和“梅”相关的意象，如酴（酒）、（梅）花、（梅）枝等。最相关的是梨，梨花同属春天花开，诗词中将梨花和梅花对比的也不少，苏轼在西江月•梅花，有“高情已逐晓云空。不与梨花同梦。”

通过关联词的分析，可见；诗词中类比和借景抒情的手法用的非常多。触景生情、借景抒情也是大多数诗词主题，词向量分析很好地反映了这个特点。

4.2 探究诗词中的七情

在词向量分析的基础上，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…

两个情绪词汇的相似词：

```
model.wv.most_similar("悲", topn=10)
```

```
[('哀', 0.8270668387413025),
 ('悽', 0.6780203580856323),
 ('伤', 0.6488035917282104),
 ('嗟', 0.6021252870559692),
 ('凄', 0.5996484160423279),
 ('呻', 0.5191422700881958),
 ('怆', 0.5104295015335083),
 ('感', 0.4921872913837433),
 ('悼', 0.4905660152435303),
 ('哽', 0.485240638256073)]
```

```
model.wv.most_similar("喜", topn=10)
```

```
[('贺', 0.5157285332679749),
 ('忭', 0.47273269295692444),
 ('悦', 0.471987247467041),
 ('忻', 0.4704793691635132),
 ('欣', 0.4658914804458618),
 ('幸', 0.46543556451797485),
 ('赏', 0.44436633586883545),
 ('好', 0.4424256384372711),
 ('庆', 0.42807865142822266),
 ('乐', 0.4247027635574341)]
```

根据以上词向量相似度分析和人工挑选得到的情绪关键字典。在诗词内容中匹配这些关键字，给诗词加上对应情绪标签，使用神经网络模型对诗词进行情感分类。

4.2.1 CNN pair BiLSTM

为了对诗词进行情感分类，我们搭建了一个简单的深度神经网络。CNN pair BiLSTM，将两个部分的输出连接到一起。

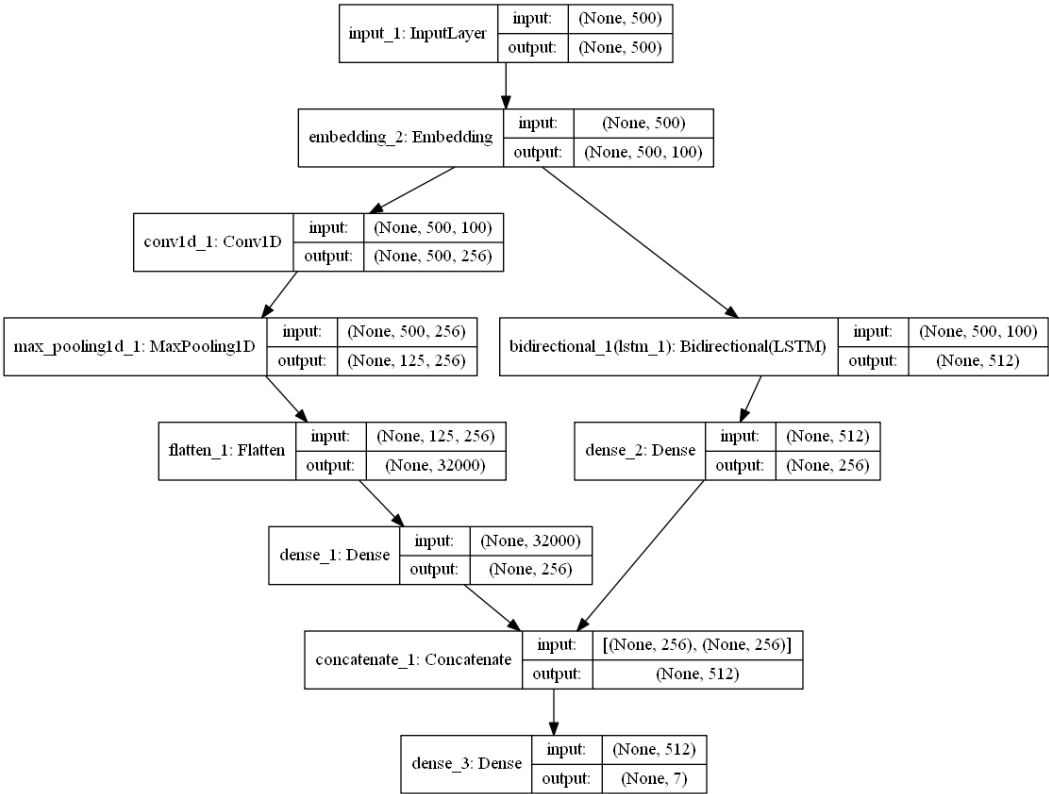
TextCNN 是 CNN 用于自然语言处理的一般方法。将句子映射到嵌入向量，并以矩阵的形式输入到模型中。再适当地使用不同大小的卷积核对所有输入的词执行卷积操作。最后使用最大池化层处理得到的特征映射，将提取到的特征进行浓缩或汇总。TextCNN 能较好的提取关键字信息，在我们的情感分类任务中，分类标签就是通过关键字构建的，因此我们首先想到这种方法。

LSTM 的全称是 Long Short-Term Memory，它是 RNN 的一种。LSTM 由非常

适合用于对时序数据的建模，如文本数据。BiLSTM 是 Bi-directional Long Short-Term Memory 的缩写，是由前向 LSTM 与后向 LSTM 组合而成。两者在自然语言处理任务中都常被用来建模上下文信息。BiLSTM 是目前对给定文本进行情感倾向分类任务中常用的模型。用 LSTM 模型可以更好的捕捉到较长距离的依赖关系。因为 LSTM 通过训练过程可以学到记忆哪些信息和遗忘哪些信息。但是利用 LSTM 对句子进行建模还存在一个问题：无法编码从后到前的信息。而诗词中大量比兴手法的应用，都需要前后的信息的对比才能更好的捕捉诗词的情感。例如“自古逢秋悲寂寥，我言秋日胜春朝”，这类诗词就需要考虑前后的对比分析，通过 BiLSTM 可以更好的捕捉双向的语义依赖。

考虑到情感分类的特点和这两个模型的优点，我们设计了 CNN pair BiLSTM 的模型结构对诗词精细情感分类。

(1) 网络结构图



(2) 模型训练

我们使用上一步训练得到的词向量模型对单词进行编码，转换为对应的 id，并

将得到的 `embedding` 传给模型的 `embedding` 层，将编码的句子和得到的标签传入模型进行训练。下面是模型代码：

```
model = Sequential()
model.add(Embedding(vocab_size, embedding_dim, input_length=max_sequence_len,
                    weights=weights, trainable=True))
sentence_input = Input(shape=(max_sequence_len,), dtype='float64')
embed = Embedding(vocab_size, embedding_dim, input_length=max_sequence_len)(sentence_input)
cnn = Convolution1D(256, 3, padding='same', strides=1, activation='relu')(embed)
cnn = MaxPool1D(pool_size=4)(cnn)
cnn = Flatten()(cnn)
cnn = Dense(256)(cnn)
rnn = Bidirectional(LSTM(256, dropout=0.2, recurrent_dropout=0.1))(embed)
rnn = Dense(256)(rnn)
con = concatenate([cnn, rnn], axis=-1)
output = Dense(num_classes, activation='sigmoid')(con)
model = Model(inputs=sentence_input, outputs=output)
return model
```

4.2.2 各朝代诗词的情感倾向

使用上面训练得到的模型对全部诗词进行情绪分类，得到下表：

朝代	情感分类	代表诗人	情感分类
先秦	思	诗经	思
汉	忧	刘向	悲、思
魏晋	思	曹植	思、乐
隋	思	李世民	思、乐
唐	悲	白居易	思、喜
五代	悲	李煜	思、悲
宋	思	陆游	思、喜
辽	思	萧观音	思、悲
金	思	李俊民	思、悲
元	思	刘基	思、悲
明	思	王世贞	思、喜
清	思	丘逢甲	思、悲
近代	悲	黄浚	思、悲

民国	思	曹家达	思、悲
当代	思	卢青山	思、悲

国家不幸诗家幸。大多数朝代的优秀诗歌的情绪都偏向“思”、“悲”，毕竟贬官文化是古代诗歌非常重要的一环，中国人思乡之情、落叶归根，很多优秀的诗作，李白的静夜思，苏轼的水调歌头，都是描写思念故乡亲情的佳作。

可能出乎很多人的意料，代表大唐气象的唐诗应该以积极昂扬的情绪为主，怎么会是“悲”、“思”、“忧”这样的情绪占据主流呢？而“喜”、“乐”这样的情绪却占据末流呢？

接下来，着重来分析下“悲”这个情绪占据主流的原因：

从常见的唐诗写作题材上说，带有“悲”字基调的唐诗较多，也多出名诗佳句，比如唐诗中常见的几种情结，如“悲秋情结”、“别离情结”、“薄暮情结”和“悲怨情结”，都体现出浓重的“悲情”色彩。

古人云：“悲愤出诗人”，它点破了人的成就与所处的环境、心境有某种关系。就像司马迁所说：“夫《诗》、《书》隐约者，欲遂其志之思也。昔西伯拘羑里，演《周易》；孔子厄陈、蔡，作《春秋》…大抵贤圣发愤之所为作也。此人皆意有所郁结，不得通其道也…”回顾古今中外的著名的诗人和作家，几乎无一不是曾有一段被排挤，诽谤，不得志和身处逆境之经历，有些甚至还很悲惨。正是在这种悲难，恶劣环境中，才使得其奋发图强。

重要的是，唐诗中的“悲”不仅仅是做“儿女态”的悲，更是具有超越时空、怜悯苍生以及同情至美爱情的大慈大悲：

陈子昂的《登幽州台歌》，“前不见古人，后不见来者。念天地之悠悠，独怆然而涕下。”从时间与空间两个角度把悲凉拉长了。李白的《将进酒》中“君不见明镜高堂悲白发，朝如青丝暮成雪”，以及《梦游天姥吟留别》中“世间行乐亦如此，古来万事东流水”让人唏嘘！还有《长相思》第一首中“天长路远魂飞苦，梦魂不到关山难。长相思，摧心肝。”杜甫的《登高》中“无边落木萧萧下，不尽长江滚滚来。万里悲秋常作客，百年多病独登台。”老病残躯，孤苦无依独登台，心中悲凉陡然而生。白居易的《长恨歌》末尾“七月七日长生殿，夜半无人私语时。在天

愿作比翼鸟，在地愿为连理枝。天长地久有时尽，此恨绵绵无绝期。”相爱而不能相聚，生死遗恨。

此外，各朝代 top 情绪分类中，仍能发现一些特点，但总体时代特征并不是十分明显。或者说我们的感情千百年来，在于朝代为单位的周期中并没有十分明显的波动，也许每个时代的人都活得辛苦，也或许悲愤思想更能触发诗人的情绪灵感，“思”，“悲”等是每个朝代都共同的情感。各朝数目最多的诗人其情绪主要集中在“喜”，“悲”，“思”，“乐”，也许也和写诗最多的诗人多出现在超代中期有关。由于诗歌数量较大，大多数诗词还是偏向“思”和“悲”这两个比较容易触动情绪的情感。情绪方面，并无明显朝代特征。

5 结果分析

通过高频词分析、特有高频词分析、高频双字词分析、TF-IDF 特征字分析和 Word2vec 词向量分析、神经网络情绪分析，对历代诗词，用词特点、情感变化有了初步的了解。

先谈用词特点。从先秦到当代，诗词特点较为明显的朝代有：秦汉、唐宋、元朝。从用词特点来看，先秦时期君子、曰、天下等词汇用得较多。唐宋时期为中国古代诗歌文化最为发达的两个时期，涌现大量写景诗词，并持续影响后世。元朝是另一个用词特点十分明显的朝代，特点是大量运用俗语，与元曲特点十分相似。共现双字词是最能代表用词特点的，时代特征非常明显，符合我们的认知。根据目前对历史的划分，中国几千年历史可以大致分为三个阶段：先秦至秦汉是上古，魏晋南北朝隋唐宋是中古，元明清是近古。不同时代背景特点不同。正如白居易所言：“文章合为时而著，歌诗合为事而作”，诗词的用词特点，很好地吻合当时的时代环境。我们也从中捕捉到了一些当时主流文学体裁对诗词创作的影响，汉元的诗词用词就极具汉赋元曲的用词特点。

情感特点。情感的表达必然是基于一定词汇，词汇特点也很好的反映了情感特点。“国家不幸诗家幸，赋到沧桑句便工。”，“思”、“悲”在诗词中始终是一个主流情绪，某种程度上这两个词可与“家”、“天下”相对应。“思乡”、“悲国”，千百年来，这两种情感始终我们的文化种占据重要地位，这也许就是我们千百年来一脉相承的民族文化。

总结：各朝代用词特点有见明显的变化，意料之外的是，除了个别持续时间较短朝代，在以朝代为单位的周期变化中，诗词中体现出的情绪并未出现明显变化。

王国维在《宋元戏曲考》首次提出：凡一代有一代之文学，楚之骚，汉之赋，六代之骈语，唐之诗，宋之词，元之曲，皆所谓一代之文学，而后世莫能继焉者也。此可谓一代有一代之文学。本文仅从诗词的角度出发探究其用词风格和情感变化，难免偏差。

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
- [2] 古代诗人所写的那些唐诗宋词是如何传播出去并且流传至今的？ - 司马乎的回答 - 知乎
<https://www.zhihu.com/question/28977567/answer/56432896>

附录A

代码地址: <https://github.com/WangShengguang/data-mining>