中央财经大学

Central University of Finance and Economics

论文题目 基于机器学习的歌曲分类

学　　校 中央财经大学

姓　　名 王龙飞

王思雨

司徒雪颖

指导老师 刘苗

论文日期 2018/03/20

目录

一、问题描述

音乐流派或者说风格是音乐的重要特征，以 QQ 音乐为例，歌曲流派分为流行、轻音乐、摇滚、民谣、R&B、嘻哈、电子、乡村等。传统的分类中，使用的是人工标记的方法对歌曲流派打标签。人工标记分类的方法准确度高，但是效率低，代价高无法满足海量出现的音乐作品分类需要，且随着歌曲数目的激增，传统的歌曲组织入库形式已经无法满足发展的需要。因此，必须让计算机自动识别歌曲的流派，判断歌曲的类型，所以以计算机程序为核心的歌曲流派自动分类技术是未来发展的必然趋势。而歌词是反映歌曲流派的重要特征，不同流派的歌曲，常用词汇、词汇使用频率往往具有明显差异，因此基于歌词的歌曲流派分类有重大商业价值和可行性。

本文以 2012 第 1 周年至 2018 年第 10 周美国 billboard 榜单上的英文歌曲数据为样本，先对不同流派的歌词进行描述统计分析，以此发现不同流派的歌曲歌词特点，从歌曲所使用的词汇中来挖掘不同流派的歌曲的用词特点和包含的情感。以流行词汇刻画每年的流行趋势，发掘每一年的用词规律。同时找出 2012 年到 2018 年间最热门的歌手以及专辑，并对这些热门歌手和专辑的歌曲风格进行了描述性分析。然后建立机器学习模型和神经网络模型，根据歌词文本对歌曲进行流派分类。分类结果显示，流派为 Pop、Rap/Hip Hop、Country 三种类型的歌曲歌词均与非此类型的歌曲歌词存在较大差异，分类效果较好，尽管其他类型的歌曲由于样本不平衡问题，分类效果较差，本文仍对歌曲流派自动分类有一定的意义。

二、歌曲数据说明与歌词文本预处理

（一）歌曲数据来源与说明

本文使用的数据来自 qq 音乐网站美国公告牌榜（https://y.qq.com/n/yqq/toplist/108.html），采用爬虫的方式进行获取。美国 billboard 榜单是美国乃至欧美国家流行乐坛最具权威的一份单曲排行榜，每周更新一次，能上该榜

单的歌曲必然会受到广泛关注，歌曲流派标签有一定的权威性，且该榜单上绝大部分歌曲为英文歌曲，收集英文歌曲数据较为方便。因此我们选择了最有代表性的美国 billboard 榜单上的歌曲作为歌词流派分类的样本。

具体爬取过称为：在 qq 音乐网站美国公告牌榜页面会显示每一期 100 首上榜歌曲的信息，利用 python 爬虫在此页面获取到歌曲名称，歌曲上榜时间和歌曲 url 链接。采用同样的方式获取到 2012 年 1 月至今的所有上榜歌曲的歌曲名称，歌曲上榜时间和歌曲 url 链接，然后利用歌曲的 url 爬取歌曲的歌词、流派、发行公司、发行时间等详细信息。具体说明如表 1 所示

表 1 变量说明表

变量名称	变量类型	取值范围	备注
歌曲名称	文本数据	God's Plan 等	
歌手	文本数据	Drake 等	
专辑	文本数据	Scary Hours 等	存在缺失值
语种	离散变量	英语、西班牙语等	存在缺失值
流派	离散变量	Pop、Rap/Hip Hop、Country、R&B、Dance 等	存在缺失值
发行公司	离散变量	环球唱片、索尼音乐等	存在缺失值
发行时间	日期数据	2018-01-20 等	
上榜时间	日期数据	2012 第 1 周至 2018 年第 11 周	

（二）歌词文本预处理

由于抓取到的歌词数据除了存在重复样本、非英语歌曲、无歌词的问题外，歌词文本是掺杂着标点，特殊符号，及对文本含义无意义的语助词，不能直接被计算机理解，在做分析前需进行文本预处理。文本预处理主要分为删除数字和非英文字符、去除停用词、转换成小写、词形还原。

1.删除数字和非英文字符

我们使用 python 的 re 模块，利用正则表达式去除数字和非英文字符，如中文、外文、标点符号等。

2.去除停用词

在英文文本中有很多无效的词，比如“a”，“to”，一些短词等，这些我们不想在文本分析的时候引入，因此需要去掉，这些词就是停用词。我们使用python中nltk及其提供的停用词表将歌词文本中的停用词去除。

3.去除低频词

由于存在大量无意义的低频词，比如拼写错误的单词（我们定义出现的频率仅为1次的为低频词，有9696个，总词典词汇数为21250，占总词典数的45.63%）可能会降低分类精度，因此对去除停用词后的文本再删除低频词。

4.转换成小写

由于英文单词有大小写之分，我们期望统计时像“Home”和“home”是一个词。因此一般需要将所有的词都转化为小写。这个直接用python的API就可以搞定。

5.词形还原

词干提取(stemming)和词型还原(lemmatization)是英文文本预处理的特色。两者其实有共同点，即都是要找到词的原始形式。只不过词干提取(stemming)会更加激进一点，它在寻找词干的时候可以会得到不是词的词干。比如"imaging"的词干可能得到的是"imag", 并不是一个词。而词形还原则保守一些，它一般只对能够还原成一个正确的词的词进行处理。我们需要保留有意义的单词，因此使用词型还原而不是词干提取。

在nltk中，使用WordNetLemmatizer，即wordnet词形还原方法对歌词文本中的英文单词进行词形还原。

三、歌曲数据的描述统计分析

（一）歌手角度

选取2012年到2018年每一周的billboard美国公告牌的top100榜单数据，在这六年间统计每个歌手的上榜曲目，根据每个歌手的上榜曲目数量来衡量在这六年间，歌手的音乐创作创作情况以及流行程度状况。通过图3-1上榜歌曲数量的歌手排名柱状图我们可以看到，美国当红炸子鸡Drake在这几年间的创作表现比较好。他所演唱的歌曲有93首都进入了排行榜，更有很多歌曲在top10中出现，创作才华和演唱才华远远的超过了其他歌手。在他之后的上榜曲目最多的歌手是加拿大音乐天才Justin Bieber，乡村音乐小天后Taylor Swift排名第三。但他们与drake的上榜曲目差距并不小。相比之下上榜曲目top10的其他歌手们差距并不大。

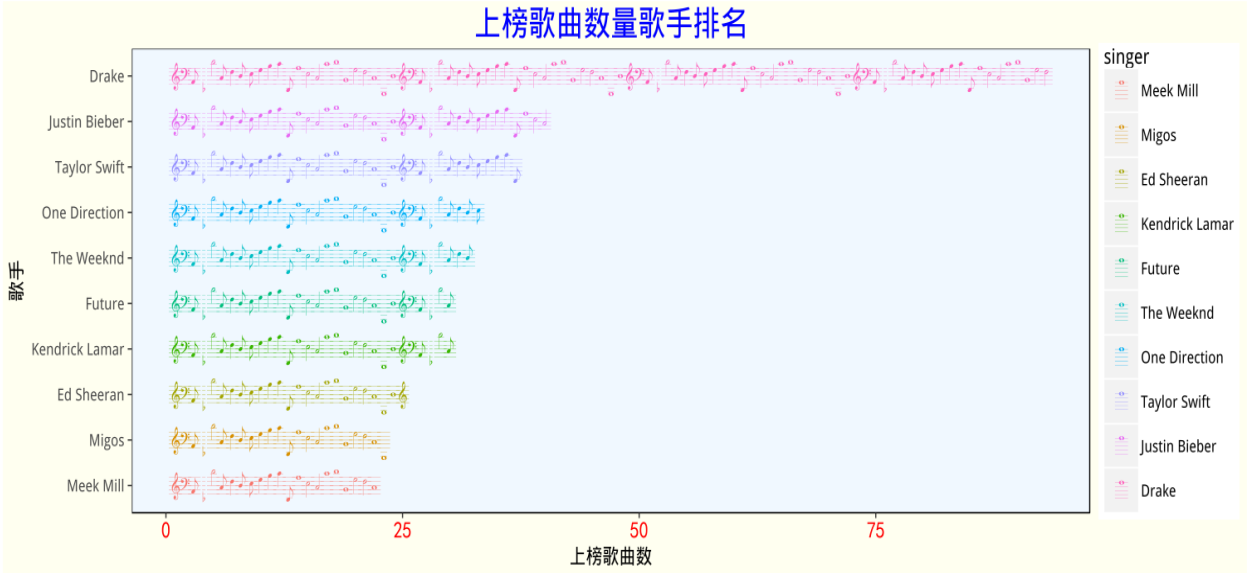

图3-1 上榜歌曲数量的歌手排名柱状图

接下来我们就看一下前四名歌手的音乐风格是怎样子的。如图3-2所示，展示了上榜曲目前四名的歌手的音乐上榜曲目的音乐风格统计。

从图3-2可以看到，当红歌手们的音乐风格都是比较单一的，也就是说歌手们都是有着自己固定的音乐风格的，他们更擅长在一个领域中钻研较为透彻。如当红炸子鸡Drake的上榜曲目的音乐风格只有两种Rock和R&B，在其中摇滚音

乐又占了绝大数，所以可以将 Drake 划分为摇滚歌手。而 Justin Bieber 的音乐风格就更单一了，全部是流行的 pop 音乐。音乐风格较为丰富的是 Taylor Swift，除了流行音乐 pop 之外还有很大的一部分是乡村民谣音乐，并且含有部分 soundtrack 音乐，这展现了她丰富的创作才华和多变的音乐风格。从这个瞳中我们也能清晰的看出，欧美的音乐流行趋势的主流是 pop 音乐，但多元音乐都掺杂其中。

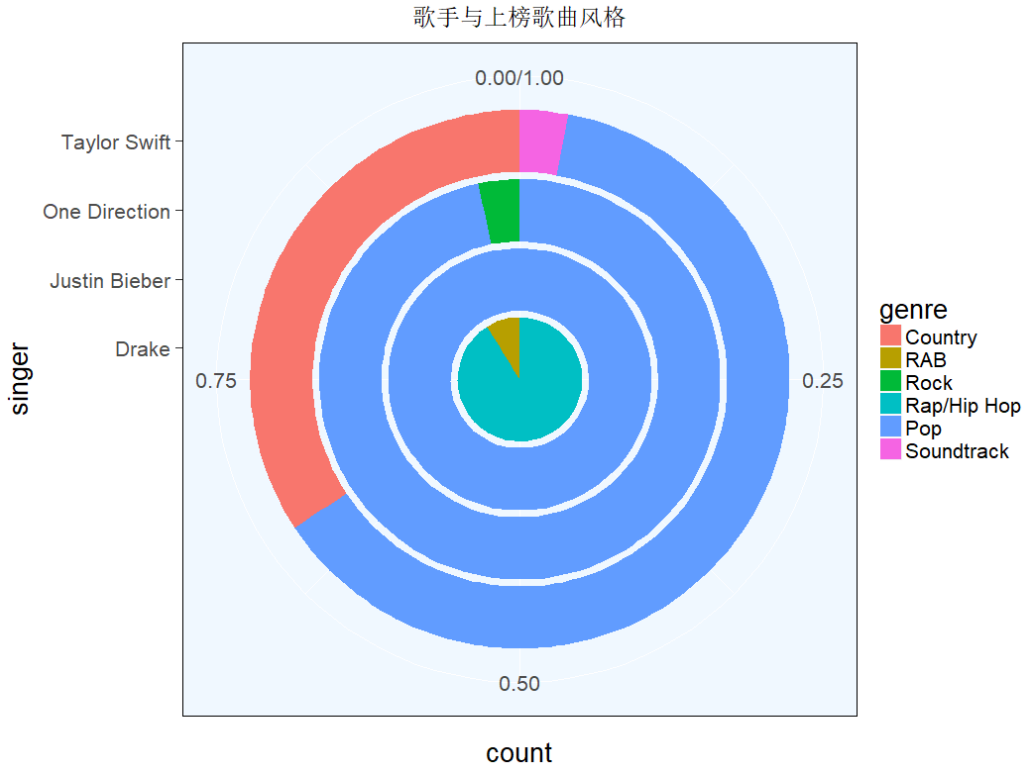

图 3- 2 当红歌手音乐风格统计

（二）公司角度

众所周知，每一个音乐公司都有他们鉴赏音乐的独特风格，他们所发行的音乐也就有这自己的曲风。在本节统计了 2012 年到 2018 年间的 4 大唱片公司的在 billboard 榜单的上榜歌曲的音乐曲风情况，统计结果如图 3-3 所示：

从图中可以看出不同公司的上榜音乐的曲风是有着明显的差别的，例如华纳 PLG 公司，这家公司所上榜的歌曲几乎收拾 Dance 和 Rap 歌曲是有着明显的公司特征的，并没有出品当下比较流行的 POP 音乐。而其他几家公司的情况也不尽相同，但都是会拓展自己的公司的音乐风格，在任何音乐上都有尝试，其中在 POP 上和 Rap 音乐上出品较多，乡村音乐上的比例相差不多。

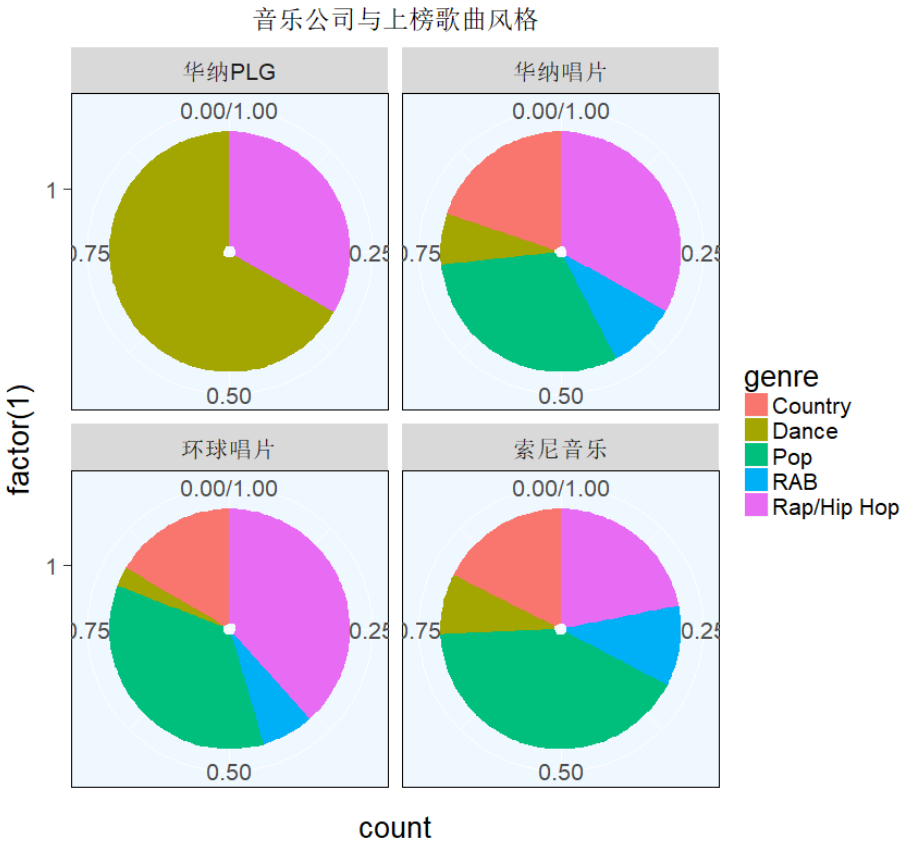

图 3- 3 音乐公司与上榜音乐曲风风格饼图

（三）从专辑角度

对上榜的专辑进行统计，统计出上榜音乐最多的专辑。如图 3-4 所示，统计出了上榜曲目最多的前十的专辑。

从图 3-4 可以看出 8 名歌手的 10 张专辑都进入了排行榜。上榜曲目最多的专辑为 Darke 的《More Life》，这张专辑有着 21 首歌都曾进入过 top100。是比较成功的专辑了。其次，是歌手 Justin Bieber 的专辑《Purpose》，有 18 首歌都进入了榜单，也是比较成功的专辑了。只得注意的是，美国摇滚小王子 Drake 在 2012 年到 2018 年间，有三张专辑都进入了专辑排名，说明 Drake 在音乐上的高产和高效，展现了惊人的才华和在欧美乐坛的统治能力。

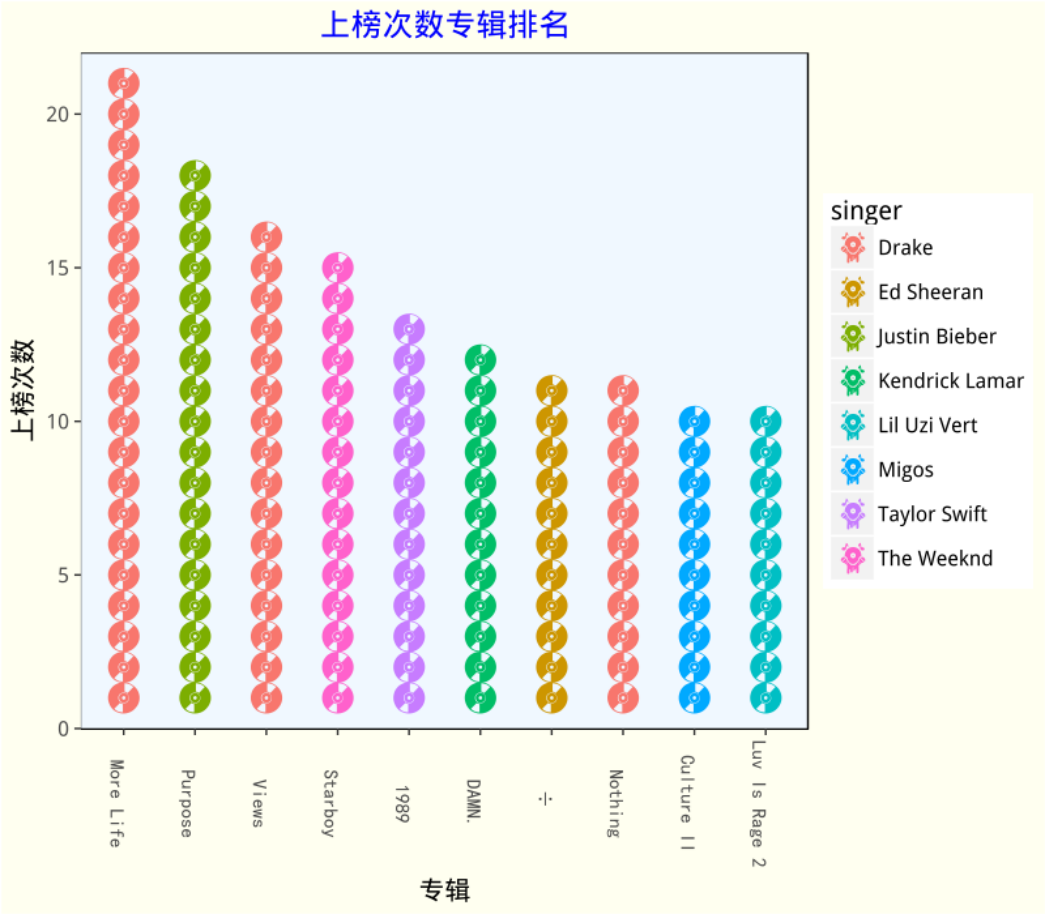

图 3- 4 上榜曲目最多的专辑柱状图

（四）歌词部分描述统计

1.根据年份的词频统计

图 3-5 中，从年份的词频统计来看，从 2012 年到 2018 年间的 billboard 榜单歌曲里面的重点的词汇并没有极大的改变，在每个年份中的词频都是 like 和 love 最多，其次就是 and 和 but 这样的连词，说明欧美流行歌曲大多数还是表达爱意，其次亦可以看出在表达爱意上比较直接。在 2016 年中可以看到 yeah 这个词汇出现的次数较多，可能说明在这一年份中 rap 一类的饶舌音乐在欧美比较火热。

并且还有一点比较值得注意的是，“she”，“girl”，“baby”出现的次数比较多，说明大多数歌曲是由男士进行创作。“you”的词频也比较多，说明大多的歌曲都是在对自己心爱的女孩进行歌曲诉说，也可以从另一个侧面表现出可能男孩子的感情更加细腻。

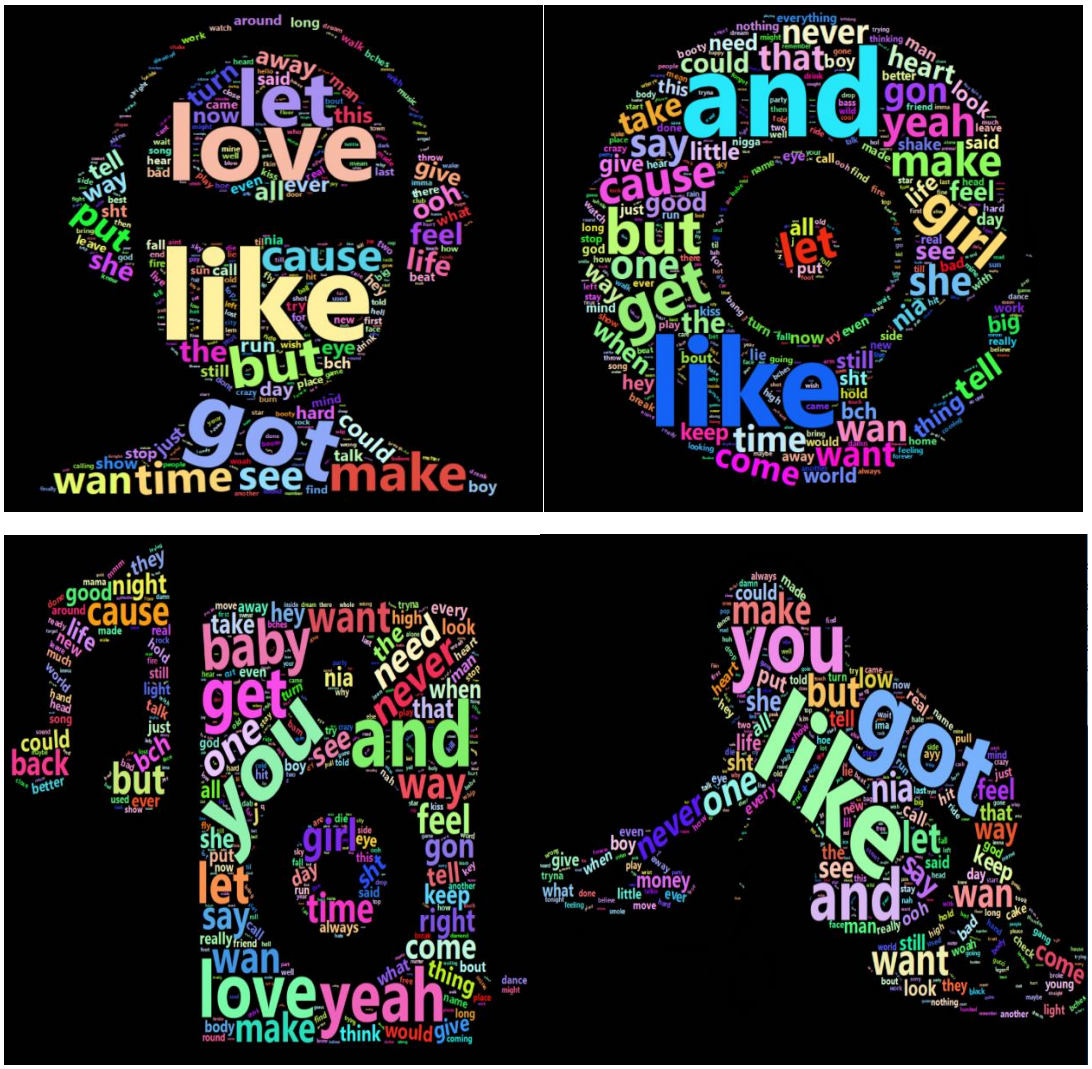

图 3- 5 年份词频统计图（从左到右，从上到下依次为 2012 年，2014 年，2016 年，2017 年）

2.根据曲风的词频统计

从图 3-6、3-7 可以看出，由于 rap 饶舌口语化的特点，所以 yeah,nia 等语气词比较多，每首歌曲 like,got,yeah,get 出现的词频也高于其他三种类型的歌曲。

R&B 歌曲节奏的曲调转换比较快，但从词汇的角度来看也可以看出，转折词汇 but 出现的频率较高，所以可以看出 R&B 歌曲的曲风不光，歌曲的曲调变化较快，而且歌曲的所表达的感情也是千回百转。

由于使用的高频词汇都是一些冠词和接续词，表达语句关系的，所以要看每首歌曲所表达的内容还应该，可能不应该分析更高级的词频，应该研究一些低频词汇，这样更能反映不同风格歌曲的特点。

图 3- 6 不同曲风歌曲词云图

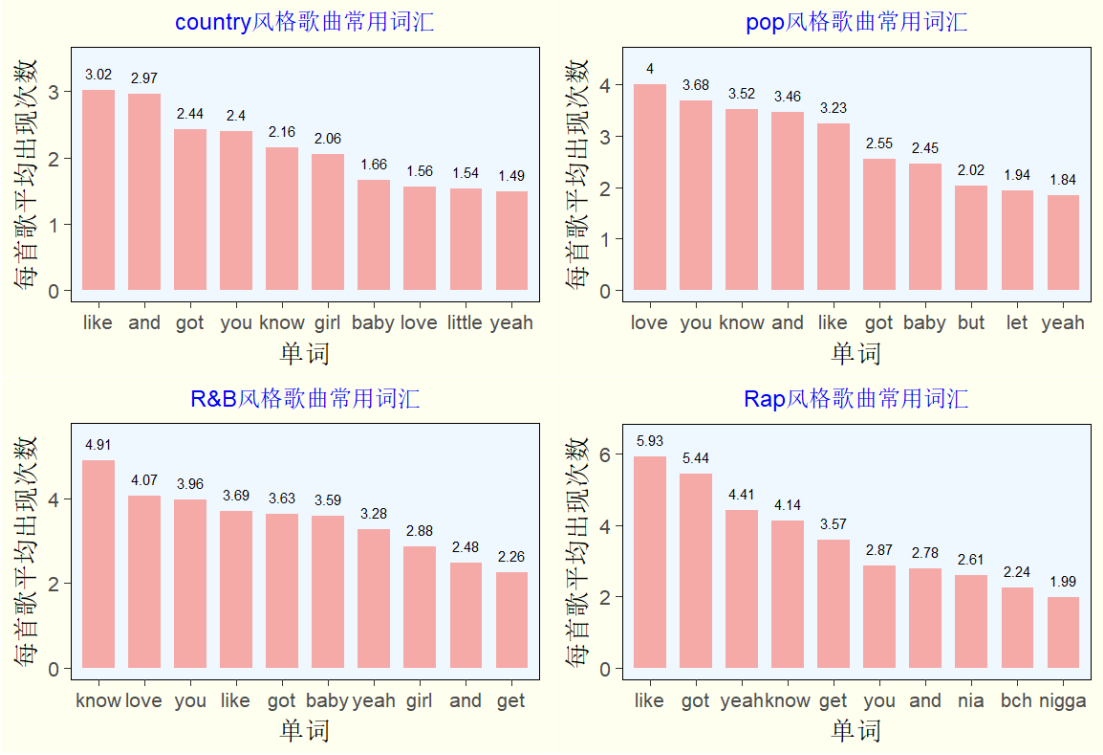

图 3- 7 不同类型歌曲的词频分布

从每首歌出现的词频来看 pop 歌曲是偏向于表达爱意的，不仅 love,you,like,baby 出现的频率高，need,want,feel 出现的次数也很高。Country 类型的歌曲高频词的出现频率要低于其他三种类型歌曲高频词的出现频率,这说明 country 类型歌曲单词分布比较均匀,且单词较少。歌曲比较舒缓，表达的感情不是很强烈。Town,road 等具有明显乡村风格的单词出现次数也较多。

四、歌词分类模型的建立

（一）歌词文本向量化

1.歌词文本处理

首先对歌词文本进行分词、去除停用词和在所有歌曲中出现次数少于 5 次的低频词，总共保留了 4161 个单词。然后计算每个歌曲中单词的 TF-IDF，得到所有歌曲单词的 TF-IDF 矩阵。

2.TF-IDF 介绍

TF-IDF 是自然语言处理常用的一种方法，用以评估单词对于一个文档的重要程度。单词的重要性与它在文档中出现的次数成正比，与它在语料库中出现的频率成反比。TF-IDF 的主要思想是：如果某个词在一篇文档中出现的频率高，并且在其他文档中出现次数很少，可以认为此词具有很好的区分能力，该单词在文档中的重要性就越高。TD-IDF=TF*IDF，单词 t_i 在文档 d_j 中的频率为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}}$$

以上式子中 $n_{i,j}$ 是单词 t_i 在文档 d_j 中的出现次数，而分母则是在文档 d_j 中所有单词的出现次数之和。

逆向文件频率（IDF）是一个词语普遍重要性的度量。某一特定词语 t_i 的 IDF 为：

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

其中，$|D|$是语料库中的文档总数，$|\{j : t_i \in d_j\}|$是包含词语t_i的文档数目。

某一特定文档内的高词语频率，以及该词语在整个文件集合中的低文档频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

3.歌曲类型 One-hot 编码

原始数据集中歌曲类型共有 19 中，但有些歌曲类型的频率较低，预测意义不大，不是本文关注的重点。本文只保留了出现频率最高的 5 种歌曲类型，将其余类型的歌曲标记为其他。在神经网络建模之前，需要将因子型的变量转换为数值型变量。因此将六种歌曲类型编码为 6 个 0-1 变量，然后将数据集划分为训练集和测试集，训练集和测试集的比例分别为 70%和 30%。

（二）基于机器学习方法的歌词文本流派分类

1.朴素贝叶斯（Naïve Beyes）

朴素贝叶斯的平滑参事 alpha 取值为 0.001， 测试集总体分类准确率为 54.60%,如表 4-1 朴素贝叶斯的混淆矩阵表所示，流派为 Dance、R&B 的歌词难以被正确分类，原因之一是它们样本数最少；流派为 country 的歌词分类效果也不好，60%的样本被预测成 pop 和 others；流派为 others 的歌词由于混杂了多种类别的歌词，因此分类效果不好可以预见；流派为 pop 和 Rap/Hip Hop 的歌词样本数最多，分类效果较好，而 Rap/Hip Hop 的分类效果可以达到 82.20%，说明 Rap/Hip Hop 歌词与其他流派的歌词有明显的差异。

表 4- 1 朴素贝叶斯分类混淆矩阵

真实/预测	Country	Dance	pop	R&B	Rap/Hip Hop	others	真实样本数	各流派预测准确率
Country	28	0	16	0	2	21	67	41.79%
Dance	0	2	14	0	6	7	29	6.90%
pop	8	0	129	5	19	42	203	63.55%
R&B	2	0	8	3	17	9	39	7.69%
Rap/Hip Hop	0	0	10	1	157	23	191	82.20%
others	21	2	68	4	40	94	229	41.05%

2.随机森林（Random Forest）

以同样的训练集建立随机森林模型，树的数目取值为 1000，对同样的测试集进行预测，测试集总体分类准确率为 54.70%，与朴素贝叶斯模型分类效果大致相同，如表 4-2 随机森林模型的混淆矩阵所示，随机森林更倾向于把样本预测为数目多的那几类，因此，数目少的流派，如 Country 歌词的分类效果比朴素贝叶斯低了很多，R&B 歌词甚至一个也没有分出来，但是流派为 pop 和 Rap/Hip Hop 的歌词分类准确率上升。

从这一点也可以看出来模型的预测效果各有侧重，歌词流派的分类更适合使用模型集成方法。

表 4- 2 随机森林分类混淆矩阵

真实/预测	Country	Dance	pop	R&B	Rap/Hip Hop	others	真实样本数	各流派预测准确率
Country	6	0	13	0	1	47	67	8.96%
Dance	0	2	18	0	3	6	29	6.90%
pop	0	0	144	1	12	46	203	70.94%
R&B	0	0	19	0	15	5	39	0.00%
Rap/Hip Hop	0	0	18	0	166	7	191	86.91%
others	3	0	88	1	40	97	229	42.36%

3.支持向量机（Support Vector Machine）

第三种机器学习方法方法是带 SGD 的线性 svm，该模型大致思想是 one versus all。假如有 K 个类，每次针对其中一个类建立一个二项的分类器，将样本分类属于一个类的和不属于这个类的。对所有 K 个类进行上述操作，得到 K 个分类器。在对新数据进行分类时，计算样本到各个分类器中超平面的距离，选取

距离最大的一个类作为这个样本的分类。模型正则项采用 L2 范数，避免过拟合。模型分类思想十分适用于歌词文本分类。

测试集总体分类准确率为 51.10%，从表 4-3svm 的混淆矩阵可以看出，流派为 country 的歌词分类效果在三个模型中表现最好，达到 46.27%，其他分类效果并不如上面两个模型，介于两个模型之间。

表 4-3 svm 分类混淆矩阵

真实/预测	Country	Dance	pop	R&B	Rap/Hip Hop	others	真实样本数	各流派预测准确率
Country	31	0	17	0	5	14	67	46.27%
Dance	1	1	12	1	5	9	29	3.45%
pop	15	0	115	2	31	40	203	56.65%
R&B	3	0	13	2	17	4	39	5.13%
Rap/Hip Hop	2	0	17	1	158	13	191	82.72%
others	25	0	76	2	50	76	229	33.19%

4.模型集成

使用同样的训练集建立以上三种模型，对同样的测试集进行分别预测 5 次，每次的随机种子不同，共得到 15 个预测值，取预测值的的众数作为最终的预测值，与真实值作比较，得到下表 4-4 集成模型分类混淆矩阵。

测试集总体预测准确率达到 57.39%，各流派分类效果均达到甚至优于三个模型的最好的预测准确率。

表 4-4 集成模型分类混淆矩阵

真实/预测	Country	Dance	pop	R&B	Rap/Hip Hop	others	真实样本数	各流派预测准确率
Country	29	0	14	0	2	22	67	43.28%
Dance	0	2	17	0	3	7	29	6.90%
pop	5	0	145	1	17	35	203	71.43%
R&B	2	0	16	2	16	3	39	5.13%
Rap/Hip Hop	1	0	18	0	163	9	191	85.34%
others	14	1	76	1	43	94	229	41.05%

（三）基于神经网络模型的歌词流派分类

1.初始参数设置

首先尝试单隐藏层神经网络，输入层的神经元个数等于单词数量 4161 个，隐藏层的神经元数量初始设置为 10，优化算法选择 adam,输出层激活函数选择 softmax 函数，损失函数选择交叉熵。初始预测准确率为 51.5%。

2.优化算法选择

常用的优化算法有 SGD, RMSprop, Adagrad, Adam, Nadam 这 5 种，分别使用他们作为优化算法，得到预测准确率如图 4-1 所示。Adagrad 的预测准确率最高，达到了 55.5%，其次是 Nadam,准确率为 54.5%，SGD 的预测准确率最低，仅有 45%，选择 Adagrad 作为最佳优化算法。

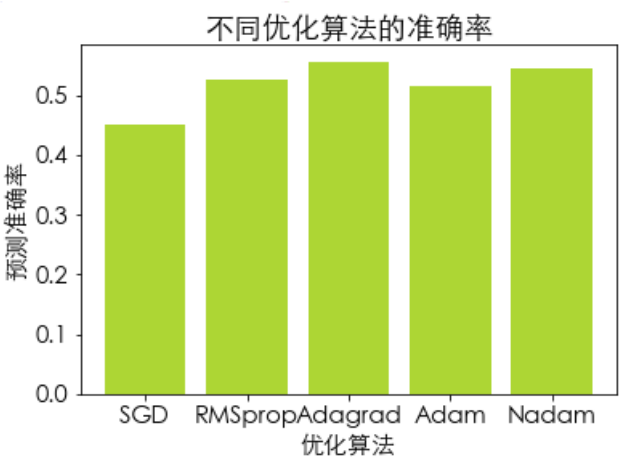

图 4-1 不同优化算法的比较

3.隐藏层神经元数目

接下来考虑隐藏层神经元数目对预测准确率的影响，神经元数目取值范围为 6 到 30，间隔为 3。得到预测准确率如图 4-2 所示，当隐藏层神经元个数为 15 时，预测准确率达到最高，最高为 56.4%。

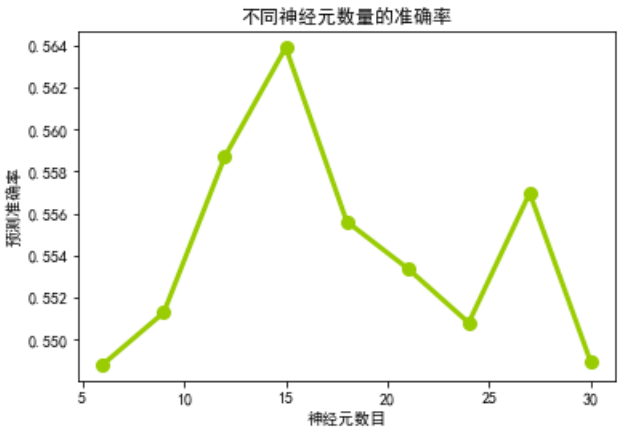

图 4-2 不同神经元数目的比较

4.隐藏层层数

隐藏层层数也是影响神经网络性能的一个因素，将其取值范围设置为 1 到 5，对应的预测准确率如图 4-3 所示。当隐藏层数目为 5 时，预测准确率达到了最高，但准确率并没有显著提升，5 层与 1 层的准确率差异并不明显。

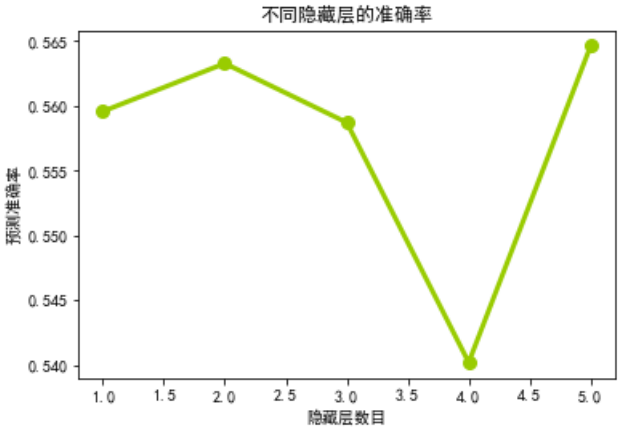

图 4-3 隐藏层层数的比较

5.激活函数

常用的激活函数有有 relu, tanh, softplus,linear 等，分别使用他们作为激活函数，得到预测准确率如图 4-4 所示。tanh 的预测准确率最高，达到了 57.7%，高于其他三种激活函数的预测准确率。

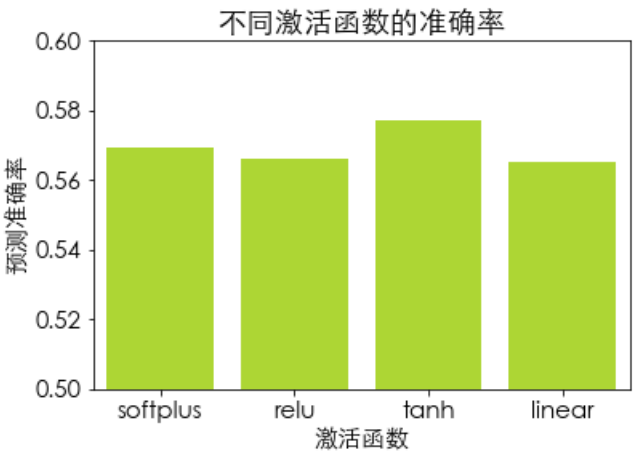

图 4-4 激活函数的比较

6.正则化

为了防止模型过拟合，通常需要设置正则化项对参数进行控制。正则化参数取不同数值对应的预测准确率如图 4-5 所示，当正则化参数为 6 时，预测准确率最高，达到 58.5%。

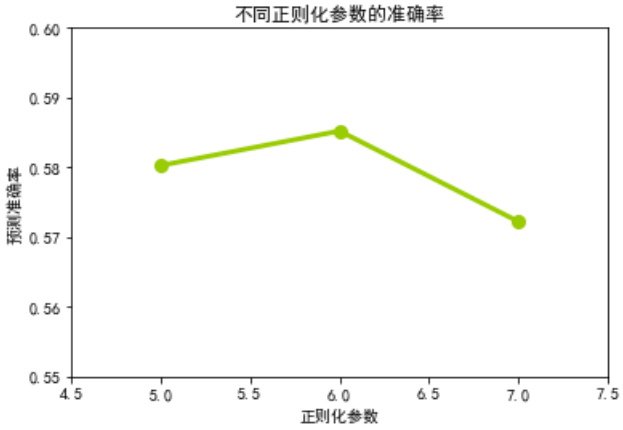

图 4-5 正则化参数的比较

经过上述优化后，神经网络模型的预测准确率达到了最高，约为 59%。尽管还有其他参数可以调试（如学习率和批大小等），但预测精度无法进一步提高。因此将神经网络的参数设置如下：选取 Adagrad 作为优化算法，隐藏层数目设置为 1 层，隐藏层神经元数目为 15，激活函数选择 tanh，正则化参数设置为 6。利用上述参数建立神经网络模型，使用训练集训练神经网络，并对测试集进行预测，此时测试集总体分类准确率为 58.5%。预测结果如表 4-5 所示，流派为 Dance、R&B 的歌曲难以被正确分类，可能是它们样本数最少导致的；流派为 country 的

预测准确率接近 60%；流派为 others 的；流派为 pop 和 Rap/Hip Hop 的歌词样本数最多，预测准确率最高，超过了 70%，说明 Pop 和 Rap/Hip Hop 歌词与其他流派的歌词有明显的差异。

表 4- 5 神经网络分类混淆矩阵

预测值 / 真实值	Country	Dance	Others	Pop	R&B	Rap/Hip Hop	真实样本数	各类型预测准确率
Country	55	0	1	31	0	9	96	57.3%
Dance	0	1	1	19	0	8	29	3.4%
Others	8	0	9	45	0	6	68	13.2%
Pop	12	4	10	143	0	28	197	72.6%
R&B	3	0	0	22	1	16	42	2.4%
Rap/Hip Hop	2	0	0	22	0	151	175	86.3%

五、结论与建议

（一）结论

本文基于 Billboard 排行榜歌曲的歌词文本信息，进行了描述性统计分析和数据模型建模工作。在描述性统计分析过程中，分析了近年来比较流行的歌手、专辑和公司的音乐风格，以及通过统计各年排行榜的歌曲所使用的词汇频率来分析在 2016 年中有一个 Rap 音乐的爆发年。最后对曲风的词频统计中，我们发现了，Rap 音乐是经常会频繁的使用语气词汇，而 Country 音乐对 love 和 like 的使用频率明显低于其他的曲风类型，而 pop 音乐经常使用 baby，girl 一类的词汇，说明大部分 pop 歌曲都是男性为女性创作的。

在文本数据建模过程中，对歌词文本进行预处理，分别是删除数字和非英文字符、去除停用词、转换成小写、词形还原。然后对处理后的文本进行描述性统计，从歌手和歌曲类型两方面分析了歌词的差异，并从时间角度挖掘歌曲的流行趋势。接着分别使用机器学习方法和神经网络模型对歌曲类型进行预测。朴素贝叶斯是文本分类最为常用的机器学习方法之一，在 6 种歌曲类型中，其总体预测准确率达到 54.60%，随机森林的预测准确率与朴素贝叶斯相差不大，但更倾向将样本预测为样本数目多的类别，svm 模型在预测流派为

country 的歌词时准确率较高，因此集成三种模型，博采众长，将预测准确率提升至 57.39%。这与神经网络模型（单隐藏层）的预测准确率（59%）较为接近。本文推测造成神经网络提升分类效果不够明显有以下几种原因：

1. 样本量较小，神经网络的优势无法体现。
2. 神经网络的输入是词频和 TF-IDF 矩阵，这些是歌词文本处理之后的信息，丢失了文本中单词顺序等重要信息。

从预测结果来看，6 种歌曲的预测准确率并不相同，Pop、Rap/Hip Hop、Other 这三种类型的预测准确率较高，而 Country、R&B、Dance 这三种歌曲类型的预测准确率较低。这可能是样本不平衡导致的，后续将通过爬取更多样本数目少的类别的歌曲歌词或采用 SMOTE 算法解决这一问题。

（二）改进方向

无论是神经网络还是机器学习方法，预测准确率都有待进一步提高。未来尝试从两个方面进行改进:

1. 对于样本不平衡问题，通过增加样本数目少的类别的歌曲歌词，或者 SMOTE 算法从数据层面解决，或者对不同类型歌曲的预测错误率赋予不同的权重。
2. 对于神经网络模型，保留原始歌词文本信息，通过卷积神经网络对歌曲类型进行预测。另外，爬取更多的歌曲数据，扩大样本数量。

参考文献

[1] 孙向琨. 音乐内容和歌词相结合的歌曲情感分类方法研究[D]. 苏州大学, 2011.

[2] 孙向琨, 邓伟. 结合 TF-IDF 的歌曲情感多标记分类[J]. 计算机工程, 2011, 37(19):189-190.

[3] 刘鲁, 刘志明. 基于机器学习的中文微博情感分类实证研究[J]. 计算机工程与应用, 2012, 48(1):1-4.

[4] 张键锋, 王劲. 基于文本挖掘与神经网络的音乐风格分类建模方法[J]. 电信科学, 2015, 31(7):80-85.