中文信息处理课程 Project

MC RAP 说唱歌词生成器

曹景辰

14307130003

2017 年 1 月 7 日

1 概述

本 project 是一个自动生成中文说唱歌词的系统。说唱是一种流行音乐形式，MC RAP 是说唱大类别中的一个分支，以不是很快的语速和对节奏感的较强要求为特色。本系统爬取了大量说唱语料对其加工训练，后用了特殊的生成算法生成出一篇说唱歌词。用户可以自定义说唱歌词的韵脚、歌词长度和中心内容。

我把生成的歌词发了朋友圈，大获好评！


●●●○○ 中国移动
下午10:52
92%
消息
详情
卢颖
昨天 下午10:35
666666666666新时代作词人
许振龙
昨天 下午10:35
求欣赏
刘超颖
昨天 下午10:36
太恶性了
曹景辰
昨天 下午10:36
回复许振龙: 我发的这个图是我跑了好多遍之后少数几个还能看的，要再改进....
豪君
昨天 下午10:38
回复曹景辰: 你已经够了。。。
陈雪兵
昨天 下午10:49
你做的是什么Project？
吴韫聪
昨天 下午10:54
景辰加油，你是最棒的
曹景辰
昨天 下午11:09
回复陈雪兵: 自动写歌，natural language process
陈雪兵
昨天 下午11:11
回复曹景辰: 这么高级？写哪种题材的歌？
周金晶
昨天 下午11:28
char-rnn吗
林禹全
昨天 下午11:51
老铁双击666
回复Daniel Wang:

2 算法细节

2.1 数据爬取与清洗

训练集是对生成歌词结果质量有着最重要影响的因素。我选择爬取了 www.xunmai888.com 上的大量中文说唱个词。这个网站上的歌词几乎都是民间 RAP 爱好者自己创作的。没有那些明星的御用词人帮忙创作，所以在语料库中，会出现很多沾染了世俗气息的不雅词汇。同时，因为网友们都是自己自愿创作上传的歌词，所以爬取的语料中混有很多冗余和无关的信息。第一步就是要清洗数据。我采用了正则表达式的方式把我不需要的内容删去并进行了数据整理。这个步骤的细节非常多，再这里就不一一列举了。挑几个比较有代表性的

来说。

- 删去混在在语料中的广告、脏话。形如 `[XXX]+`（`XXX` 为违禁词）匹配后删除整小节内容
- 规范大量语料中的标点符号，删减空行。
- 把作者的注释和非歌词部分删去。这个环节比较麻烦，要自己粗略地看一遍语料库，针对每个情况重写规则后删改
- 删除单句太长或者太短的内容。形如 `^[/w]{0,5}$` 即可

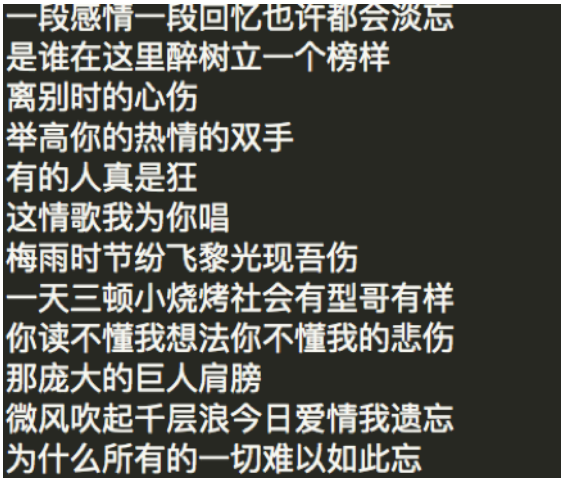
在这个程序中，为了提高歌词的艺术感，我对爬取到的内容分了两个部分。其中 `part1` 是平价较高的高人气用户的创作，`part2` 是广大普通群众的创作。显然，高人气用户的作品更有参考的价值，但也不能单单从高人气用户的创作中训练，需要少量的相对不是那么好的训练数据让生成结果更多元化。（这个思路我是类比神经网络中的过拟合训练而自行发明的，不知道有没有足够权威的理论支持）

2.2 确定韵律规则

好的 `RAP` 一大特色就是很有节奏感，这个节奏感有多重方面来权衡。我们这里排除音乐的元素，就单单歌词而言，歌词需要平仄对应。非常不幸的是，因为没有相关的创作经验，所以我也很难说一段歌词的平仄是否符合我们对“好听”“带感”等评价的期望。同时，因为多音字的关系，平仄对照部分用计算机完成比较困难。所幸这是歌词韵律中相对不是最重要的一部分。节奏感最重要的是押韵！

所以我采用了查字典的方法找出对应的每句歌词最后一个字的韵脚。字典的算法用了最朴素的比较，`python` 内部使用了红黑树实现，效率很高。根据用户输入的歌词韵脚，我找出对应韵脚的每一句话加入集合，完成训练集的初步筛选。

如图，这里采用了韵脚“ang”

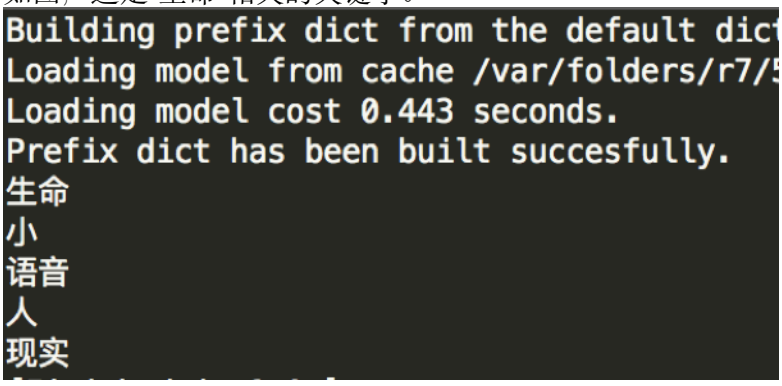


2.3 歌词主题的确定

用户可以输入一个关键词作为创作歌曲的主题。这个关键字建议不要太偏门，不然很难找到对应的特征。比如“希望”、“梦想”等比较常见的艺术主题等。程序会遍历一遍所有语料，对所有含有关键字的句子分词、确定词性。把单词和对应词性加入候选特征集合中去。其中，我们只考虑有意义的词性，比如名词形容词等。代词和“了呢着”等无意义的词汇出现频率很大，我们必须把他从特征集合中剔除。随后对特征集合做 naive bayes 归类，取出最靠前的 5 个词作为关键词特征。

拥有特征之后开始第二次语料遍历，把所有特征相关度高的句子加入训练集中去。在这个过程中，特别在意句子是否符合用户自定义的韵脚。如果符合，则额外增加该句子在训练集中的权重。其余句子的权重则是该句子和特征相关的程度。

如图，这是“生命”相关的关键字。



公式：相关度 $= \sum_{i \in Feature}^{j \in F[i]} sent \times Feature[j]$

另外，我们也可以认为，凡事出现了用户指定关键词的句子附近的句子，也是很有高可能性是相关的有价值的句子。我在这里用了一个正态分布，对出现了关键词的句子及其周围的句子分配了权重。

上文提到的所有选择操作都加入了相关系数，系数越高越有被选中的可能性。这个系数是我自己做了好多次试验后得出的比较可靠的数字。具体系数含义我在程序中写了注释。

```
myRythem = 'an' # 韵脚
myLength = 20   # 歌词句数
neighbour = 10  # 相邻加入特征的个数
hitPara = 20    # 选中概率[0-40]
keyWord = "王"  # 关键词
keyWordPercentage = 0.9 # 关键词相关比例（0-1）
keyWordRythemPercentage = 0 # 关键词相关句中，符合韵律的比例
keyWordNeighbour = 2 # 关键词相邻区间
final = []
```

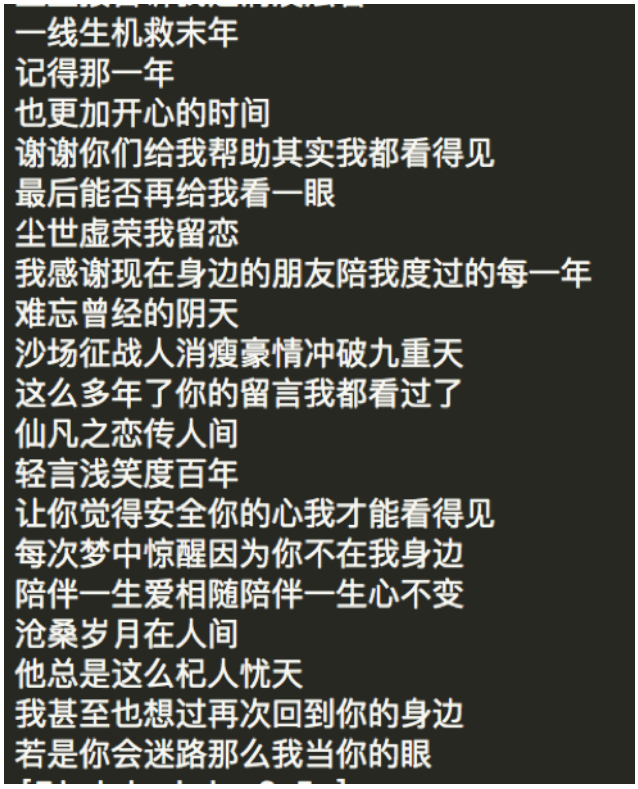
2.4 完整歌词的生成

经过大量筛选，我们拥有了一个比较可靠的训练集。现在我们根据权重和用户自定义的输入条件生成出对应长度的歌词。对每个被选中加入结果的歌词，我再抽取这个句子的句法结构。然后随机地讲非韵脚和关键词相关的部分，用其他相同词性的内容替代。在这个环节中，只有名词、动词和形容词会被替代。因为如果替代了代词等内容，很容易产生让人无法读懂的句子，但是换一个谓语或者宾语，很多时候不影响对句子是否有意义。比如“我在吃香蕉”和“小明在吃红烧肉”都是有意义的句子，但是“我过吃香蕉”就是没有意义的句子。(“过”和“着”的词性相同）被替换的内容也是随机找的。对全文按照词性分词后，用了 jieba 自带的马尔可夫模型分析了词语搭配，随后填入对应的被替换内容。当然，并不是所有符合条件的词性都会被替换，这里我也设置了一个概率参数控制替换的程度。(参数是我自己调的，很难解释出实际的数学或者语言学意义）

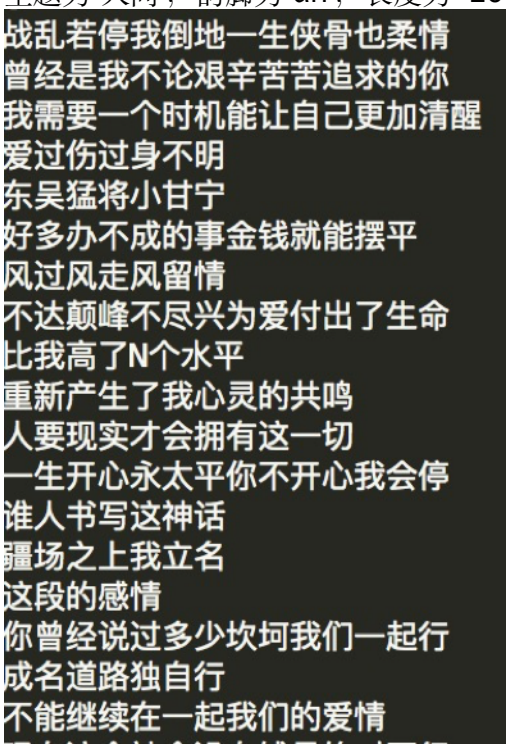
终于，最后的歌词生成啦！

3 操作演示

主题为“感情”，韵脚为“ing”，长度为 18 的歌词



主题为“人间”，韵脚为“an”，长度为 20 的歌词



4 一些想法

1. 目前生成的歌词没有很好的上下文关系，感觉要确定一个上下文关系比较困难，可能需要神经网络的方法来获得上下文关系特征。

2. 个人感觉对数据的清洗和爬取是最重要的部分。着直接决定了生成歌词是否有“美感”可言。而且清洗过程非常繁琐，我提交的都是清洗好的数据，中间的大量步骤比较枯燥，却关系了歌词的质量。

3. 自动写歌相比自动写作，难点在于韵律的把握。但是对句法结构的要求比文章低，因为歌词可以更有弹性。

4. 本 Project 的另一个难点在于训练的参数选择。我用了大约 6 个左右的独立参数，调参的过程同样非常的枯燥。如果老师上课可以讲好的调参方法就更好了。比如不同的训练模型的参数，在语言处理中对哪些因素有着比较大的影响之类的。

5. 祝老师春节快乐