北京航空航天大学软件学院

工程实践结题报告

课程名称：古诗词生成器的设计与实现

姓 名：王雨清 陈立元 吴迪

组 别：

学 号：

专业方向：人工智能研究方向

所属院系：北京航空航天大学软件学院

指导老师：

实验起止时间：

目录

[1 项目概述](#_Toc587218793)

[1.1 项目来源](#_Toc1698895986)

[1.2 项目意义](#_Toc402266190)

[2 项目需求分析](#_Toc609334574)

[2.1 总体需求](#_Toc1884156322)

[2.2 重点需求说明](#_Toc221445192)

[3 项目总体方案](#_Toc240181693)

[3.1 项目总体结构](#_Toc1611941538)

[3.2 关键技术及解决方案](#_Toc1395222261)

[4 项目测试(实验验证)和效果分析](#_Toc1126599034)

[4.1 测试(实验验证)情况说明](#_Toc386648839)

[4.2 运行效果分析](#_Toc121521251)

[5 个人工作总结](#_Toc150717260)

[5.1 吴迪个人总结](#_Toc1221769007)

[5.2 陈立元个人总结](#_Toc33068035)

[5.3 王雨清个人总结](#_Toc1723683319)

1 项目概述

1.1 项目来源

自然语言处理是教会机器如何去处理或者读懂人类语言的系统，目前比较热门的方向，Long Short Term Memory (LSTM)是循环神经网络的改进，在自然语言处理方面有明显的优势，在当前卷积神经网络流行的时期依然担当者重要角色。

中国是诗歌的王国,五千年的优秀文化熔铸了不计其数的脍炙人口的优秀诗篇。许多诗流传千古,哺育了一代代人的成长,是我国极其珍贵的精神财富。研究用自然语言处理技术进行诗歌自动生成，可以帮助普通人进行诗歌创作以及激发创作灵感。但是早起期的中国传统诗歌研究，主要集中在艺术和语言分析领域，直到20世纪90年代才开始有中国传统诗歌自动生成的相关研究。最近几件，以深度学习为代表的人工智能技术取得了越来越多的突破，将人工智能技术用于诗歌自动生成引起了越来越多研究者的注意。本人主要是通过LSTM深度学习技术进行自动生成藏头诗和五/七言律诗的研究。

国外在诗歌自动生成的相关研究开始的比较早，积累了不少方法技术。国内直到20世纪90年代才有中国传统诗歌自动生成的相关研究，起步比较晚，因此还有很多提高和进步的空间。诗歌自动生成不仅可以帮助普通人进行诗歌创作，而且也能以此为契机探索计算机进行艺术创作的可能，因为这方面的研究很有价值和必要。

本研究属于自然语言生成（Natural Language Generation,NLG）领域，它是人工智能在自然语言处理领域的重要应用。Long Short Term Memory (LSTM)是循环神经网络的改进，在自然语言处理方面有明显的优势，在当前卷积神经网络流行的时期依然担当者重要角色[2]。自然语言生成主要以计算语言学和人工智能为基础，研究人类创作自然语言文本的过程[3]。早起的NLG系统一般采用基于模板（Template-Based）的方法来进行文本生成。它的主要原理是先用基于正则表达式的模式匹配技术进行模板匹配，再将模板中的槽位进行替换以得到最终的句子。基于模板的方法由于技术简单并且生成结果非常可靠而得到广泛引用，但是它依然有很多局限性：一方面需要人工来进行繁琐的模板涉及，另一方面人工涉及的模板缺乏灵活性，不能适应复杂的生成情况。为了解决这些问题，人们开始探索其他的NLG方法，以更好的应用不同领域、不同情况下的文本生成任务。

随后有研究者提出基于规则（Rule-Based）的方法[4]。它主要是以树结构来进行文本表达，相比基于模板的系统，可以覆盖更多的自然语言场景和情况，但是基于规则的方法仍然需要人工来决定哪些统计特征可以用于优化目标。整个系统的表现受限于事先定义的语法规则，而在新的应用领域中，仍然需要人工来添加规则。

随着NLP技术的不断发展，基于短语（Corpus-Based）的方法[5]逐渐引起了人们的注意。它定义了一个灵活的学习结构并通过生成模块和重排序模块直接从数据中进行学习，它的最终结果通过对生成器的候选结果进行重排序得到。直接从数据中学习的方法似系统生成的结果更加自然，不然需要提前定义规则并使系统更容易跨领域应用。

近几年有学者提出PPG（Planning based Poetry Generation Model）引入到自动生成诗歌。它包含诗歌规划和诗歌生成两个部分。诗歌规划部分可以利用诗歌语料和外部知识，提前规划处由关键词组成的写作大纲。诗歌生成部分使用改进的基于注意力机制的编码器-解码器，可以同时编码关键词和历史生成内容来主句生成诗歌。

1.2 项目意义

人与人之间的日常交流和沟通一定离不开自然语言以及各种符号表达类语言。因此语言在日常的生活,学习以及工作等活动中有着不可或缺的作用。另一方面,计算机已经开始在多个领域逐步的替代或者简化一些工作,目的就是可以减轻人类的负担。所以人们也在思考计算机能否在自然语言处理与表达上达到模仿人类,或者超过人类的语言表达水平,从而起到辅助人类的作用。

诗歌是一种特有的表达内心情感的文学体裁。通过使用高度凝练的语言、跳跃的语句结构以及一定的节奏和韵律,从而集中反映出当时的社会生活以及表达出个人思想感情,是人类智慧和文明发展在语言表达层面的优秀展现。中国素来享有诗歌之国之称,诗歌文化也是源远流长,但是总结之前工作,主要是从人类进行文学创作的方向上研究,直到20世纪90年代中期才逐渐着眼于从计算语言学角度对进行研究山。因为诗歌创作有着一定的难度并且对创作者的自身文化功底有一定基础要求,所以一直以来诗歌创作就是文人墨客的专属,也是人类智慧的体现。然而如今计算机技术的逐步前进发展,人工智能产品应用越来越广,其正在成为一种潮流,特别是在自然语言处理方面取得的巨大进展,让人们不禁开始思考是否可以让计算机自动或者辅助创作诗歌。

深度学习技术旨在模仿人脑的机制,从数据中学习岀高层抽象的分布式表示。近些年深度学习技术在语音,图像和自然语言处理领域都取得了很大的突破,探索将深度学习技术应用到诗歌生成中有很大的研究价值。我们希望借助深度学习技术,设计出一套比较完善的诗歌生成模型,让计算机具有人一样的表达和写作能力，即能够根据一些关键信息以及在机器中的表达形式，经过一个规划的过程，从而自动生成高质量的藏头诗和五言/七言古诗。它可以为广大诗歌文学爱好者提供很大便利，同时也有助于汉语言的传播与宣扬。

2 项目需求分析

古诗生成系统支持用户录入关键字，通过后台搭建的LSTM训练出的网络生成藏头诗和古诗。

2.1 总体需求

藏头诗是一种常见的诗歌类型,通过将要表达内容的每个字按顺序藏在每句诗之首的方法,来传达作者的思想。藏头诗在藏字的同时,也需要保证诗词格律的要求。

以“美丽华中大”为头生成的藏头诗：

五言诗：

美酒贱人少，上月同辞帷。

丽柳晴光禁，黄云敞照桥。

华丘相见去，罗晰得馀看。

中水复成泰，朱林测不通。

大颜之上安，二妃岂知情。

古诗生成和藏头诗比较接近，差别在于藏头诗是指定一个字，然后使用训练好的LSTM网络预测每一句诗歌的内容。而生成古诗不指定诗歌内容（可以限定是5言或七言），诗歌第一个字有程序随机生成，然后接下来全部由LSTM网络生成，生成的诗歌长度也是不一定的。

五言诗：

花偏君亦长，一别少看花。

项小黔州路，天边山已深。

渡长淮河上，月夜南山分。

情世不可识，歌枝一少愁。

七言诗：

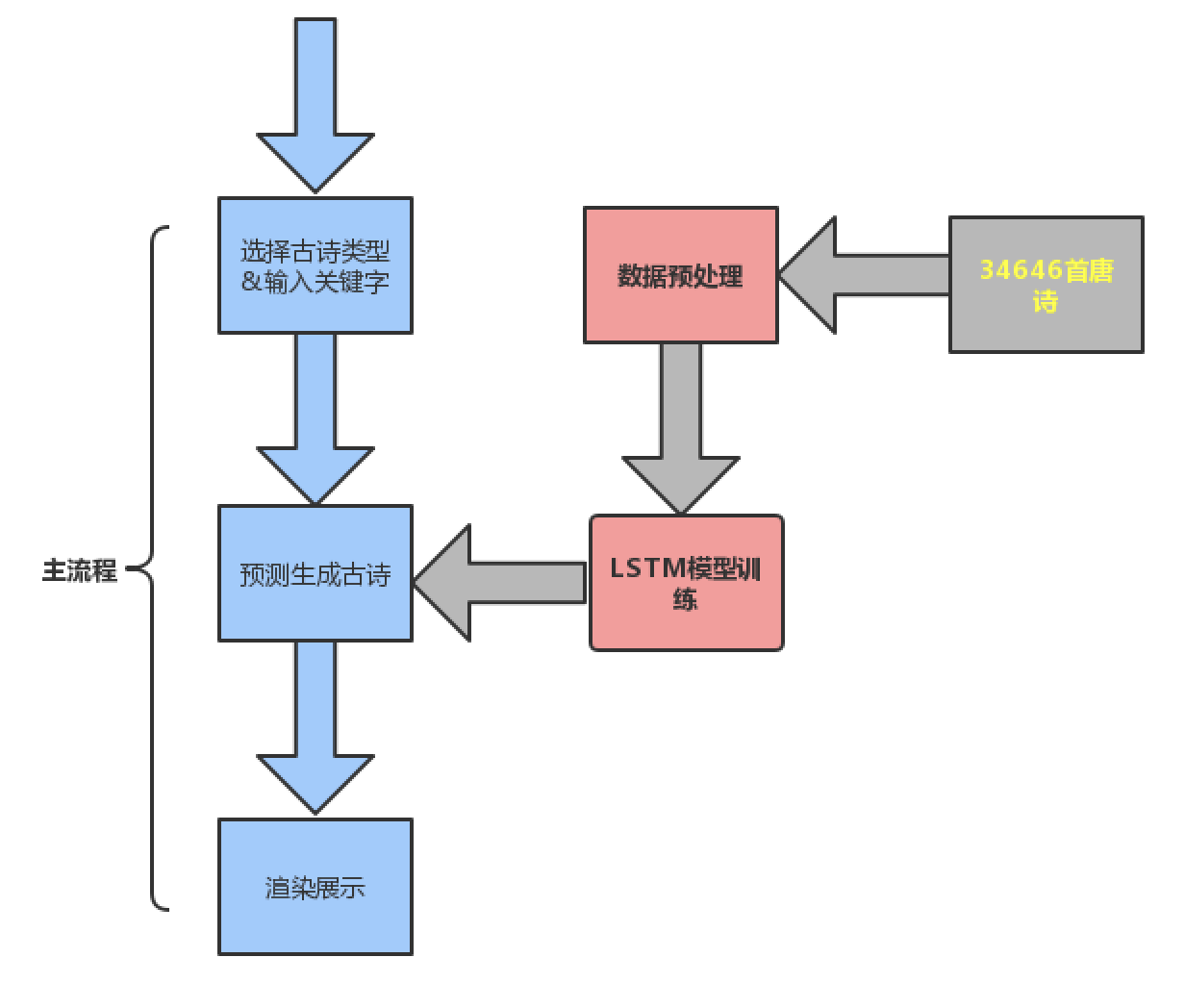
彼处闻寒溜泉频，竹峰蛛网木浓阴。

先生成性思成远，白发幽人事肯醒。

惟有月圆心便寝，起经徒到意无身。

跂襟藉笋丛青菊，声价同亲奈欲何。

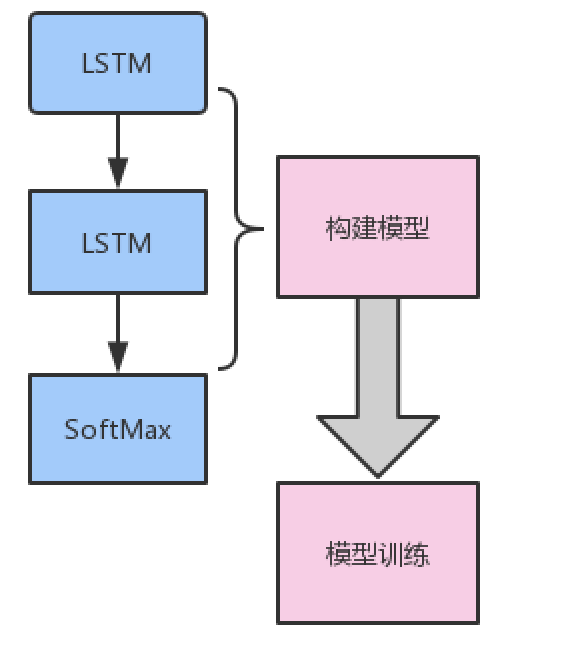
本需求的主要功能是根据前端传入的类型(藏头诗、五言还是七言古诗)和欲生成古诗关键字，利用LSTM生成的网络模型训练最后预测生成相应类型的古诗。其中训练网络模型的部分是本需求重点部分，主要包括数据预处理和使用LSTM模型训练两个部分。



2.2 重点需求说明

2.2.1 **LSTM网络结构**

模型训练过程主要分为搭建模型和模型训练两部分。其中搭建模型主要采用了双层LSTM结构，后接SoftMax函数归一化处理。



LSTM（Long Short-Term Memory）是长短期记忆网络，是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。在介绍LSTM网络之前，需要先介绍一般的循环神经网络（RNN），结构如图（2）。

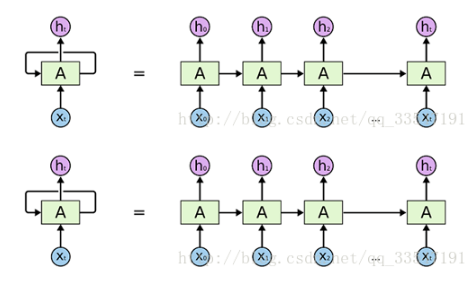


图2、普通的循环神经网络结构

在程序中运行时，循环神经网络采用的是参数共享的方式，也就是实际上一层循环神经网络只有一个cell（图中的A区域），在深入学习研究之后，发现cell内部存在的权重也是在训练的，但这部分权重不需要程序员去定义，而且也不能直接查看，而在模型加载时会加载这部分权值，这是循环神经网络和卷积神经网络很大的一个区别。

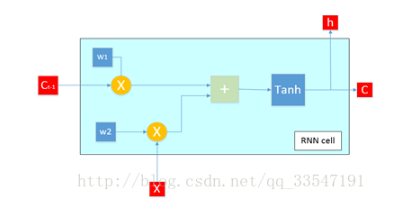


图3、RNN cell结构

从图（3）我们给出了当循环神经网络输入数据维度为一维，内部单元数只有一个时的cell内部结构。该图详细阐述了RNN内部的参数传递。图四是当输入数据维度变为2，cell内部单元数增加为2时的结构，从结构图中可以看到单元之间数据是独立的而每个单元的上一个状态是之间cell内部每个单元的状态集合。也就是单元之前看似独立又相互联系。

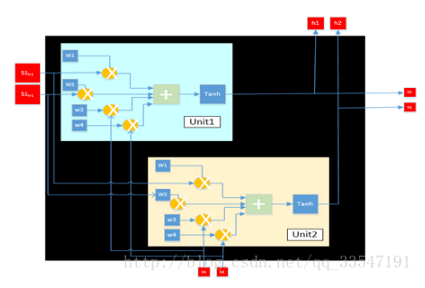


图4、增加单元后循环神经网络结构

而我们使用的LSTM网络和循环神经网络也是采用的链式结构，只是在cell内部设计更为复杂，LSTM在结构上增加了三个门，包括输入门，遗忘门，输出门。LSTM内部参数较多，但是参数之间的联系紧密。

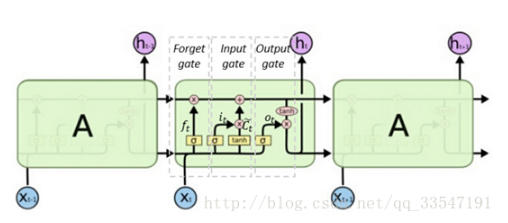


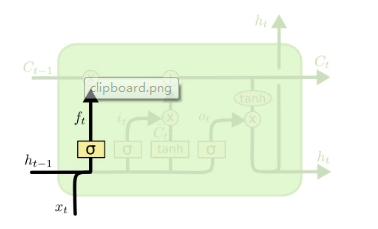
图5、LSTM主要结构

RNN 的关键点之一就是他们可以用来连接先前的信息到当前的任务上，例如使用过去的视频段来推测对当前段的理解。如果 RNN 可以做到这个，他们就变得非常有用。有时候，我们仅仅需要知道先前的信息来执行当前的任务。例如，我们有一个语言模型用来基于先前的词来预测下一个词。如果我们试着预测“the clouds are in the sky” 最后的词，我们并不需要任何其他的上下文 —— 因此下一个词很显然就应该是 sky。在这样的场景中，相关的信息和预测的词位置之间的间隔是非常小的，RNN 可以学会使用先前的信息。

LSTM是一种 RNN 特殊的类型，可以学习长期依赖信息。LSTM 由Hochreiter & Schmidhuber【8】提出，并在近期被Alex Graves【9】进行了改良和推广。在很多问题，LSTM 都取得相当巨大的成功，并得到了广泛的使用。LSTM 通过刻意的设计来避免长期依赖问题。记住长期的信息在实践中是 LSTM 的默认行为，而非需要付出很大代价才能获得的能力，所有 RNN 都具有一种重复神经网络模块的链式的形式。在标准的 RNN 中，这个重复的模块只有一个非常简单的结构，例如一个 tanh 层。

LSTM 的关键就是细胞状态，水平线在图上方贯穿运行。细胞状态类似于传送带。直接在整个链上运行，只有一些少量的线性交互。信息在上面流传保持不变会很容易。LSTM 有通过精心设计的称作为“门”的结构来去除或者增加信息到细胞状态的能力。门是一种让信息选择式通过的方法【10】。他们包含一个 sigmoid 神经网络层和一个 pointwise 乘法操作。Sigmoid 层输出 0 到 1 之间的数值，描述每个部分有多少量可以通过。0 代表“不许任何量通过”，1 就指“允许任意量通过”。LSTM 拥有三个门，来保护和控制细胞状态。

LSTM遗忘门（forget gate）顾名思义，是控制是否遗忘的，在LSTM中即以一定的概率控制是否遗忘上一层的隐藏细胞状态。遗忘门子结构如下图所示：



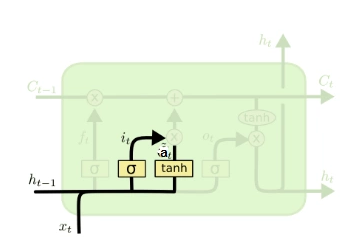
图中输入的有上一序列的隐藏状态h(t−1)和本序列数据x(t)，通过一个激活函数，一般是sigmoid，得到遗忘门的输出f(t)。由于sigmoid的输出f(t)在[0,1]之间，因此这里的输出f^{(t)}代表了遗忘上一层隐藏细胞状态的概率。用数学表达式即为：  
　　　  
　　　　f(t)=σ(Wfh(t−1)+Ufx(t)+bf)  
　　　　  
其中Wf,Uf,bf为线性关系的系数和偏倚，和RNN中的类似。σ为sigmoid激活函数。

LSTM之输入门（input gate）负责处理当前序列位置的输入，它的子结构如下图：

从图中可以看到输入门由两部分组成，第一部分使用了sigmoid激活函数，输出为i(t),第二部分使用了tanh激活函数，输出为a(t), 两者的结果后面会相乘再去更新细胞状态。用数学表达式即为：  
　　　　  
i(t)=σ(Wih(t−1)+Uix(t)+bi)  
a(t)=tanh(Wah(t−1)+Uax(t)+ba)

其中Wi,Ui,bi,Wa,Ua,ba,为线性关系的系数和偏倚，和RNN中的类似。σ为sigmoid激活函数。

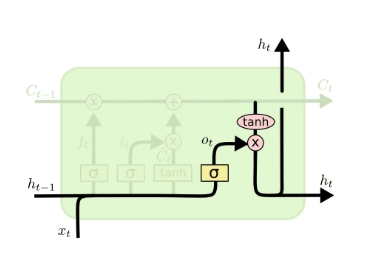
LSTM之细胞状态更新：在研究LSTM输出门之前，我们要先看看LSTM之细胞状态。前面的遗忘门和输入门的结果都会作用于细胞状态C(t)。我们来看看从细胞状态C(t−1)如何得到C(t)。如下图所示：



细胞状态C(t)由两部分组成，第一部分是C(t−1)和遗忘门输出f(t)的乘积，第二部分是输入门的i(t)和a(t)的乘积，即：  
　　　　  
C(t)=C(t−1)⊙f(t)+i(t)⊙a(t)

其中，⊙为Hadamard积，在DNN中也用到过。

LSTM之输出门：有了新的隐藏细胞状态C(t)，我们就可以来看输出门了，子结构如下：

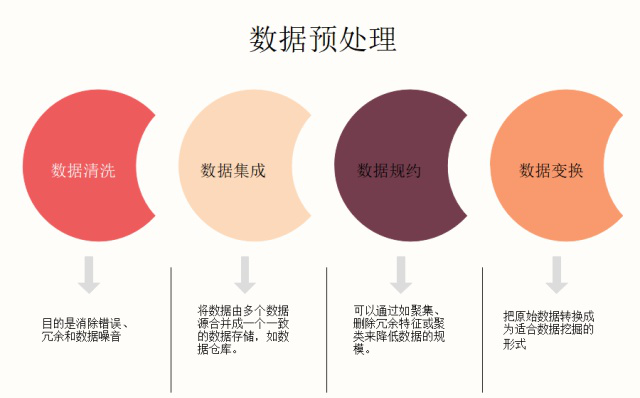


从图中可以看出，隐藏状态h(t)的更新由两部分组成，第一部分是o(t), 它由上一序列的隐藏状态h(t−1)和本序列数据x(t)，以及激活函数sigmoid得到，第二部分由隐藏状态C(t)和tanh激活函数组成, 即：  
　　　　  
o(t)=σ(Woh(t−1)+Uox(t)+bo)  
h(t)=o(t)⊙tanh(C(t))

### 2.2.2数据预处理

因为原始数据往往是不完整的，含有噪声的以及不一致的，采用这些数据进行机器学习往往得到的结果是不准确的。所谓不完整是指缺少相应的属性值；含有噪声是指数据中存在错误的、异常的数据；而不一致的数据是指数据中可能存在在编码上或者命名上的差异重复。

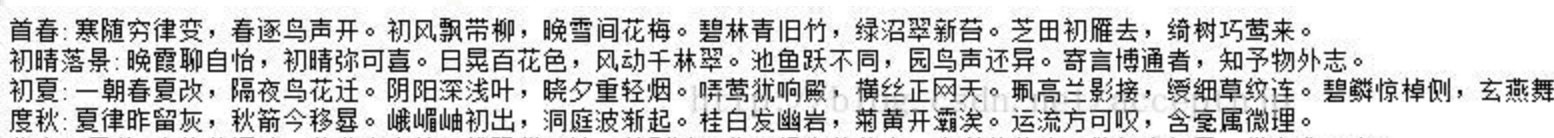
**数据预处理主要包括以下步骤：数据清洗，数据集成，数据规约以及数据变换**



图

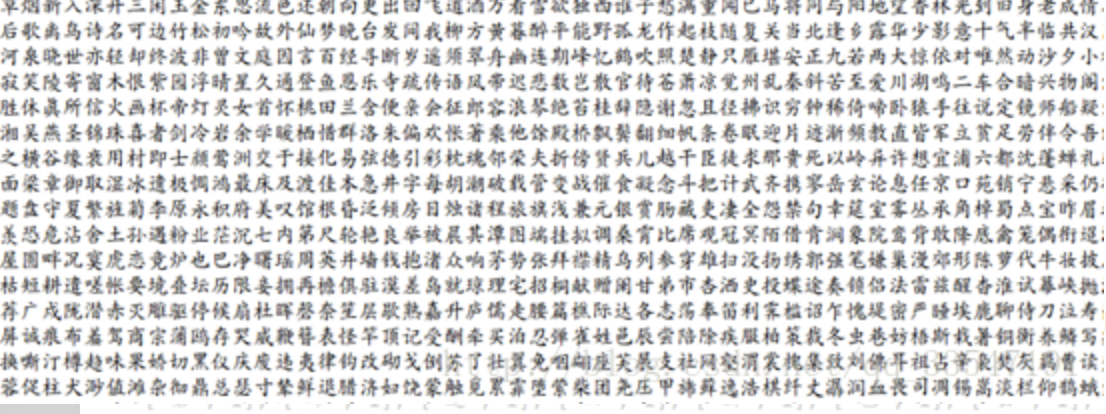
本文数据预处理步骤具体包括数据清洗、统计词频、生成字库、转换向量四个主要步骤。

藏头诗自动生成采用的是全唐诗，一共包含34646首唐诗。数据格式为“题目：诗句”，如图（9）。所有诗歌存放在一个txt文档中，每一行存储一首诗。



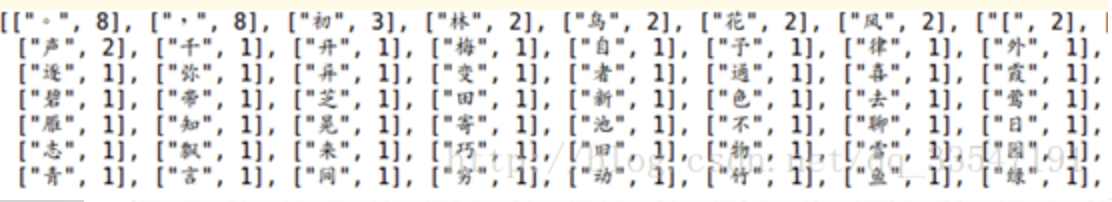
1. 进行原始数据清洗（这里只进行去除诗歌题目）

过滤一些不好的训练样本，包含特殊符号、太少或太多的字，最后在诗的前后分别加上开始和结束符号，用来告诉LSTM这是开头和结尾，这里用方括号表示。



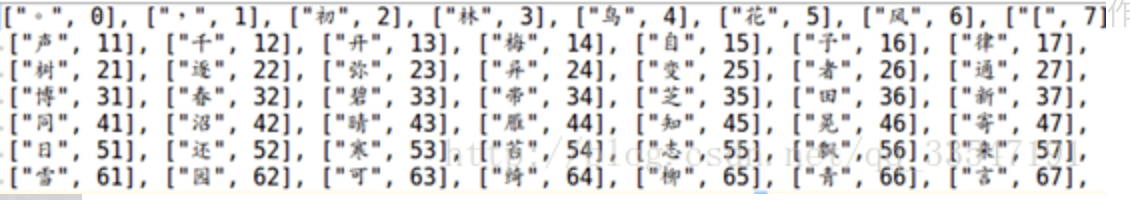
2、统计词频

诗歌中就是统计字频（即每个字出现的次数），并删除出现次数较少的生僻字



3、生成字库

根据字出现的次数从大到小排序，并建立字到ID的映射（这里索引从0开始）



4、转换向量

根据字库把每一首诗转化为一个向量。这个向量以字库中的汉字数作为维度，每个维度中通过0和1来表示这个汉字在这首诗中是否出现。

3 项目总体方案

本节是对人工智能软件内容设计文档的综述，说明该项目所涉及的总体设计方案。

3.1 项目总体结构

说明人工智能软件结构设计、算法原理

3.2 关键技术及解决方案

对系统的原理及其关键算法的解决方案，或系统的关键技术和重要模块的设计与实现。

4 项目测试(实验验证)和效果分析

本节是对软件系统实现内容测试情况的说明，并介绍软件实际运行效果。

4.1 测试(实验验证)情况说明

介绍针对项目所作的测试情况:实施了哪些测试、通过测试发现的问题已经如何解决。

4.2 运行效果分析

介绍该项目实际的运行情况，可以截取 1~2 个界面有针对性地进行说明。

5 个人工作总结

针对项目组的每一个成员通过一小节来说明该成员在该项目中的工作情况。

5.1 吴迪个人总结

说明个人在项目中所承担的角色，主要负责完成的工作，以及对该项目的体会。

5.2 陈立元个人总结

说明个人在项目中所承担的角色，主要负责完成的工作，以及对该项目的体会。

5.3 王雨清个人总结

说明个人在项目中所承担的角色，主要负责完成的工作，以及对该项目的体会。