

**数据挖掘导论大作业**

题目Title：K-means算法

院 系

School (Department)：数据科学与计算机学院

专 业

Major： 软件工程专业

学生姓名

Student Name： 李平

学 号

Student No.： 16340121

时间：2019 年 06 月 14 日

Date: Month Day Year

**【摘 要】**

中国古诗词是中文独有的一种文体形式，通过深度学习以自动生成中文古诗词是自然语言处理领域一项具有挑战性的任务。XXX...

**【关键词】：**深度学习；古诗词生成；循环神经网络；XXX

[ABSTRACT]

Chinese poetry generation is a very challenging task in natural language processing. Chinese poetry has its unique tune and structure, and thus it needs to train the model with a large variety of existing Chinese poetries via supervised learning. In this paper, XXX…

**[Keywords]:** deep learning; poetry generation; recurrent neural network;XXX;

目 录

第一章 概述/引言 5

1.1 研究基于深度学习的古诗词自动生成算法的背景和意义 5

1.2 问题的描述 6

1.3 本文的工作 6

1.4 论文结构简介 7

第二章 相关工作综述 8

2.1 诗歌自动生成相关发展 8

2.2 循环神经网络(Recurrent Neural Network,简称：RNN) 8

2.3 长短时记忆网络(Long Short Term Memory Network,简称：LSTM) 10

2.4 门控循环单元(Gated Recurrent Unit, 简称：GRU) 14

2.5 编码器-解码器结构 15

2.6 Attention机制 15

第三章 完成古诗词自动生成算法的方法 17

3.1 基于深度学习的古诗词自动生成算法的问题描述 17

3.2 逐步完成算法 17

**3.2.1** **训练数据集的处理以及构建训练batch** 17

**3.2.2** **LSTM模型的搭建** 18

**3.2.3** **模型训练** 19

**3.2.4** **自动写诗** 20

第四章 仿真/实验结果与分析 21

4.1 数据集 21

4.2 评价 22

4.2.1 评价标准 22

4.2.2 结果 22

4.2.3 人机诗词盲测 23

4.2.4 生成示例 23

第五章 总结与展望 25

参考文献： 26

致 谢 28

# 概述/引言

## 研究基于深度学习的古诗词自动生成算法的背景和意义

语言对于地球上的生物而言至关重要，能够掌握语言也是人类相较于地球上其他的物种之间的显性特征。人类的智慧与语言密切相关，可以说没有语言的话，人类的智慧可能会下降许多。人类在日常生活中的思考，规划等等都需要基于语言来作为载体，人们也会通过语言传递不同的想法，传递人类所需要的信息，传播人类发展赖以生存的知识，传播新的思想。所以，语言是人工智能的重要和核心部分。

XXX…

## 问题的描述

基于深度学习的古诗词自动生成算法就是要通过深度学习方法来学习已有的古诗词，来得到一个语言模型，再通过这个模型自动生成古诗词。XXX…

## 本文的工作

本文选择循环神经网络的特型长短时记忆网络来学习古诗词数据集，通过长短时记忆网络能够比基本的循环神经网络更加注重长期依赖，XXX…。

## 论文结构简介

本文第一章主要是阐述研究基于深度学习的古诗词自动生成算法的背景和意义以及本文所要大致工作内容；第二章主要综述本文在完成古诗词自动生成算法的过程中所作的相关工作以及选择的考量；第三章将会介绍我在解决问题过程中所作的工作以及所提出的方法和实现算法的步骤；第四章是实验结果的评估；第五章是对本文的总结以及对接下来研究方向的展望；最后是参考文献以及致谢内容。

# 相关工作综述

## 诗歌自动生成相关发展

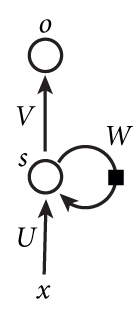
诗歌的自动生成一直是过去几十年很受欢迎的研究课题（见Colton等人(2012)[1]及其中的参考文献）。XXX...

## 循环神经网络(Recurrent Neural Network,简称：RNN)

循环神经网络(Recurrent Neural Network, 简称RNN)首先用于自然语言处理领域，并且可以通过使用RNN可以对语言模型进行建模。XXX...

图1是一个我们简化过的循环神经网络，我们可以将其视为一个单元或者细胞，这个细胞由几个部分组成。

XXX...



**图 SEQ 1：简单的循环神经网络图示。**

## 长短时记忆网络(Long Short Term Memory Network,简称：LSTM)

长短期记忆网络，通常称为“LSTM” (Long Short Term Memory Network)，是一种特殊的RNN，能够学习长期依赖性。XXX...

## 编码器-解码器结构

XXX…

## Attention机制

XXX...

# 完成古诗词自动生成算法的方法

## 基于深度学习的古诗词自动生成算法的问题描述

XXX...

## 逐步完成算法

### **训练数据集的处理以及构建训练batch**

XXX...

### **LSTM模型的搭建**

XXX...

### **模型训练**

XXX...

### **自动写诗**

XXX...

# 仿真/实验结果与分析

## 数据集

XXX…示例如表格1。

**表格 SEQ 1：训练数据格式**

|  |
| --- |
| 宿龙潭寺::白居易::夜上九潭谁是伴，云随飞盖月随杯。明年尚作三川守，此地兼将歌舞来。 |
| 对雪献从兄虞城宰::李白::昨夜梁园里，弟寒兄不知。庭前看玉树，肠断忆连枝。 |
| 过西堡塞北::温庭筠::浅草干河阔，丛棘废城高。白马犀匕首，黑裘金佩刀。霜清彻兔目，风急吹雕毛。一经何用厄，日暮涕沾袍。 |
| 望宅中树有所思::卢照邻::我家有庭树，秋叶正离离。上舞双栖鸟，中秀合欢枝。劳思复劳望，相见不相知。何当共攀折？歌笑此堂垂。 |
| 偶作::白居易::红杏初生叶，青梅已缀枝。阑珊花落后，寂寞酒醒时。坐闷低眉久，行慵举足迟。少年君莫怪，头白自应知。 |

## 评价

### 评价标准

XXX…

### 结果

1. 人类评价的结果展示在表格2中。XXX…

**表格 SEQ 2：人类对于三种语言系统的评价**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **模型** | **诗意** | | **连贯** | | **流畅** | | **意义** | | **平均** | |
| **五言** | **七言** | **五言** | **七言** | **五言** | **七言** | **五言** | **七言** | **五言** | **七言** |
| SMT | 3.15 | 3.19 | 2.77 | 2.83 | 2.98 | 3.01 | 2.53 | 2.74 | 2.86 | 2.94 |
| RNNLM | 2.77 | 2.45 | 3.41 | 3.29 | 3.13 | 3.24 | 2.86 | **3.12** | 3.04 | 3.02 |
| LSTM(our) | **3.21** | **3.43** | **3.52** | **3.55** | **3.25** | **3.28** | **3.11** | 2.98 | **3.27** | **3.31** |

### 人机诗词盲测

XXX...结果如表格3所示。

**表格 SEQ 3：人类盲判以判断诗词是人写的还是机器写的结果**

|  |  |  |
| --- | --- | --- |
| 无法分辨是人写的还是机器写的 | 错将机器写的当成人写的 | 成功辨认出是人写的 |
| 13.3% | 33.3% | 53.4% |

### 生成示例

#### 自动生成

**表格 SEQ 4：自动生成示例**

|  |  |  |
| --- | --- | --- |
| 银城晓视凌江玉，  舫宿蓬榭碣洞庭。  敢遇仰朝陪育鲁，  五移苍泽母荒皋。 | 旧居疏密入东窗，尚有庭枝雨满林。  造物小传箬帐饮，那知赤忍岂无钱。  绝风踏破江衢外，买陂青乌万古风。  恰见铁坛遥弛处，十载香风寄同来。 | 了无时暑方相见，  不必离眠忽不知。  犹植劫间底骨髓，  壁间一画等佳人。 |

#### 藏头诗生成

**表格 SEQ 5：藏头诗生成示例**

|  |  |  |
| --- | --- | --- |
| Input：墨雨南怜 | Input：麦地初茶 | Input：树深见鹿 |
| 墨白三掌青，  雨露秋牵月。  南浦偏精刊，  怜我身更觉。 | 麦熟青精又早花，  地寒无尽暖寒香。  初开夏亩修篱外，  茶芭仗茶笋笋箬。 | 树云万里上山峰，  深念幽人十欲攀。  见说黑云来隐钥，  鹿如甘袭旧枝浸。 |

# 总结与展望

本文中主要解决了基于长短时记忆网络来训练语言模型并自动生成古诗词的方法…

XXX…

# 参考文献：

1. Simon Colton, Jacob Goodwin, and Tony Veale. 2012. Full-FACE Poetry Generation. In Proceedings of the International Conference on Computational Creativity, pages 95–102, Dublin, Ireland.