**数据结构实验设计报告**

**题 目：­­**基于马尔科夫链的写作机器人

**专 业：**计算机类

**小组成员：**计科A2001 邓岳

计科A2001 熊圳

计科A2001 赵千慧

计科A2001 赵凯

# 任务书

|  |  |  |  |
| --- | --- | --- | --- |
| 课程名称 | 数据结构实验 | | |
| 设计题目 | 基于马尔科夫链的写作机器人 | | |
| 指导教师 | 彭四伟 | 时 间 | 2022年1月3日-7日 |
| **一、教学要求**  1. 巩固学生使用程序设计语言进行程序设计、编写和调试的能力； 2. 掌握数据结构常用经典算法原理、实现及其应用； 3. 提高计算机专业技能的自学能力； 4. 掌握书写程序设计说明文档的能力；  **二、设计要求及成果**  1. 分析课程设计题目的功能需求，完成一个题目； 2. 写出详细设计说明（至少**包括功能实现分析**和**模块流程图**） 3. 编写程序代码，调试程序使其能正确运行（代码书写要规范，标示符要见名知意，要有必要的注释）； 4. 设计完成的软件要便于操作和使用，在软件界面上有作者版权信息； 5. 设计完成后提交设计报告（请严格按照模板进行排版）；  **三、评分标准**  1. 使用的算法（20分）：使用了哪些数据结构课程中学到的经典算法；  2. 程序功能完成情况（20分）：功能是否能用，是否易用，是否完备；  3. 程序界面（15分）：界面是否美观大方，设计合理；  4. 课程设计报告（30分）：报告是否内容充实，图表丰富，格式正确，没有错别字；  5. 程序健壮性（15分）：运行过程中发现程序一个BUG，扣3分，扣完为止；  **四、郑重提示**  原则上，只要编码并实现了一个可用的基本程序即可及格！但如果发现非自己编写、拷贝他人（含网络），直接不及格！ | | | |

# 目 录

[任务书 I](#_Toc9021)

[目 录 II](#_Toc27954)

[第一章 需求分析 3](#_Toc11897)

[1.1 引言 3](#_Toc2148)

[1.2 任务概述 3](#_Toc28625)

[1.3 数据描述 3](#_Toc11953)

[1.4 功能需求 3](#_Toc21642)

[1.5 运行需求 3](#_Toc13764)

[1.6 任务计划 3](#_Toc27446)

[第二章 概要设计 4](#_Toc2865)

[2.1 总体设计 4](#_Toc18233)

[2.2 面向对象设计 4](#_Toc26726)

[2.3 运行界面设计 4](#_Toc18813)

[第三章 详细设计 5](#_Toc19218)

[3.1 系统设计与实现 5](#_Toc16310)

[3.2 模块设计 5](#_Toc32206)

[第四章 测试分析 6](#_Toc5868)

[第五章 用户手册（可选） 7](#_Toc6928)

[第六章 课程设计总结 8](#_Toc800)

[附录1 工作量及成绩自评表 9](#_Toc1195)

[附录2：程序代码（仅电子版需要、纸质版不需要） 10](#_Toc19781)

# 需求分析

## 引言

## 自动生成文本是自然语言处理中一大研究领域，目前常用的生成方法包括深度学习和马尔可夫链等。在本次数据结构实验设计中，我们利用马尔可夫链作为基本原理制作了一个写作机器人。

## 马尔可夫链是描述可能事件序列的随机模型，其中每个事件的概率仅取决于前一个事件中获得的状态。马尔可夫链文本生成的思想与此相同，即试图找出某个词出现在另一个词之后的概率。为了确定转换的概率，我们用一些文本来训练模型。它使用历史文本，将其分割成单个单词（或单词集），然后随机选择一个单词，之后基于之前选择的单词序列随机选择一个相似的单词。这个过程一直继续，直到生成一段由规定数量的句子组成的文本段落。

## 利用马尔可夫链文本生成方法，我们可以用一个轻便简易的方式来模仿智能，生成一些看起来有意义的文本段落。随机文本生成本来是个比较麻烦的问题，但是运用马尔科夫链的数学统计模型知识重新进行表述后，可以使整个问题大大简化。

## 任务概述

本实验通过对一段输入文本进行分词、词频分析、构建散列表数据结构后，采用适当的随机方法，自动生成并输出一段与原文章在词法、语义、主题上近似的文段。

## 数据描述

数据以.txt文件形式保存，UTF-8进行编码，是英文形式的自然语言字符串。内置文学主题包含：哈利波特七部曲、银河系漫游指南；同时接受用户自定义的输入文章。

## 功能需求

用户可以选择使用内置的演示文档和自行上传文档，实现对两类文档的文本解析，生成相应数据集之后，自动返回生成的结果文本到网页上。

## 运行需求

开发环境：Python=3.9、Flask==2.0.2

开发工具：Pycharm 2021.3(Professional Edition)、Visual Studio Code、Postman

## 任务计划

后端

1. 解析文本：利用马尔可夫链算法，对已有文本进行分词，统计频率
2. 语段生成：根据解析的文本，进行随机文本生成
3. 返回前端：将生成的文本输出在网页上

前段

1. 设计网页，完成选择、上传、提交等功能

第二章 概要设计

## 总体设计

## 主要技术问题

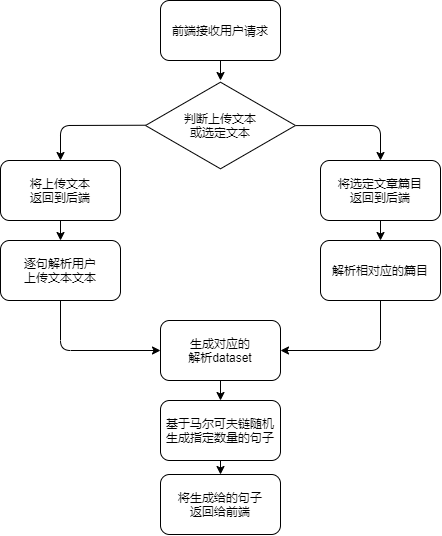
1. 对于输入文章的格式化与文章句段的切分问题
2. 对于汉语的分词问题
3. 在语段分析后，建立高效读写数据结构的问题

## 运行界面设计

# 第三章 详细设计

## 系统设计与实现

总体框架



接口设计

1. 获取生成文字

1）功能说明

通过选定的主题和句子数量生成相应

2）接口说明

请求方式：POST

请求地址：http://ip:port/login

请求参数： 无

3）请求参数

示例

127.0.0.1:5000/login

4）响应参数

|  |  |  |
| --- | --- | --- |
| 参数名 | 类型 | 说明 |
| result | string | 返回生成的文本 |

示例

{

"result": "Relaxing in his manic desperation and covered with white furry creature bounded through think it's could neither judge nor have died this bobbing, glistening in the din "Most of muscle movements…"

}

## 模块设计

前端

后端

解析文本模块

解析收到的txt格式文本文件，首先消除文本中的\n和类似“Page”字样，然后利用正则表达式将文本根据文本结束符（句号、感叹号、问号等）分隔句子。对每个句子添加”^”作为句首标记，”$”作为句尾标记，循环遍历每个句子，对每个单词和其后的单词，如果这样的组合从未出现，则将其放入字典中，并将count设置为1；如果这样的组合已经出现，则将对应的count加1。这样可以得到一个以字典为基础的数据集，

生成文本模块

针对生成的数据集，设定一个句子的句首必须为”^”，以此为首单词，根据马尔可夫链的思想，生成下一个单词，随机生成的过程如下：首先，对某一单词的下一单词列表的count值进行遍历求和，得到下一单词出现的频数总和，生成一个随机数范围在1到频数总和之间，继续遍历下一单词的字典，得到下一单词，直到生成的下一单词为”$”，停止生成。

【各模块的具体设计与实现】

# 第四章 测试分析

【程序运行展示】

【测试方法、发现的问题及解决方案】

【对最终成果的评价】

# 第五章 用户手册（可选）

# 基于马尔可夫链的写作机器人

A Writing bot based on Markov Chain

该写作机器人基于马尔可夫链设计，利用flask框架在前端进行演示。

马尔可夫链是随机过程中的概念，基于该算法可以根据已有的文本生成一些看似真实的文字，每生成一个词，它可以根据下一个词出现的频数随机选择下一个出现的字，直到把他们连成一个句子。

### Feature

* 共有六篇文章/小说作为演示文档，其中三篇中文文章，四篇英文文章
* 用户可以上传.txt文件，在线进行文章解析和数据集的生成
* 用户可以选定生成的文章句子数量

### Usage

要使用该程序，讲文件克隆到本地后，使用该命令，配置好环境

pip install -r requirements.txt

使用该命令进入演示网页

python main.py

在网页中，你可以在表单中选择我们提供的演示文章，或者自己上传本地的.txt文件进行解析。接着填写需要生成的文章的句数，点击生成即可生成由马尔可夫链随机生成的文章。

# 小结

【设计初衷、实践过程中遇到的困难及解决方案、课程收获等】

# 附录1 工作量及成绩自评表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **班级** | **学号** | **姓名** | **工作量简述**  **（代码量）** | **成绩自评** |
| 计科A2001 | 2019080184 | 邓岳 |  | A+ |
| 计科A2001 | 2020020360 | 熊圳 |  | A+ |
| 计科A2001 | 2019010025 | 赵千慧 |  | A+ |
| 计科A2001 | 2017060315 | 赵凯 |  | A+ |
| 备注：  　1、成绩自凭栏可选填：A+、A、A-、B+、B、B-、C+、C、C-、D、F | | | | |

# 附录2：程序代码（仅电子版需要、纸质版不需要）