基于机器学习的英语单词智能打卡系统

摘 要

随着人工智能和机器学习技术的发展，基于机器学习的英语单词学习应用程序兴起。这些应用程序可以自动检测学习者的语音或笔迹输入，给予即时反馈与评分。本研究设计并实现了一套基于机器学习的英语单词智能打卡系统。学习者可以通过输入单词，系统根据相关度对语料库中类似的单词和例句进行检索，给予结果并在系统中记录学习者的学习记录。相比现有冗余的英语学习应用，本系统提供更加个性化和简便，有潜力成为英语学习领域一款有价值的人工智能产品。本研究为基于机器学习的英语学习应用提供了一种新的设计思路。未来，在系统功能和模型的不断优化下，这一基于机器学习的英语单词学习方法有望被应用于更广阔的学习场景与群体中。机器学习技术可以实现英语学习内容的自动生成与优化，这一设计思路为开发人工智能英语学习工具提供了全新的视角，有望在提高学习效率与体验的同时，进一步推动教学方法的革新。

关键词　机器学习；单词打卡；自然语言处理

A Machine Learning Based Intelligent Clock in System for English Words

**Abstract**

With the growth of artificial intelligence and machine learning technologies, machine learning-based English word learning applications have emerged. These apps can automatically detect learners' speech or handwriting input, giving instant feedback and grading. In this study, we design and implement a machine learning-based intelligent English word punching system.Learners can input words by handwriting, and the system searches similar words and example sentences in the corpus according to the degree of relevance, gives the results and records the learners' learning records in the system. Compared with the existing redundant English learning applications, this system provides more personalization and simplicity, and has the potential to become a valuable artificial intelligence product in the field of English learning. This study provides a new design idea for machine learning-based English learning applications. With the continuous optimization of system functions and models in the future, this machine learning-based English word learning method is expected to be applied to a wider range of learning scenarios and groups. With the machine learning technology enabling automatic generation and optimization of English learning contents, this design idea provides a new perspective for developing artificial intelligence English learning tools, which is expected to further promote the innovation of teaching methods while improving learning efficiency and experience.

**Keywords　Machine learning, Word punch, Natural language processing**

目 录

摘要 I

Abstract II

[第1章绪论 1](#_Toc7400)

[1.1课题背景 1](#_Toc3488)

[1.1.1研究意义 1](#_Toc19121)

[1.1.2国内外研究现状分析 3](#_Toc2500)

[第2章相关技术介绍 4](#_Toc15845)

[2.1Python 4](#_Toc4323)

[2.2Django 4](#_Toc16244)

[2.3机器学习 5](#_Toc7185)

[2.4Mysql 6](#_Toc28798)

[第3章需求分析 7](#_Toc2782)

[3.1功能需求 7](#_Toc12663)

[3.2非功能需求 8](#_Toc26223)

[3.3需求模型 8](#_Toc28526)

[3.4可行性分析 10](#_Toc13867)

[第4章系统设计 11](#_Toc28794)

[4.1系统的基本数据 11](#_Toc15087)

[4.1.1实体关系 11](#_Toc16535)

[4.1.2数据设计 13](#_Toc19703)

[4.2系统的概要设计 16](#_Toc10882)

[4.2.1功能模块 16](#_Toc14798)

[4.3系统的详细设计 19](#_Toc17489)

[4.3.1架构设计 19](#_Toc11015)

[4.3.2模型设计 22](#_Toc5586)

[4.4系统的核心算法 26](#_Toc9113)

[4.4.1词根匹配 26](#_Toc20226)

[4.4.2自然语言处理 28](#_Toc20733)

[4.5本章小结 29](#_Toc30487)

[第5章系统编码及运行 30](#_Toc5780)

[5.1系统开发涉及的软件 30](#_Toc239)

[5.2系统运行界面及结果 30](#_Toc11480)

[第6章系统测试 36](#_Toc22499)

[6.1测试分析 36](#_Toc15197)

[6.1.1测试目的 36](#_Toc20561)

[6.1.2对比测试 36](#_Toc19387)

[6.2结果分析 40](#_Toc348)

[6.3本章小结 40](#_Toc9421)

[结论 41](#_Toc31148)

[致谢 42](#_Toc22767)

[参考文献 43](#_Toc3894)

# 第1章绪论

## 1.1课题背景

随着人工智能和机器学习技术的发展，基于机器学习的英语学习应用程序兴起，如无人英语老师和AI英语家教等。这些应用程序可以自动检测学习者的语音或笔迹输入，给予即时反馈与评分，实现个性化和高效的学习。相比传统英语学习方法，这些新应用程序有望极大优化学习体验，激发学习兴趣，提高学习效果。但是，现有英语学习应用程序仍存在一定局限。大多系统需要人工录入大规模的单词、词性、例句与翻译，无法实现学习内容的自动生成，增加了开发工作量，难以扩展。另外，现有系统主要基于规则引擎与统计机器翻译，学习内容与效果难以做到高度个性化，无法充分满足不同学习者的需求。基于此，研究开发一套基于机器学习的英语单词智能打卡系统具有重要意义。通过对海量公开语料的采集、清洗与向量化，机器学习模型可以实现学习内容的自动生成与优化。与规则引擎和统计方法相比，机器学习模型可以建立单词语义空间，提供高质量的近义词、例句与翻译，实现学习内容的高度个性化。系统可以实现海量英语语料的下载、清洗与学习内容的自动生成，提供个性化的学习反馈与评分，优化学习体验与效果。研究成果有望为英语学习应用的开发与优化提供新思路，推动机器学习在英语教学领域的应用与发展。本课题旨在通过机器学习技术，研发一套基于机器学习的英语单词智能打卡系统。系统可以实现学习内容的自动生成与优化，提供高度个性化的学习体验，这一设计思路将为基于人工智能的英语学习工具开发开拓新视野，有望提高学习效率与学习效果，为英语教学方法的革新与发展注入新动力。

### 1.1.1研究意义

许多专家学者研究表明，词汇量与语言能力及认知能力密切相关。如美国著名语言学家诺曼·莱文斯指出:“词汇是思维的工具，词汇是思想的武器[1]。”我国语言学家桂诗春也研究发现，词汇量大的人语言水平不一定高，但语言水平高的人词汇量都比较大；词汇量测试和智力测试高度相关。由此可见，记忆词汇是学习英语最基本的要求。只有不断丰富英语词汇，听、说、读等语言能力才能提高。通过单词打卡系统，学习者可以随时随地学习记忆词汇。系统分析学习者的记忆遗忘规律，在关键时间点智能提醒，将短期记忆转化为长期记忆，从而提高学习效率。 总之，专家学者的研究结果显示，词汇量是反映语言与认知能力的关键指标。要提高英语水平，词汇量的积累至关重要。单词打卡系统可持续监测学习者的记忆情况，在关键时间点开展干预，使词汇从短期记忆进入长期记忆，这大大提高了词汇记忆的效率，对英语水平的提高起到关键作用。通过这一系统，学习者可以有效记忆词汇，为全面提高英语能力奠定基础。

本选题“基于机器学习的英语单词智能打卡系统”中的两个关键词为“机器学习”与“智能单词打卡”。

机器学习是一个跨学科融合的领域，它汇集了概率论、统计学、近似理论和复杂算法等诸多学科的知识[2]。机器学习运用计算机作为工具，致力于模拟人类学习的方式，对知识进行结构化处理，以提高学习效率。它的研究对象是如何使用计算机模拟或实现人类的学习活动，这也是人工智能领域最具智能特征和最前沿的研究方向之一[3]。本系统在打卡任务生成、打卡评判两部分应用到机器学习、自然语言处理等技术对打卡任务进行处理。

本研究开发的系统，实现了英语词汇学习。该研究探索了一种以学习者为中心、打卡式学习以及自我控制学习进度的新模式[4]。此外，本文通过提供基于具体语境的初级词汇打卡学习，并使用词义之间的聚合将词汇分类，有助于学习者构建系统的词汇结构，使其能够更好地学习词汇[5]。软件推出后，学习者可以选择不同的方式学习英语词汇，这大大提高了词汇学习的方便性以及记忆效率。软件利用大数据分析，找到背诵单词的效率规律和遗忘曲线，然后根据曲线来规划记忆时间，安排单词出现的几率，确保背诵单词的最佳效率。

#### 1.1.1.1理论或实际应用方面的价值

（1）理论意义

随着互联网技术的发展,越来越多的学习者开始接受基于互联网的学习方式[6]。本研究通过设计和实现一个词汇学习平台,进一步明确了互联网学习的概念,有助于互联网学习理论的进一步发展和完善。

（2）实践意义

本研究通过实现一种基于机器学习的英语词汇学习方式，探索了一种以学习者为中心、自我控制学习进度以及个性化学习的新途径[7]。此外，本文还提供了基于具体语境的初级词汇学习，并通过词义的聚合将词汇分类，从而帮助学习者构建系统的词汇结构，达到更好地学习词汇的目的。最后，本系统主要针对二年级以上大学以下学生，对于中小学生高效率的提升词汇量，具有一定的现实意义与价值。

### 1.1.2国内外研究现状分析

#### 1.1.2.1国内研究现状

伴随着社会的发展，互联网网络技术的持续创新与发展，教育技术的改革与深化，导致了各种各样的网络学习方式和平台的出现，并将其应用于教育培训组织、学习、工作和生活等方面的软件也在持续增多。为了适应人们生活、工作、学习等各方面的需要，科技工作者也在不停地进行着软件的开发与升级。英语作为国际交流中最主要的语言，同时，它也是我国考生复习、准备考试的重点，从而影响到他们的教学策略与学习方法。在现代社会,词汇学习是掌握一门语言的基础。在当今网络发达的社会中,单词软件学习成为学习英语词汇的主要方式之一[8]。在这种社会背景下,单词软件知识学习的改革正在实际的教学和研究活动中发生。教育科研人员早就关注移动学习在教育领域的应用,其中语言学习是移动学习领域的重要研究方向之一。

英语词汇学习的传统工具是由词汇手册、教科书、字典等纸面资料构成的。这种方法的局限性是学习者只能按词汇的次序进行直线式学习，不能与学习者互动，也不能针对学习者的特征做出相应的调整与反馈[9]。但是，随着词汇教学软件的普及，英语词汇教学软件的数量也越来越多，其发展趋势是能够按照“遗忘规律”的规则来对海量词汇进行准确的复习。

#### 1.1.2.2国外研究现状

美国斯坦福大学成立了世界上第一个移动语言教学实验室，这标志着国外移动语言学习研究的开展。在2000年,他们开始了一项关于语言学习的重要项目,他们通过手机进行西班牙语学习的研究,结果证明使用单词软件功能可以明显提高语言学习效果[10]。

随后，世界各地都开始了对手机语言学习的研究。在日本，Thornton和Houser[11]对英语单词的实际教学做了一次调查。他们每天三次用手机发三条简明扼要的课程学习内容给学生。本研究发现，使用短信息辅助词汇学习的学生，其英语学习成效明显优于使用常规教学法的学生。

#### 1.1.2.3总结对比分析

移动语言学习是一种不受时间空间与限制的信息化学习方法，它不但可以为学习者提供个性化的学习方案，而且还可以把传统的课堂学习延伸到人们的生活中去，并且可以对学习者进行及时的反馈，从而提升学习效果[12]。由于其在语言学习上的诸多优势，因此，移动教学被认为是一种最受欢迎的教学方法。随着科技的日臻完善，移动语言教学与英语教学的关系将会日益紧密，两者的发展前景将会更为广阔。

# 第2章相关技术介绍

## 2.1Python

Python是一种高级编程语言，由Guido van Rossum于1991年创建[13]。Python的设计目标是简单、易读、易学，同时具有强大的扩展性和可嵌入性。Python以其清晰简洁的语法、丰富的内置库和第三方库，以及友好的开发环境等特点，成为当今最受欢迎的编程语言之一[14]。Python的特点包括：

1. 简单易学的语法 Python的语法设计简单明了，易于掌握和理解。Python通过使用缩进而不是花括号进行代码块分割，从而使代码更加美观、整洁。Python还提供了丰富的标准库和第三方库，使得开发者能够轻松实现复杂的功能和应用。
2. 面向对象编程 Python是一种面向对象编程语言，支持类、封装、继承、多态等面向对象编程概念[15]。这使得Python能够更好地组织和管理代码，并提高了代码的可重用性和可维护性。
3. 开源的免费软件 Python是开源的，任何人都可以免费使用和修改Python的源代码。这意味着Python社区非常活跃，有着大量的资料和支持工具，例如在线文档、社区论坛和开源项目等。
4. 丰富的标准库和第三方库 Python拥有丰富的标准库和第三方库，包括数值计算库NumPy、Pandas数据处理库、matplotlib绘图库和机器学习框架TensorFlow等。这些库为Python提供了强大的功能和扩展性，使得Python可以在多个领域广泛应用。
5. 跨平台兼容性 Python是一种跨平台语言，可以运行在多个操作系统上，例如Windows、Linux、MacOS等。这使得Python成为一个非常灵活和通用的编程语言，可以在不同的平台上实现相同的功能。

综上所述，Python是一种简单易学、面向对象、开源免费、拥有丰富库支持、跨平台兼容性强的高级编程语言。在各种应用场景中都具有广泛的应用前景。

## 2.2Django

在本文系统中我们采用了Django框架来实现系统的开发和部署。

Django是一个基于Python的Web框架，它提供了完整的MVC模式支持和一系列易用的API，使得开发人员可以快速构建高效、可扩展、易维护的Web应用程序。我们选择Django框架作为系统的开发和部署平台，主要基于以下几个方面的考虑：

1. 强大的模板引擎和表单处理功能 Django提供了强大的模板引擎和表单处理功能，使得我们可以轻松地实现前端页面的设计和交互功能。
2. 简单易用的ORM框架 Django的ORM框架能够将数据库操作抽象成Python类的操作，并提供了丰富的查询API，使得我们可以快速地实现数据管理。
3. 优秀的安全性和可靠性 Django框架具有优秀的安全性和可靠性，可以有效地保护用户数据和隐私。
4. 社区活跃度高 Django拥有庞大的社区和广泛的应用范围，在系统开发和部署过程中，可以快速地获取到各种优秀的开源库和技术支持。

## 2.3机器学习

机器学习是人工智能领域的一个分支，旨在让计算机通过学习从数据中发现规律和模式，并利用这些规律和模式来进行预测和决策。机器学习的主要目的是建立一个可以自动适应数据的模型，使得这个模型可以不断地改进其性能和准确度。它的主要特征是:

1. 使用数据来训练算法。机器学习利用大量的数据来训练算法模型,使其能够学习和预测。这不同于传统的专家系统,后者依赖人工设计的规则。
2. 算法会随着数据的增加自动优化。机器学习算法可以自动检测复杂的数据模式,并根据数据不断优化模型,而无需人工干预。
3. 可用于预测和分类。机器学习可以用于对未来事件作出预测和判断,也可以用于对数据进行分类归组。这使其在许多领域都有广泛的应用。
4. 包括监督学习和无监督学习。监督学习使用标注数据来训练算法,无监督学习不需要标注数据而是发现数据中的隐藏模式。这两种学习方式可应用于不同的任务。
5. 常用算法包括神经网络、决策树、SVM等。这些算法可以应用于回归、分类、聚类、降维等不同类型的机器学习问题。
6. 需要大量的数据。机器学习的性能取决于使用的数据量和质量。较少的数据会导致"过拟合"，无法准确地推广到新数据。
7. 不需要预先编程规则。机器学习可以直接从数据中学习和检测模式,不需要人工预先定义规则。这使其能够解决复杂而不规则的问题。

## 2.4Mysql

MySQL是一种开源的关系型数据库管理系统（RDBMS），最初由瑞典MySQL AB公司开发，现在属于Oracle公司旗下产品[16]。MySQL广泛应用于各种大型Web应用程序中，并且在小型个人网站和企业级数据仓库等场景中也受到欢迎。MySQL具有以下特点：

1. 可靠性：MySQL支持数据的完整性和安全性、备份和恢复、事务处理等高可靠性和可靠性功能。
2. 可扩展性：MySQL支持分布式数据库架构，能够轻松地扩展为更大的应用程序。
3. 可移植性：MySQL可以运行在多种操作系统上，例如Windows、Linux、MacOS等，具有良好的可移植性。
4. 性能优越：MySQL的设计和实现使其具有出色的性能，可以应对各种高负载环境。
5. 开源免费：MySQL是一款开源软件，用户可以免费使用和修改其源代码，而且拥有庞大的社区支持。

MySQL支持多种编程语言，例如C、C++、Java、PHP、Python等，同时还提供了多种客户端工具，例如MySQL Workbench、phpMyAdmin等来方便用户管理和操作数据库。MySQL不仅具有广泛的应用领域，还具有强大的扩展性和可定制性，因此是一个非常优秀的数据库管理系统。

# 第3章需求分析

## 3.1功能需求

该毕业设计是开发一个智能单词打卡系统，该系统主要有四个功能模块，分别是注册登录模块、打卡任务模块、遗忘曲线分析模块、单词导入模块。用户通过注册获得账号，然后进行登录，可以进行个人信息修改，在学习任务中主要功能有系统自动生成打卡任务，查看用户学习情况，遗忘曲线分析。在遗忘曲线分析模块中，主要的功能有根据用户单词记忆情况分析出记忆遗忘曲线，根据曲线遗忘程度着重安排单词出现频率，最大程度的提升背单词效率以及单词的记忆程度。单词导入模块中，用户可以自行选择要学习的单词，以单词本的形式导入，导入的单词自动解析，与系统数据库中的单词进行对比，去除重复词语，并自动生成中文例句填词处理。根据需求，制作系统功能模块划分结构图，如图3-1所示：



图 3-1系统功能模块划分结构图

各个模块功能如下：

1. 注册登录模块：

用户通过注册获得账号，通过账号进行登录使用系统，用户登录后要进行身份认证，通过账号密码等信息验证用户的身份，不同的身份所展示的效果不同，不同的身份都可以进行个人信息修改。

1. 打卡任务管理：

登录单词打卡系统后，可以学习今日的单词任务以及昨日单词的复习，也可以查看自己的遗忘曲线，每日的单词学习情况，任务完成后后台自动评判产生成绩。

1. 遗忘曲线分析模块：

登录单词打卡系统后，可以查看自己的遗忘曲线，单词每隔一段时间会再一次出现，根据完成情况系统分析出遗忘曲线，根据遗忘曲线安排后续单词出现的频率，最大程度上保证背单词的效率以及成功率。

1. 单词导入模块：

在这个模块中，用户可以自行选择要学习的单词，以单词本的形式导入，导入的单词自动解析，与系统数据库中的单词进行对比，去除重复词语，并自动生成中文例句填词处理。

## 3.2非功能需求

1. 性能需求：用户期望该软件在响应速度、结果精度和资源消耗等方面表现良好。
2. 可靠性需求：用户要求该软件尽可能不失效，并要求易恢复性，系统故障易预测且严重程度尽量降低。
3. 易用性需求：该软件需要具备易理解、易学习、易操作等特点，人机界面应简单明了，用户无需手册即可顺利进行各种操作。
4. 运行环境约束：用户对软件运行的环境有特殊要求。
5. 外部接口：用户要求该软件与其他软件系统或硬件设备具有互通性。
6. 安全性需求：该软件需要保证用户的重要信息安全，不同用户只能查看自己相关信息，具有最高权限的管理员才可修改系统，系统需提供数据备份及意外崩溃时数据恢复等功能。
7. 可保障性需求：用户要求该软件具备可配置、可扩展、可维护、可移植等特点。

## 3.3需求模型

1) 用户具备的功能有：登录、注册、个人信息修改、完成打卡任务、搜索单词，导入单词本，查看遗忘曲线，单词学习情况。

用户用例图如图3-2所示：



图3-2用户用例图

2）系统用例图如图3.3所示：



图3-3系统用例图

## 3.4可行性分析

1. 技术可行性：机器学习方法已经广泛用于自然语言处理，其在单词分类、预测和翻译方面的表现已经超越了人类。因此，我们可以使用机器学习算法来实现基于英语单词的智能打卡系统。
2. 数据可行性：目前，大量的英语单词语料库和单词库已经公开发布，可以使用这些数据来训练和优化机器学习模型，同时也可以通过数据挖掘和爬虫等技术来扩充和更新语料库和单词库。
3. 实施可行性：基于机器学习的英语单词智能打卡系统可以英语学习者提供方便、快捷、高效的学习工具，具有很高的市场需求和潜在用户群体。
4. 安全可行性：在系统的开发过程中，可以使用各种安全技术和措施来保护用户数据的安全和隐私，例如数据加密、用户认证、访问控制等，以确保系统的安全性和可靠性。
5. 综上所述，基于机器学习的英语单词智能打卡系统技术方案具有很高的可行性，可以为用户提供优质的学习体验和商业价值，同时也可以促进机器学习技术在教育领域的应用和发展。

# 第4章系统设计

## 4.1系统的基本数据

本系统的数据采用关系型数据库mysql 进行数据的存储。存储的信息包括各种事务的 id（包括分享事务，关注 id 以及用户的全局 id），用来唯一标识个体及其事务、按照用户 id 索引的用户信息（包括用户注册邮箱、 用户名以及用户密码）、按照用户名进行索引的用户信息（包括用户 id）、用户存储的单词本（包括用户 id 以及用户的单词本内容）、导入的单词本、单词发音、单词例句。将上述存储的信息设计为 mysql 中的表格，如下进行详细说明：gid 表主要用来保存各种事务和个体的 id，用来唯一识别。这里的表分别存储单词本的 id，用户的 id，以及事务的 id，对应只在同名的列限定符上取值。

### 4.1.1实体关系

在设计数据库之前，需要明确数据实体之间的关系，以便建立正确的数据库模型[17]。 数据实体之间的关系是指不同实体之间的联系和依赖性，这些联系可以是一对一、一对多或多对多等形式。正确建模实体关系可以有效地提高数据库运行效率、减少数据重复和冗余，并帮助开发人员更好地理解数据之间的联系。

本系统主要使用python开发，使用mysql来存储系统使用的数据，使用Django快速高效地开发出高质量、易于维护的Web应用程序。搭建的系统数据库实体关系图如图4-1所示。

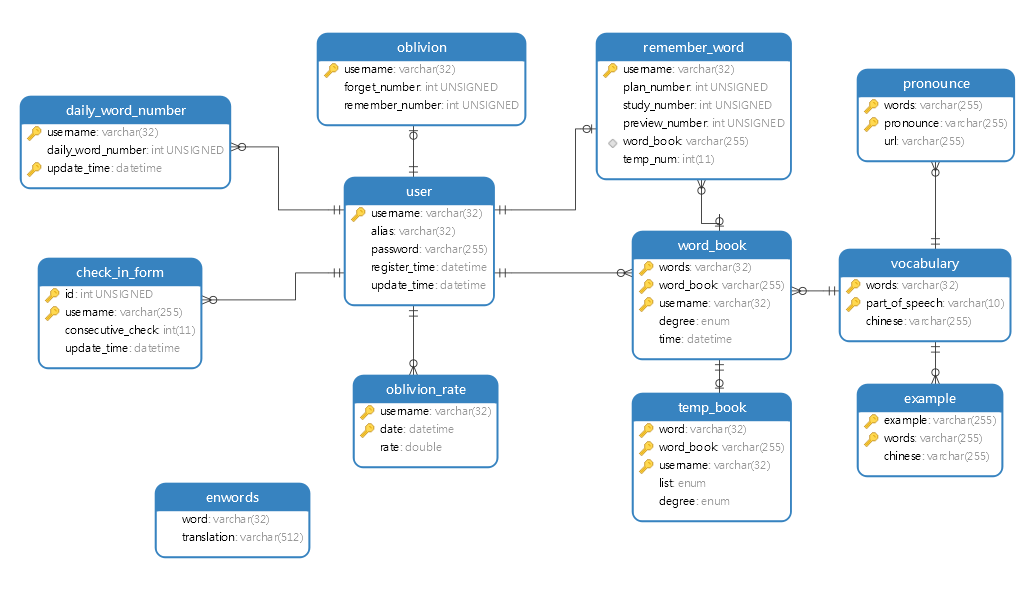


图4-1数据库实体关系图

Django搭建的框架在数据库中所对应的实体关系图如图4-2所示

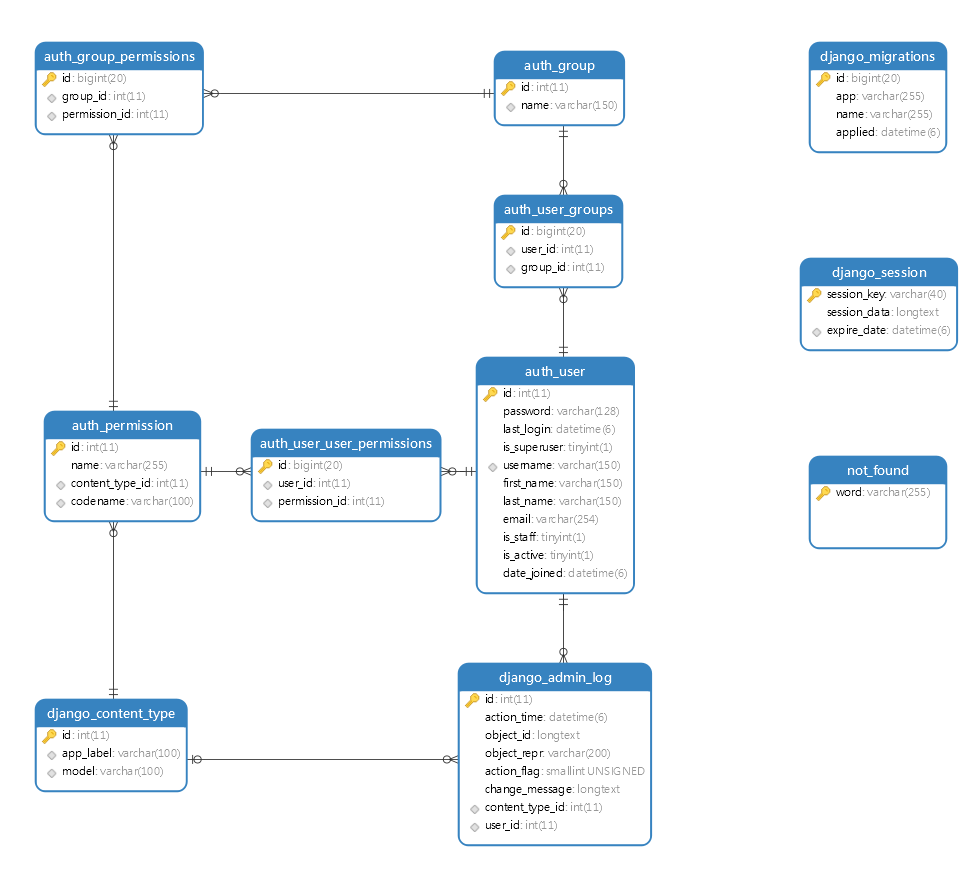


图4-2 Django实体关系图

### 4.1.2数据设计

该数据库包含以下表格：

1. Users表: 该表存储系统的用户信息，包括用户名、用户别名、密码、注册时间、更新时间。

user 表主要以用户username为索引的行键，列限定符包括 alias、password、register\_time、update\_time，都在 user 的列族之下，主要用来按用户 username索引信息使用。

表4-1 用户信息表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| username | varchar | 32 | 0 | 0 | 1 | 用户名 |
| alias | varchar | 32 | 0 | 0 | 0 | 别名 |
| password | varchar | 255 | 0 | 0 | 0 | 密码 |
| register\_time | datetime | 0 | 0 | -1 | 0 | 注册 |
| update\_time | datetime | 0 | 0 | -1 | 0 | 更新 |

1. Oblivion表: 该表存储单词记忆情况的信息，包括用户名称、忘记的单词数量和记忆单词的数量等。

表4-2 单词记忆表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| username | varchar | 32 | 0 | 0 | 1 |  |
| forget\_number | int | 10 | 0 | 0 | 0 | 忘记数 |
| remember\_number | int | 10 | 0 | 0 | 0 | 记忆数 |

1. Daily\_world\_number:表：该表存储单词的记忆信息，包括每日记录的数量和更新时间等。

表4-3 单词遗忘比例表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| username | varchar | 32 | 0 | 0 | 1 |  |
| daily\_word\_number | int | 10 | 0 | 0 | 0 | 数量 |
| update\_time | datetime | 0 | 0 | 0 | 1 | 时间 |

1. Check\_in\_form表: 该表存储用户学习单词的连续性打卡记录，包括用户ID、连续打卡次数和更新时间等。

表4-4 打卡记录表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| id | int | 10 | 0 | 0 | 1 | ID |
| username | varchar | 255 | 0 | 0 | 1 |  |
| consecutive\_check | int | 11 | 0 | 0 | 0 | 连续次数 |
| update\_time | datetime | 0 | 0 | 0 | 0 | 时间 |

1. Oblivion\_rate表: 该表存储用户学习单词遗忘比例的记录，包括用户学习的天数以及遗忘比例等。

表4-5 遗忘比例表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| username | varchar | 32 | 0 | 0 | 1 |  |
| date | datetime | 0 | 0 | 0 | 1 | 天数 |
| rate | double | 0 | 0 | -1 | 0 | 比例 |

1. Enwords表: 该表存储单词翻译记录，包括单词以及翻译等。

表4-6 单词翻译表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| word | varchar | 32 | 0 | 0 | 0 | 单词 |
| translation | varchar | 512 | 0 | -1 | 0 | 翻译 |

1. Remember\_word表: 该表存储单词的学习情况，主要包括计划学习数量，学习数量，复习数量，单词本，暂存数量。

表4-7 模型表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| username | varchar | 32 | 0 | 0 | 1 |  |
| plan\_number | int | 10 | 0 | -1 | 0 | 计划 |
| study\_number | int | 10 | 0 | -1 | 0 | 学习 |
| preview\_number | int | 10 | 0 | -1 | 0 | 复习 |
| word\_book | varchar | 255 | 0 | -1 | 0 | 单词 |
| temp\_num | int | 11 | 0 | -1 | 0 | 暂存 |

1. Word\_book表: 该表存储单词难度的参数。

表4-8 参数表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| words | varchar | 32 | 0 | 0 | 1 | 单词 |
| word\_book | varchar | 255 | 0 | 0 | 1 | 单词 |
| username | varchar | 32 | 0 | 0 | 1 |  |
| degree | enum | 0 | 0 | 0 | 0 | 难度 |
| time | datetime | 0 | 0 | 0 | 0 | 次数 |

1. Temp\_book表: 该表存储暂存的单词，用于存储每日应该复习的单词但是复习的数量小于已有的数量，该部分单词存储在此表中。

表4-9 暂存表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| word | varchar | 32 | 0 | 0 | 1 | 单词 |
| word\_book | varchar | 255 | 0 | 0 | 1 |  |
| username | varchar | 32 | 0 | 0 | 1 |  |
| list | enum | 0 | 0 | -1 | 0 | 表格 |
| degree | enum | 0 | 0 | 0 | 0 | 难度 |

1. Pronounce表: 该表存储单词发音，用于学习者在学习单词时展示单词的音标。

表4-10 发音表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| words | varchar | 255 | 0 | 0 | 1 | -1 |
| pronounce | varchar | 255 | 0 | 0 | 1 | 发音 |
| url | varchar | 255 | 0 | -1 | 0 | 来源 |

1. Vocabulary表: 该表存储单词系统中的单词，以及后续导入的的单词本中解析出的单词都存储在此表中。

表4-11 词汇表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| username | varchar | 32 | 0 | 0 | 1 |  |
| alias | varchar | 32 | 0 | 0 | 1 | 别名 |
| password | varchar | 255 | 0 | 0 | 0 |  |

1. Example表: 该表存储单词的例句，用于学习者在学习单词时展示单词的例句，通过单词的名字与单词表进行关联。

表4-12 例句表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 字段名称 | 数据类型 | 长度 | 小数点 | 是否为空 | 是否主键 | 注释 |
| example | varchar | 255 | 0 | 0 | 1 | 例句 |
| words | varchar | 255 | 0 | 0 | 1 | 单词 |
| chinese | varchar | 255 | 0 | -1 | 0 | 中文 |

这些表共同支持基于机器学习的英语单词智能打卡系统的数据管理和分析功能，从而帮助用户更好地学习英语单词。

## 4.2系统的概要设计

### 4.2.1功能模块

通过建设基于机器学习的英语单词智能打卡系统，为用户提供一站式学习平台，掌握学习情况，保障学习的主动性的效率，实现以下目标：

通过设计单词打卡系统架构，形成基于网页的学习平台，实现学习任务有效与效率处理。用户访问使用单词打卡服务系统主要通过网页浏览器访问模式，方便快捷。为了方便用户使用，我们将所有数据存储在一个数据库中，并提供了一站式、安全可信、易于使用的个人账户和密码登录方式，以实现随时随地访问数据。学习平台。

#### 4.2.1.1设计目标

任何一个为了确保系统的高效运作，系统建设必须遵守多项设计原则，包括安全性原则、系统化原则、实用性原则、稳定性原则和可维护性原则等等[18]。在本项目中，根据具体的设计需求，我们提出了以下方案：

1. 提高学习效率：通过机器学习算法的应用，实现对学习者的单词掌握程度进行智能分析，从而提高学习效率。
2. 提高系统的智能化程度：通过采用机器学习算法，对学习者的学习情况进行全面分析，从而提高系统的智能化程度。
3. 提供个性化的学习方案：根据学习者的学习情况和单词记忆情况，提供个性化的学习方案和建议，从而更好地帮助学习者掌握英语单词[19]。
4. 提供多种学习方式：除了传统的单词记忆方式外，系统还应提供多种学习方式，例如单词拼写、听力理解等，满足不同学习者的需求。
5. 实现可扩展性：系统架构应该具有可扩展性，能够方便地添加新的学习功能和模块，以适应不断变化的学习需求。
6. 提供良好的用户体验：系统应该提供良好的用户界面和交互体验，方便用户使用，从而提高学习积极性。

#### 4.2.1.2系统功能

本系统的研发目的在于提高英语学习者单词记忆的效率与效果，帮助用户养成规律而持续的背诵习惯。为实现此目标，系统将采取积极主动的策略来推动用户学习，这一理念贯穿于系统的每一功能设计之中。基于此设计理念，系统的具体功能不会是简单被动的知识展示，而是要主动监测用户的学习情况，并在需要时进行针对性地提醒与推动。系统首先需要建立用户的学习模型，追踪用户的学习热度与关注词汇，预测用户今后一段时间内的学习需求。其次，系统要根据用户模型定期推送合理的学习计划给用户，并在学习过程中随时给出提醒， 维持学习的连续性与强度。再次，系统还会根据用户的实际学习表现不断修正用户模型，实现模型的动态优化，以保证提醒与推送的准确性。系统的主要功能如下：

1. 用户管理功能：该功能提供用户注册、登录和个人信息管理等功能，以便用户可以使用系统并查看自己的学习记录。

用户注册和登录模块主要实现用户的登录、注册。登录需要输入用户名和密码，如果信息输入正确则跳转到主页，如果信息输入错误则跳转到登录页面；注册需要填写用户信息，如果信息输入正确，则返回登录页面，如果错误则提示注册失败。用户登陆注册流程图如图4-3 ，4-4所示：



图4-3 用户登录流程图



图4-4 用户注册流程图

1. 单词输入功能：该功能允许用户输入要学习或使用的英语单词，并将其传递给单词检测模块进行检测。
2. 单词检测功能：该功能使用机器学习模型和规则引擎检测用户输入的英语单词是否被正确使用。
3. 单词定义功能：该功能允许用户查看英语单词的定义、用法和语境，以便用户更好地理解和使用单词。
4. 学习记录和统计功能：该功能记录用户输入的单词和检测结果，并提供单词学习历史记录和统计信息，例如单词学习次数、错误率等。
5. 数据收集和预处理功能：该功能负责从不同来源收集英语单词数据，并将其进行预处理，例如清理数据、去除停用词、进行词干提取、词向量化等[20]。

这个基于机器学习的英语单词智能打卡系统的功能结构设计可以帮助用户学习英语单词，并提供错误提示和纠正建议，从而提高英语写作和口语的准确性和流畅性。同时，它可以为学习者提供更好的学习记录和统计信息，以便他们更好地掌握自己的学习进度和成果。

#### 4.2.1.3系统技术架构

本系统的目的在于通过机器学习技术为英语学习者提供良好的学习计划，全面提高单词记忆的效率与效果。为实现此目的，系统必须具备动态理解每个用户的学习情况并进行主动推动的能力。这就要求系统基于海量学习数据构建用户模型，并持续优化这些模型以追踪用户的最新学习状态。

系统技术架构设计图通常包括以下组件：

用户界面：提供给用户进行单词打卡、学习进度追踪、单词查询等功能的图形用户界面（GUI）。

数据库：存储单词数据、用户信息、学习记录等数据的数据库。

后端服务器：提供基于机器学习的单词学习算法，包括单词推荐、学习计划生成等功能的服务器端应用。

前端服务器：处理用户请求，与后端服务器进行交互并返回数据给用户的服务器端应用。

机器学习模块：使用机器学习算法进行单词推荐、学习计划生成等功能的模块。

系统技术架构设计图如图4-5所示：



图4-5系统技术架构

## 4.3系统的详细设计

### 4.3.1架构设计

#### 4.3.1.1总体架构

系统总体架构自顶向下主要包括４个层次，包括：

1. 数据层：该层负责管理数据的存储和处理，包括单词库、用户数据等。
2. 处理层：该层负责对用户的操作进行处理，包括用户的登录、注册、单词学习等功能的处理。
3. 机器学习层：该层负责训练和应用机器学习模型，对用户进行单词学习进度的预测和推荐。
4. 应用层：该层是整个系统的核心层，负责将机器学习层的推荐结果反馈给用户，并提供用户操作的界面。

系统总体架构图如图4-6所示：



图4-6系统总体架构

#### 4.3.1.2功能结构设计

系统工作流程为主要如下，用户进行登陆注册，系统自动生成学习任务，用户完成每日的学习任务和昨日学习单词的复习，系统根据用户背单词情况好坏安排后续的单词出现频率，还能看到自己的单词完成情况，以及错误率。系统工作图如图4-7所示：



图4-7 系统工作流程

1. 数据收集模块：该模块负责从不同来源收集英语单词数据，并将其存储到数据库中。可以使用公共数据集，或从用户生成的数据集中获取数据。
2. 数据预处理模块：该模块负责对收集的英语单词数据进行预处理，例如清理数据、去除停用词、进行词干提取等。这些预处理步骤可以提高后续模型训练的效果。
3. 模型训练模块：该模块负责使用神经网络和支持向量机等机器学习算法，对预处理后的数据进行训练，以学习单词的定义、用法和语境。可以使用已有的预训练模型，例如GPT-2、BERT等。
4. 单词检测模块：该模块负责接收用户输入的英语单词，并使用训练好的模型检测单词是否被正确使用，并提供错误提示和纠正建议。可以使用模型预测或规则引擎进行单词检测。
5. 用户界面模块：该模块负责提供用户界面，以便用户输入单词和查看错误提示和纠正建议。用户界面可以是Web应用程序、移动应用程序。具体的页面展示流程图如下图4-8所示：



图4-8 页面展示流程图

用户对于系统的操作就是学习任务的完成，其他的操作也同理，用以下的数据流图4-9来表示。



图4-9 文件操作数据流图

1. 数据存储模块：该模块负责存储用户输入的单词、模型训练等重要数据。可以使用关系型数据库进行数据存储。

总体而言，这个基于机器学习的英语单词智能打卡系统的架构设计可以帮助用户学习英语单词，并提供错误提示和纠正建议，从而提高英语写作的准确性和流畅性。

### 4.3.2模型设计

背单词基于 DHP 记忆模型，计算了不同复习反馈[时间序列](https://www.zhihu.com/search?q=%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":2758395455}" \t "https://www.zhihu.com/question/451367906/answer/_blank)下最优的复习时机。其优化目标是最小化每一位学习者记忆单词的时间成本。

在记忆算法中，每个用户的每个单词都有一个记忆稳定性来刻画该单词的遗忘速率。对于不同难度、不同记忆稳定性的单词而言，其未来复习时间成本的期望是不同的。抽取复习单词，是根据每个单词的最优复习时间确定的。在最优复习时间进行记忆，未来的复习时间成本最小。

当你对单词反馈“模糊”、“认识”或者“忘记”时，该单词的复习时间序列会更新，并重新计算记忆稳定性，然后根据模型安排最佳复习间隔。

每[个人记忆曲线](https://www.zhihu.com/search?q=%E4%B8%AA%E4%BA%BA%E8%AE%B0%E5%BF%86%E6%9B%B2%E7%BA%BF&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":2758395455}" \t "_blank)是根据所有已学单词的记忆稳定性计算而来的。每个记忆稳定性对应一个单词的[遗忘曲线](https://www.zhihu.com/search?q=%E9%81%97%E5%BF%98%E6%9B%B2%E7%BA%BF&search_source=Entity&hybrid_search_source=Entity&hybrid_search_extra={"sourceType":"answer","sourceId":2758395455}" \t "_blank)，对这些遗忘曲线进行加权平均，即可得到个人整体记忆曲线。

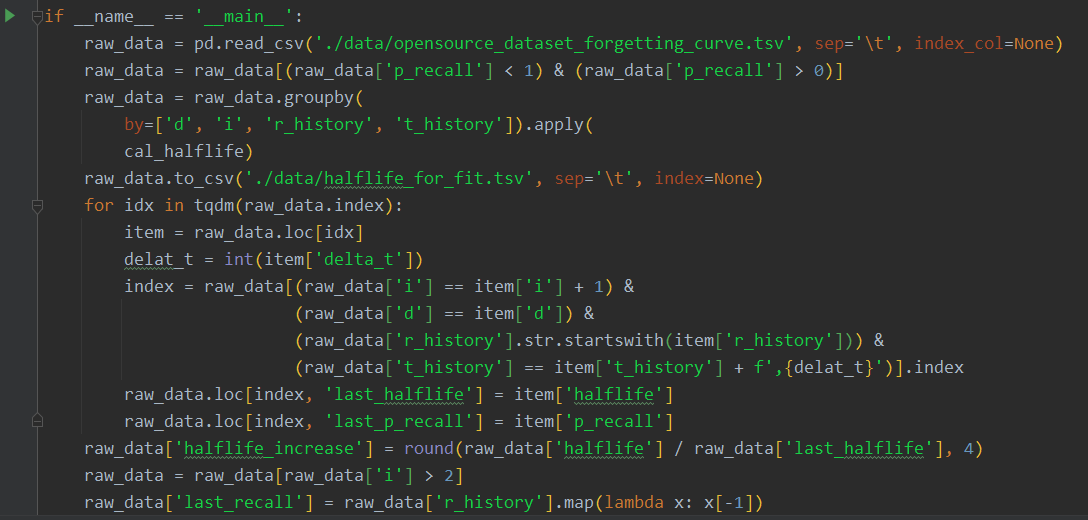
#### 4.3.2.1模型选择

随着在线学习平台的普及，大量关于学生复习的数据得以收集。这使研究人员能够设计出可训练、自适应、有保证的算法。高效的间隔重复调度算法可以节省数百万用户的时间，并帮助他们记住更多的单词。最近，一些研究采用机器学习来确定最佳复习时间。然而，由于以下三个原因，这些普遍方法并不适用于我们的情况：

* 缺少时间序列信息：一些研究，如 HLR（半衰期回归）模型和 EFC（指数遗忘曲线）模型，使用遗忘曲线将回忆概率与上次复习后的时间联系起来。但他们忽略了时间间隔对记忆强度的影响。在他们的模型中，记忆强度是关于重复次数的函数。根据间隔效应和我们收集的数据，复习间隔对长期记忆的形成过程有很大影响。
* 缺少用户感知的优化目标：HLR 和基于它改进的 C-HLR 模型，旨在准确预测记忆的回忆概率。而用户更关心复习压力、记忆效率等指标。
* 缺少可解释性：一些基于深度强化学习的算法对设计者来说是黑盒子，缺乏可解释性。可解释的学生记忆模型对教育研究更有意义。

在本文中，根据收集到的数以百万计的记忆数据，使用了用于模拟用户长期记忆的 DHP（Difficulty-Halflife-P(recall)）模型。我们设定了具有实际意义的优化目标：最小化用户形成长期记忆的成本。

为更接近自然环境的长期记忆模型，并通过现实世界的数据进行了测试。数据导入的过程如4-10所示。



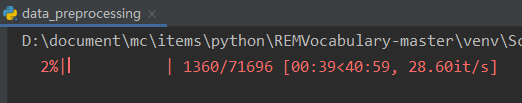


图4-10数据导入

为了验证DHP模型，在墨墨背单词上收集了其用户一个月的学习日志，从中筛选出 2.2 亿条记忆行为数据，用于训练模型来模拟学习者的记忆。由于以下原因，没有使用 Duolingo 的开源数据集：

* 它缺乏时间序列方面的信息，如反馈和间隔的序列，而对数据的分析表明，不同的序列对记忆状态有很大的影响。
* 它对回忆概率的定义有问题。Duolingo 将回忆概率定义为一个单词在复习环节中被正确回忆的次数比例，即假定同一单词在复习环节中的多次记忆行为相互独立。然而在实践中，第一次回忆会影响学习者的记忆状态和后续记忆。

从收集的记忆数据记录了学习者对单词记忆的历史信息。对于任意记忆行为，用一个四元组来表示：

其含义为一个学习者 在时刻  回忆单词 并反馈 （回忆成功 =1 ；回忆失败 =0 ）。为了便于研究记忆行为序列，我们将历史特征加入其中：

其中  表示学习者  对单词 的第 次回忆， 表示两次回忆之间的时间间隔。 和  分别表示第 1 到第 −1 回忆的间隔历史和反馈历史。

我们过滤掉其中第一个反馈为 =1 的日志，以排除学习者在使用间隔重复之前形成的记忆的影响。

为了捕捉记忆的衰退，我们需要得到二元回忆背后潜在的概率。为了得到回忆概率，忽略学习者本身的影响，用学习该单词的学习者中的回忆比例作为回忆概率 , 从而聚合得到：

通过控制 、 和 ，我们可以绘制每个   的，从而得到遗忘曲线。当  足够大时，比率  接近回忆概率。然而，共有将近10万个单词，在不同的时间序列中为每个词收集的行为事件是稀疏的。需要对单词进行分组，以便在区分不同的单词和缓解数据稀疏性之间做出权衡。鉴于对遗忘曲线感兴趣，材料的难度会明显影响遗忘速度。通过控制不同种影响因素来达到拟合模型的过程如图4-11所示。

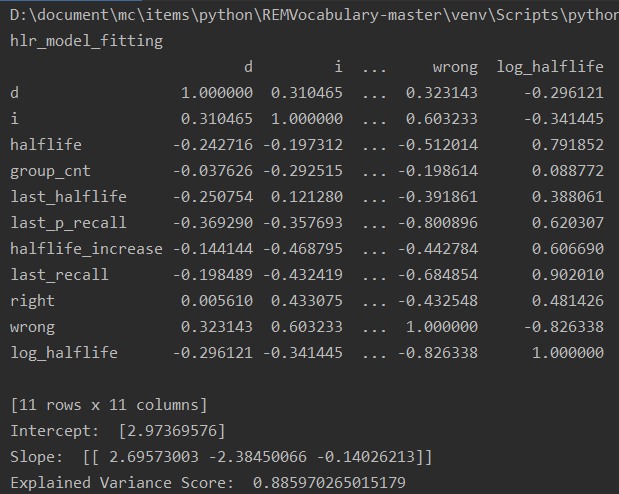


图4-11模型拟合

#### 4.3.2.2训练模型

将训练模型和使用模型区分开来是至关重要的。

使用模型意味着提供输入并接收估计或预测。当我们在训练模型以及我们或我们的客户在现实世界中使用模型时，使用模型所花的时间通常不到几秒钟。

相比之下，训练模型是改进模型性能的过程。训练要求我们在特殊循环中使用模型以及目标函数和优化器。这可能需要几分钟或几天的时间才能完成。通常，我们仅训练一次模型。训练模型后，我们可以随心所欲地使用它，无需进行进一步更改。

由于数据量过大，我们采取底层代码C++代码来运行代码，使用cmake对C++代码进行封装，封装为一个可执行文件。封装过程如图4-12所示。

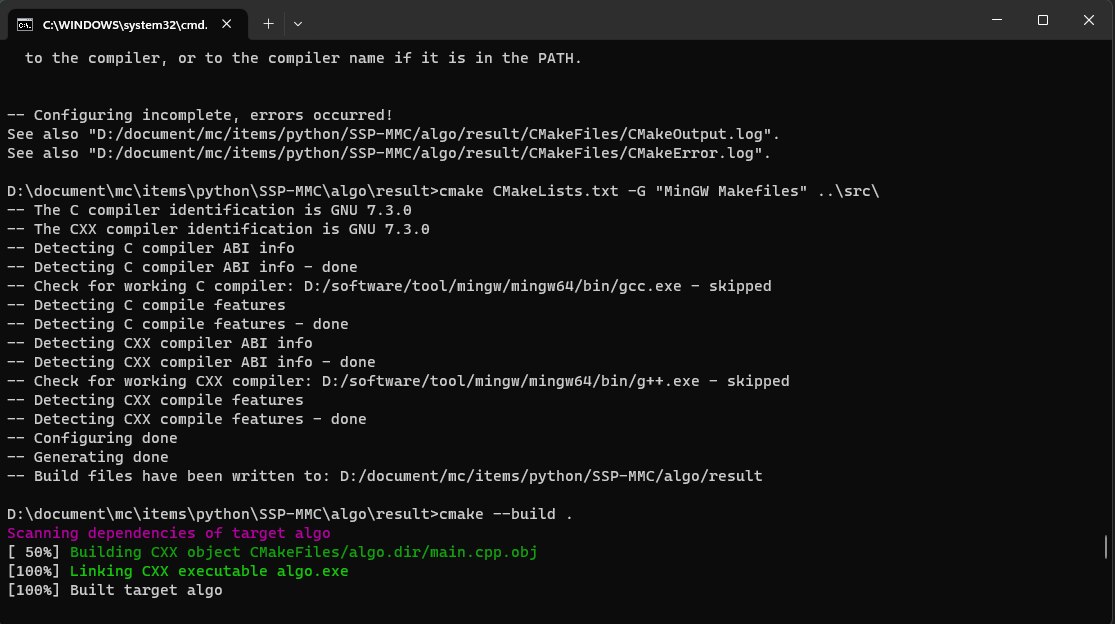


图4-12封装过程

封装为可执行文件之后，运行可执行文件进行训练数据集，训练流程如图4-13示。

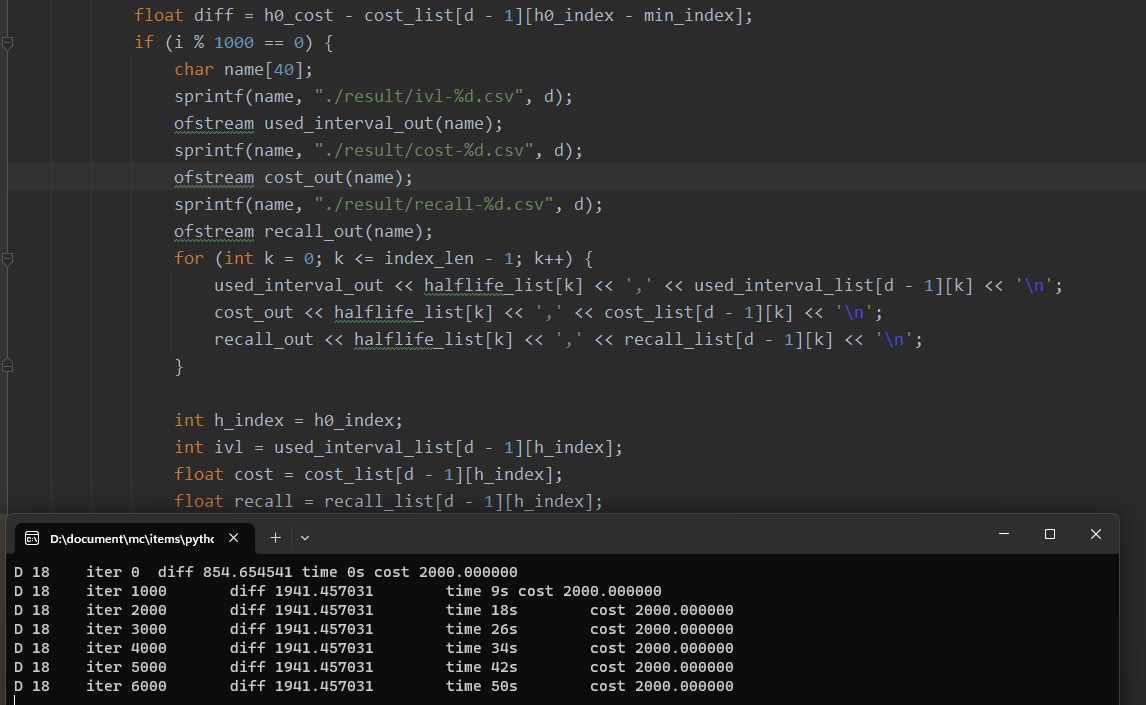


图4-13训练流程

通过在 DHP 模型的环境中训练数据，我们得到了训练好的机器学习模型并将其导出供单词系统使用。

## 4.4系统的核心算法

### 4.4.1词根匹配

在本研究中，我们采用词根匹配算法来过滤重复出现的词，获得关键词列表。词根匹配算法通过比较单词与一系列词根的匹配情况来识别含特定词根的单词，并利用词根信息过滤重复出现的词。具体而言，我们首先定义一系列词根，如act, ag, dict, duc, fac, fect, form等。然后使用正则表达式匹配文本中的单词与这些词根，得到初步的关键词列表。词根匹配算法的效果体现在两个方面:一方面它可以有效识别出一定的关键词，另一方面通过过滤重复出现的词和近义词它可以精炼关键词列表，避免冗余信息的出现。这两点效果的结合使得我们可以获得较为准确的关键词列表。本研究旨在评估词根匹配算法在识别关键词方面的效果。我们将算法应用于大单词本解析，评估其识别出的关键词列表与数据库中单词列表的重合度和相关度。对机器学习单词打卡的词库进行了处理，以减少词汇间的重复现象，提高学习效率。  
 运用词根匹配算法处理词汇时，我们首先对所有待处理的单词进行预处理，如转换为小写字母、去除标点符号等。接着，选用一种适当的词根匹配方法对预处理后的单词进行词根还原。该过程中，算法会根据预定义的规则，自动识别并剔除词缀，将单词还原至其基本形态。

在对词库使用词根匹配算法后，我们发现，重复词的数量明显减少，词库的规模更为紧凑。实际应用表明，经过词根处理的词库能够帮助学习者在更短的时间内接触到更多独立的单词，从而提高学习效果。

我们收集了每个字母开头的的词根、前缀、后缀字符文件，并与现有单词文件进行比对，清洗出适合于本单词数据库的词根、前缀、后缀字符文件，供后续系统导入单词本使用。应用算法比较词根、前缀、后缀过程如图4-14所示。

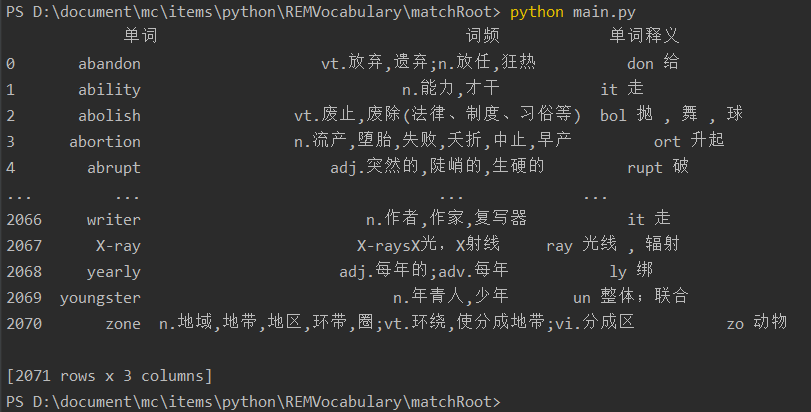
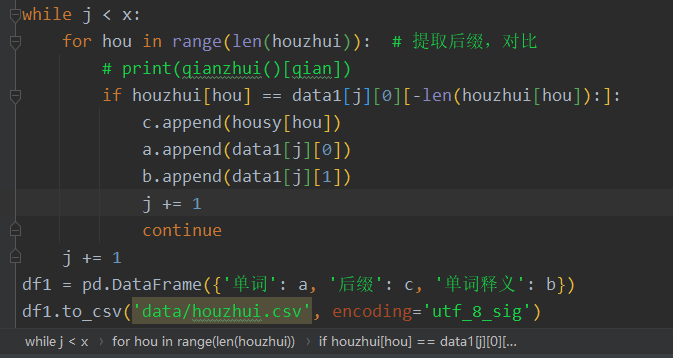


图4-14数据处理流程

并将处理后的词根、前缀、后缀字符组存入csv文件中供后续使用，存储的词根、前缀、后缀字符如图4-15所示。

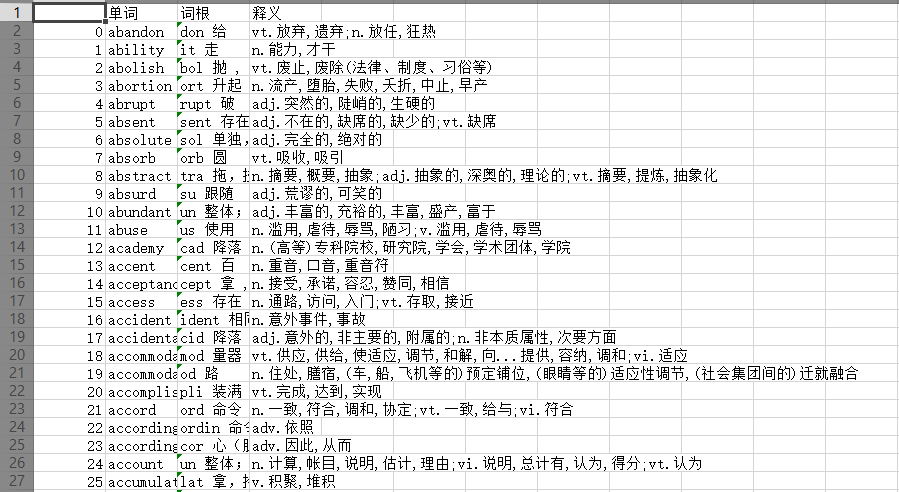


图4-15训练流程

将词根、前缀、后缀字符组和英语单词做比较，判断字符是否重复的算法。应用处理好的词根、前缀、后缀文件与单词库进行比较如下代码所示。

1. *# 读取文本文件*
2. with open('text.txt', 'r') as f:
3. text = f.read()
4. *# 定义词根列表*
5. root = open('D:\document\mc\items\python\REMVocabulary\matchRoot\data\词根.txt', 'r')  *# 读取词根文件*
6. *# 定义正则表达式匹配词根*
7. regex = re.compile("("+("|".join(roots))+")(\w\*)")
8. *# 查找文本中匹配的词根*
9. matches = regex.findall(text)
10. *# 过滤重复词根和已经使用的词根*
11. used\_roots = []
12. result = []
13. for root, word in matches:
14. if root not in used\_roots:
15. used\_roots.append(root)
16. result.append(root + word)

### 4.4.2自然语言处理

使用翻译语料库和机器学习技术可以自动生成语言学习内容，如英语填空练习。这一技术解决了人工生成语言学习内容难以实现大规模生成的问题。翻译语料库包含大量真实语境下的语言知识，是生成自然语言学习内容的重要知识源。机器学习技术可以自动分析语料库，生成对应的语言练习。考虑到两者的优势，本文提出使用语料库和机器学习技术自动生成英语填空的方案。

选用的语料库是TMX翻译记忆库，它包含源语言和目标语言的对应词与句。我们采用的机器学习技术主要包括分词、词形还原、词性标注和模板生成。这些技术可以自动分析语料库，识别关键词和句型，生成填空模板。语料库和机器学习技术的结合实现了自动生成英语填空练习的功能，这体现了两者优势的互补。

以关键词study为例详细展示自动生成英语填空的技术流程。首先从语料库中提取所有英语单词。然后选择study作为要生成填空练习的关键词。对study进行分词、词形还原和词性标注，识别出名词和动词。根据名词生成名词填空句，如The study of history。根据第一个动词生成动词原形填空句，如He studies chemistry。根据第二个动词生成动词过去式填空句，如They studied their lessons。

通过导入TMX翻译语料库，进行分词、词素分析、单词还原等操作自动生成各种不同语态、情态下的英语填空来进行单词学习。生成例句流程如图4-16所示。



图4-16训练流程

## 4.5本章小结

本章主要建立了机器学习单词打卡系统的系统设计模块，给出了该模块的具体实现细节，并对其中的关键问题进行了推导和解释。

首先，我们介绍了机器学习单词打卡系统的总体设计思路，并提出了系统设计的主要目标和考虑因素。随后，详细阐述了系统设计中各个模块的功能和作用，包括词汇表管理、数据预处理、模型训练和识别等模块，并利用流程图等形式来直观地展示各个模块之间的联系和交互方式。

总体来说，本章所建立的机器学习单词打卡系统的系统设计模块是一个稳健可靠的基础模块，为后续的模型训练和性能优化提供了重要保障。该模块的实现将为机器学习单词打卡任务的高效完成和精准识别提供有力支持。

# 第5章系统编码及运行

## 5.1系统开发涉及的软件

表3-1开发软件

|  |  |
| --- | --- |
| 操作系统 | windows11 |
| 数据库 | Mysql关系数据库 |
| 使用语言 | Python |
| 使用框架 | Django |
| 开发软件 | Webstorm、Pycharm、Navicat、Postman |

## 5.2系统运行界面及结果

登录页面为本系统的入口页面，只有通过登录才能进入系统的主页面，同时因为数据存储的设计与用户名相关，本系统在登录后的运作需要依赖与正确的用户名。登录页面如图5-1所示。

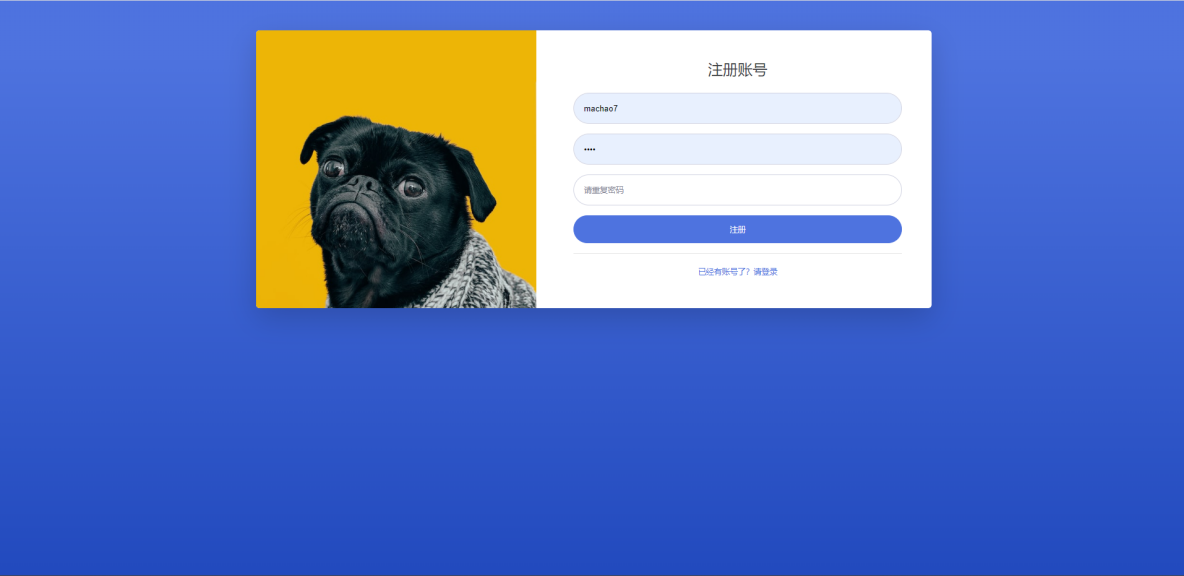


图5-1 登录界面

登陆时输入的账号密码发送登录请求到后端服务器上，服务器通过查询数据库中的查找出用户的账号和密码，进行对比，如果验证为成功，则成功跳转到主页面。

注册功能通过登录页面的按钮点击进入，本系统的注册功能在使用方面与普通的注册功能相似。图5-2为注册页面。

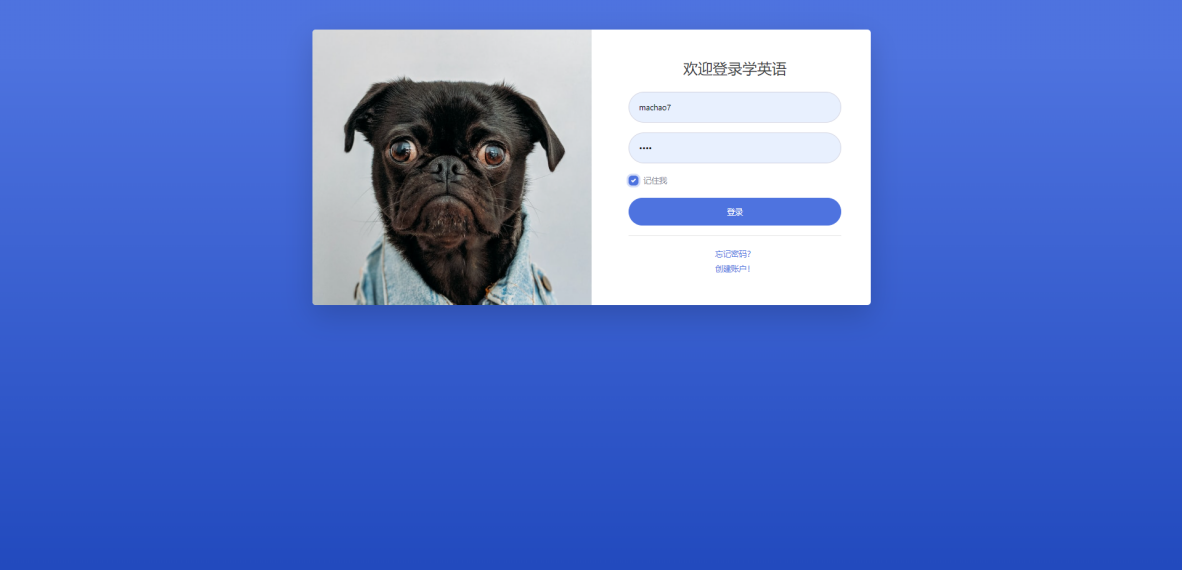


图5-2 注册界面

用户登陆后进入网站首页，在该界面中用户可以查看自己的学习打卡记录以及遗忘曲线，学习数量以及自己的学习情况。用户登陆后的首页如图5-3所示。

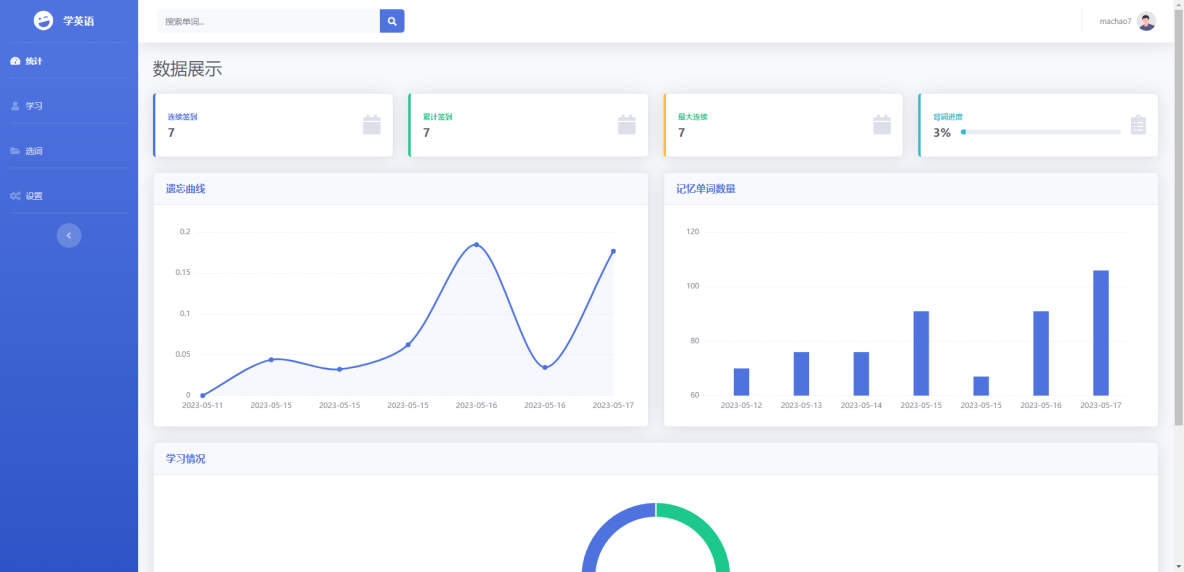
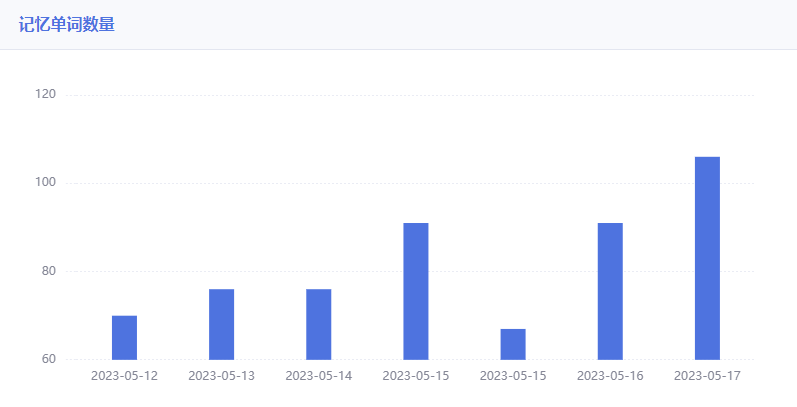
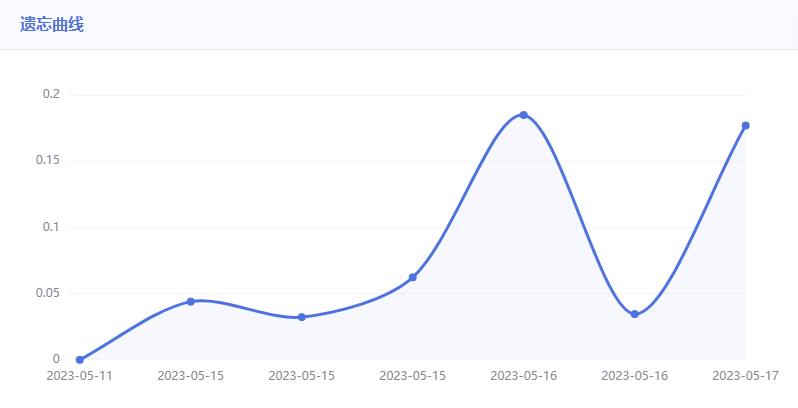


图5-3 首页界面

具体的模块都展示在此页面上，遗忘曲线，记忆单词数量，学习情况。遗忘曲线根据用户的单词学习情况系统自动分析出。记忆单词数量展示出近一周用户每天的学习单词数量。学习情况展示用户当日学习新单词的数量以及复习单词的数量，并通过饼图的形式展示。遗忘曲线，记忆单词数量，学习情况如图5-4，5-5，5-6所示。



|  |  |
| --- | --- |
| 图5-4 遗忘曲线 | 图5-5 记忆单词数量 |



图5-6 学习情况

在此页面，用户可以对个人基本信息进行修改，同时也可以退出系统。如图3-7所示。



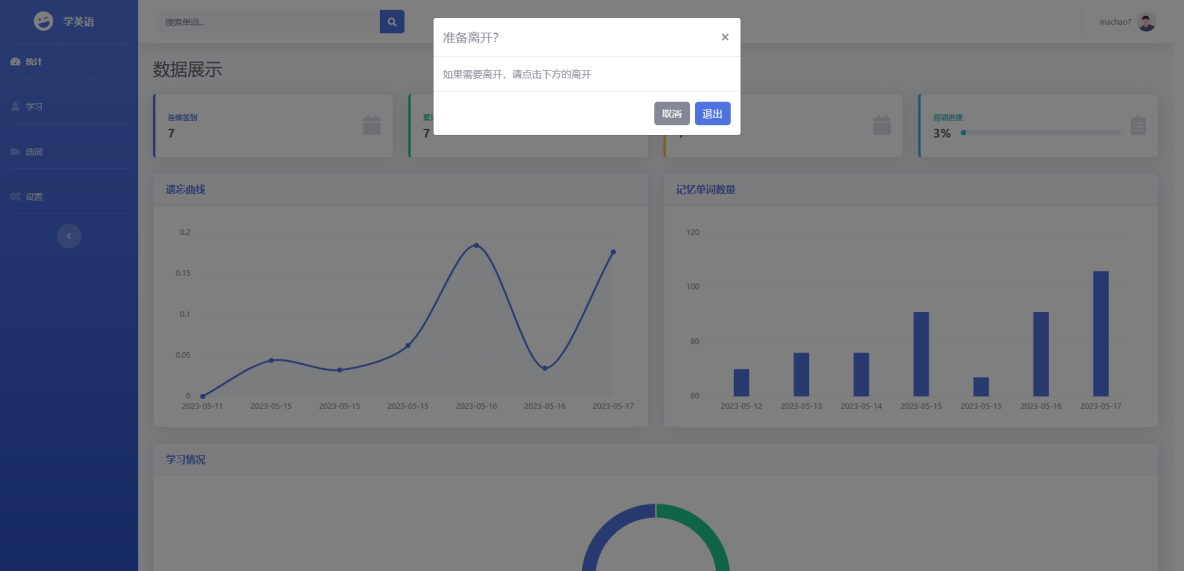


图5-7 退出系统

通过该界面企业用户可以通过搜索单词，搜索出的单词会展示出单词的英美式音标，以及中文例句供用户学习。如图5-8所示。

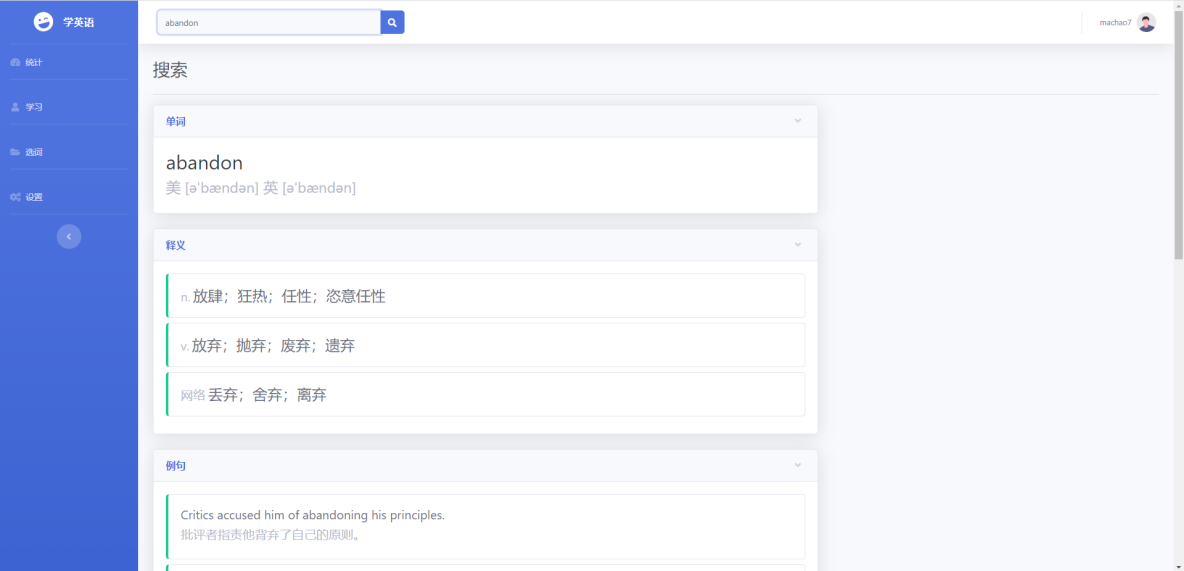


图5-8 搜索单词

用户登陆后在首页可以切换学习界面，在该界面中用户可以进行完成单日的学习计划单词，单词为系统根据机器学习模型自动生成的。在学习单词的时候，每日需要复习以及学习的单词会交替出现。在学习新单词时，出现下一个的按钮，点击后出现单词的详细情况。在复习单词时可以通过单词拼写的形式或者对这个单词很自信直接点击熟悉，切换到单词的详细界面来验证自己的记忆情况，在详细界面出现三个按钮熟悉来告知系统自己的记忆是否正确，模糊，陌生，三个按钮对应不同的记忆情况，系统会根据点击的按钮调整机器模型来安排后续的单词复习以及单词出现的情况。用户学习界面如图5-9，5-10所示。



图5-9 学习界面

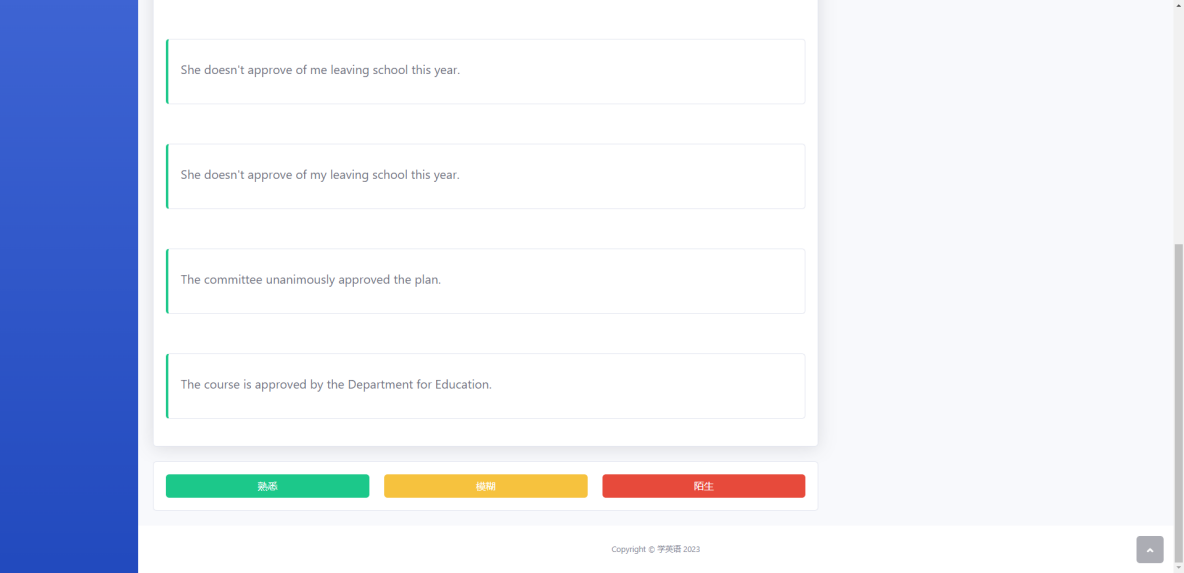


图5-10 学习界面

切换到选词界面，用户学习单词可以根据系统数据库自带的数据，也可以根据自己导入单词本的形式。导入单词本之后，系统自动根据前后缀分词以及自然语言处理等算法来处理解析单词并加入到系统数据库中供用户使用。选择单词本界面如图5-11所示。

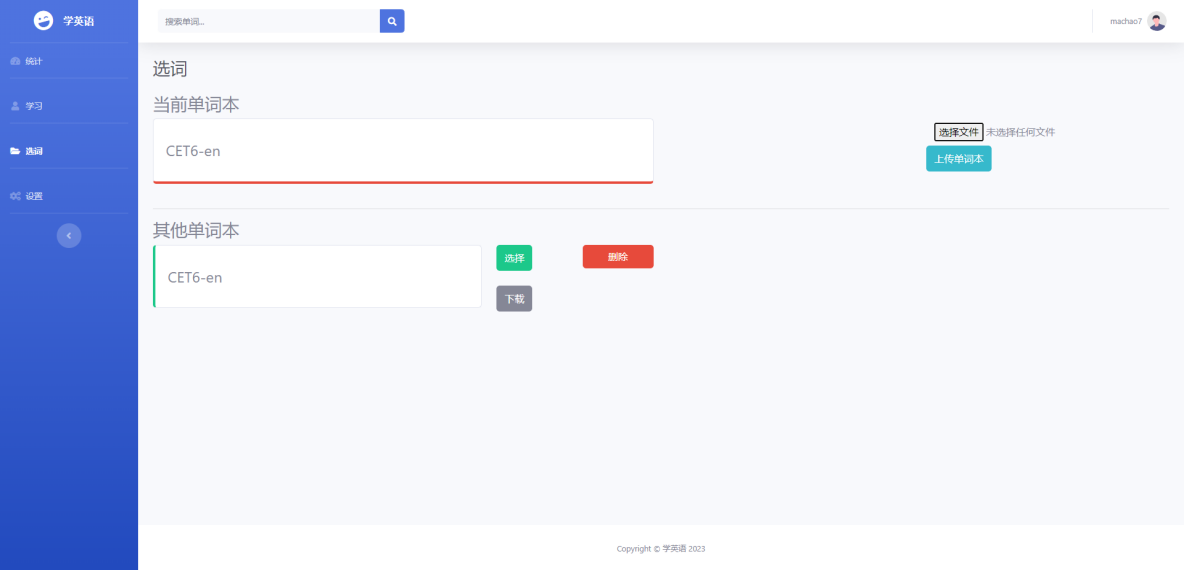


图5-11 选词界面

进入到个人信息设置界面。可以进行用户别名修改，个人密码修改以及每日学习单词数量修改功能。在这里修改之后数据会同步到系统数据库之中。修改每日学习单词数量之后会在明日的学习中体现。个人信息设置界面如图5-12所示。

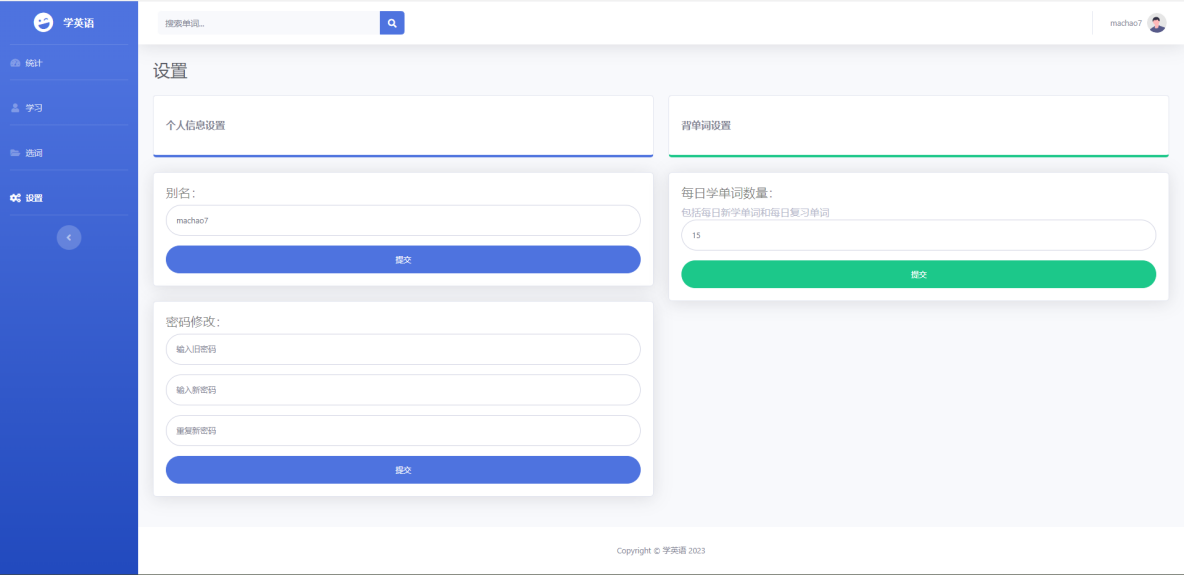


图5-12 个人信息设置界面

# 第6章系统测试

## 6.1测试分析

### 6.1.1测试目的

测试是判断一个人工智能系统质量与价值的关键手段，它使理论与技术框架实际转化为现实世界中的功能与创造价值。对于本系统来说，测试的目的在于：

第一，深入验证其设计理念的科学性与可行性。系统基于主动监测与优化学习过程这一设计理念来实现提高学习效果的目标，测试可以判断这一理念在系统实现过程中是否达成预期，如果未达成则可以发现理念实施过程中存在的失误或限制。这使理念得以在实践中进一步优化与改进，是实现理论创新与技术创造契合的关键步骤。

第二，全面评估系统技术框架的适用性与完整性。系统采用机器学习和推理技术构建自动提醒与内容推荐机制，测试可以检验这一技术框架是否能够胜任实际学习场景的需求，如果不能则可以发现其优缺点与不足之处，为技术框架的进一步优化提供依据。这使技术框架能够与实践场景完美契合，真正发挥其功效。

第三，评估系统对学习效果的提高，这是判断一个教育类系统成效的关键指标。

测试提供全面深入的评价。验证系统设计理念和技术框架的适用性，发现系统问题，为改进提供依据，衡量系统成效的关键手段。良好测试判断系统质量与特点，使理念框架系统功能转化为真实世界中创建价值。测试使产品由设想转入实践，让理论联系实际，是人工智能系统顺利实施的重要环节。

### 6.1.2对比测试

根据收集的数据集训练了 DHP 模型，并将其与 HLR 模型进行比较。还对 DHP 模型的参数进行了可视化，以获得对记忆的直观解释。然后，使用 DHP 模型作为训练环境，以获得最佳策略并将其可视化。最后，在由 DHP 模型组成的模拟环境中比较不同基线的性能。对比过程如图6-1示。

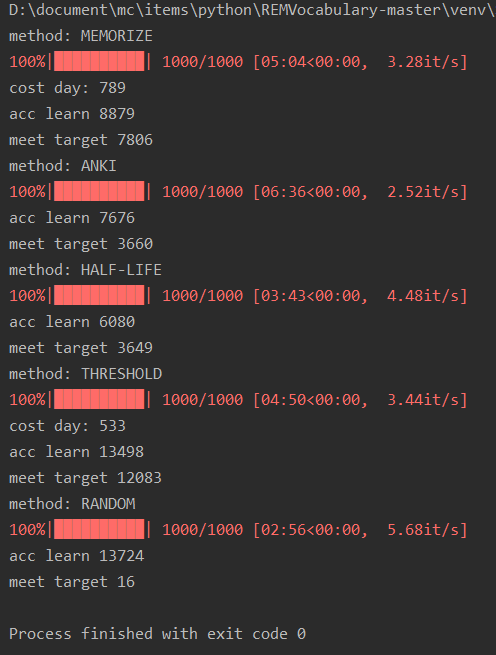


图6-1对比过程

对比过程基于5 个基线策略：RANDOM、ANKI、HALF-LIFE、THRESHOLD、MEMORIZE 和 3 个指标：目标达成量（THR）、累计学习量（WTL）、回忆期望（SPR）。

**基线**

几个基线调度算法进行对比：

* RANDOM 策略，每次从 [1，] 中随机选择一个间隔进行安排复习。
* HALF-LIFE，即直接以半衰期作为本次安排的复习间隔。
* THRESHOLD，即当 小于等于某一阈值（使用 80%）时进行复习。
* ANKI，SM-2 的一种变体。
* MEMORIZE，一种基于最优控制的算法。

**指标**

评价指标包括：

* SPR，累计回忆期望，即学习者所有记忆的单词的回忆概率之和。
* WTL，累计学习数量。
* THR，达到目标的记忆数量。

**实现细节**

我们设定 360 天（接近一年）的回忆半衰期为目标半衰期，当单词的半衰期超过这个值时，它将不再被安排复习。然后，考虑到实际场景中学习者每天的学习时间大致恒定，我们设定 600s（10 分钟）为每天学习成本的上限。当每次学习和复习过程中的累计成本超过这个上限时，无论是否完成，复习任务都会被推迟到第二天，以确保每种算法在相同的记忆成本下进行比较。我们使用学习者的平均时间，即成功回忆的时间为 3s，失败回忆的时间为 9s。然后，语言学习是一个长期的过程，我们设定模拟时间为 1000 天。

**分析**

数据拟合结果如图6-2，6-3，6-4所示。

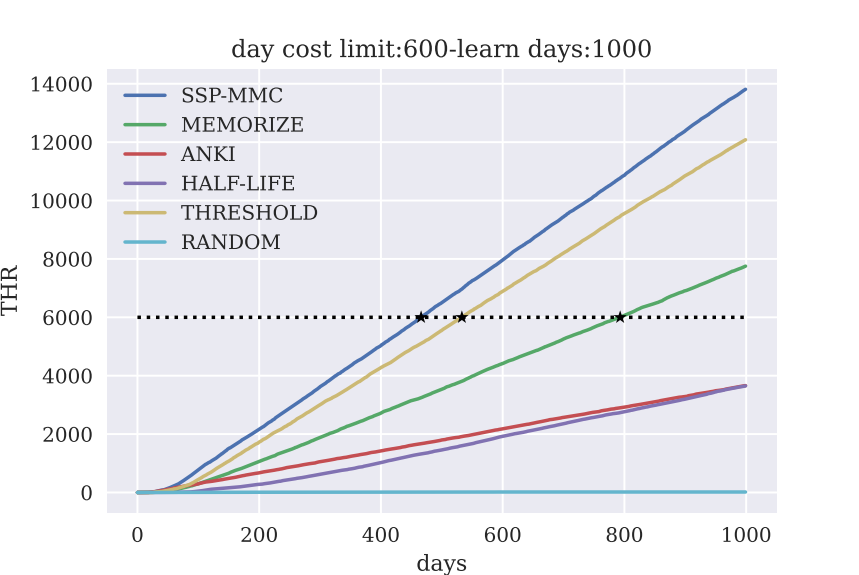


图6-2训练流程

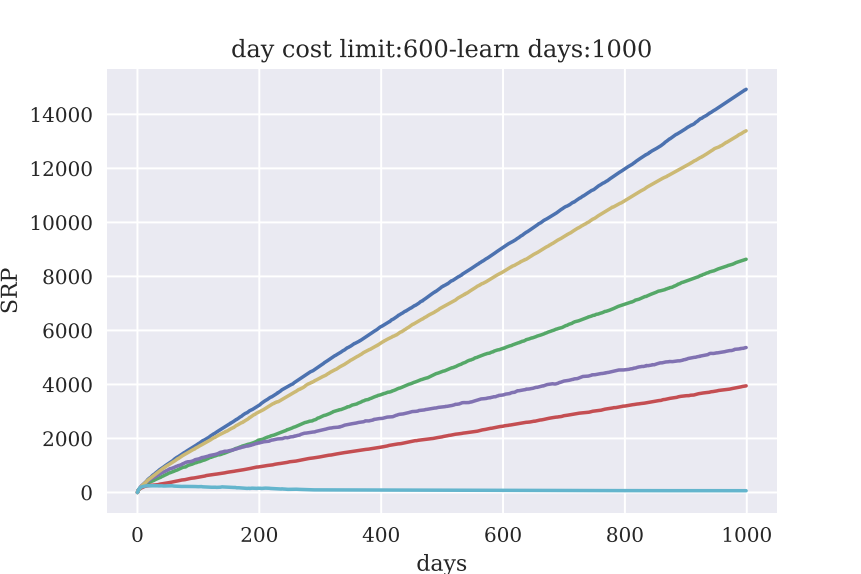


图6-3训练流程

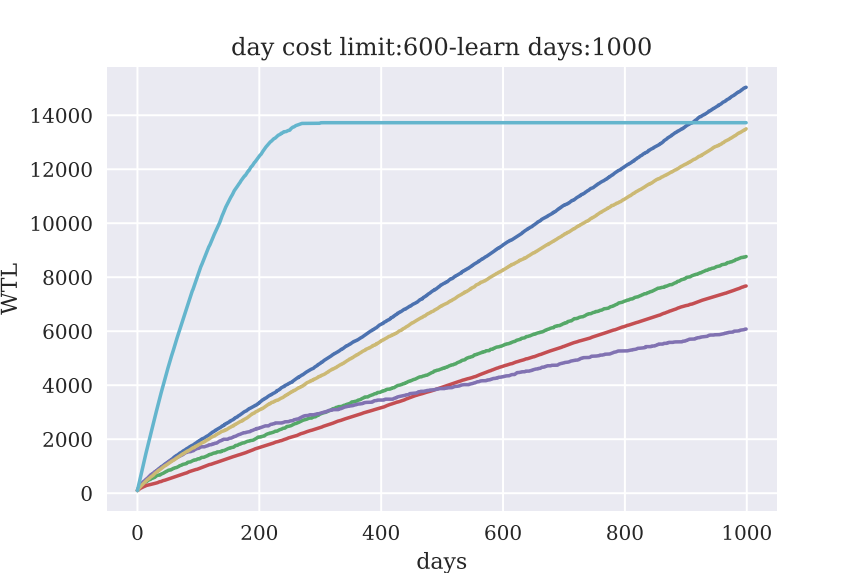


图6-4训练流程

## 6.2结果分析

上图中的仿真结果显示，在 THR 上， DHP 的表现比所有基线都好，这在意料之中。THR 与 DHP 的优化目标一致，而且 DHP 可以达到这个指标的上限。

为了量化每种算法表现之间的相对差异，我们比较了 THR = 6000 的天数（图中标记为 ⋆）： DHP 为 466 天，THRESHOLD 为 533 天，而 MEMORIZE 为 793 天。与THRESHOLD 相比，DHP 节省了 12.6％ 的复习时间。

SRP 上的结果与 THR 上的结果相似。这意味着学习者按照 DHP 安排学习会记得最多。

在 WTL 上，RANDOM 在早期阶段击败了所有的算法，因为只要调度算法不安排复习，学习者就可以继续学习新的单词，但这是以忘记已经学过的单词为代价的。此外，DHP胜过其他基线，因为它将记忆的成本降到最低，给学习者更多的时间来学习新词。

相较于传统记忆模型，考虑时序信息的模型在拟合用户长期记忆上的精度有显著提升。在各项指标上都超过了之前的记忆算法。

## 6.3本章小结

在本文中，我们提出了一种基于机器学习的英语单词学习方法，以帮助语言学习者更加有效地学习和记忆英语单词。我们提出的方法分为两个步骤：单词生成和练习生成。在练习生成阶段，我们结合单词和填空练习模板，使用机器学习算法自动生成多种类型的英语单词练习题。

为了验证我们的方法的效果，我们设计了一组实验，我们从公开的单词数据集中随机选取学习行为数据，并进行了验证。我们比较了常见的极限调度算法，并选择了性能最好的逻辑回归模型。在不同阈值下计算单词记忆的半衰期，绘制记忆曲线，并计算曲线下记忆单词需要花费的时间。同时，我们还计算了模型的准确率、半衰期等指标，以便更全面地评估模型的性能。

实验结果表明，我们提出的方法可以显著提高用户的英语单词学习效果，并且比传统的背单词方法具有更高的学习效率和更好的记忆效果。此外，通过人工评估与主观测评，我们得到了许多具有较高难度和足够多样性的练习题，这也表明了我们的方法在练习生成方面具有一定的优势。

结论

通过对翻译语料库和机器学习技术的研究与应用，我们实现了一个自动生成英语填空练习的系统。该系统可以大规模生成贴近真实语境的英语填空内容，这为英语学习者提供了丰富的练习素材，开启了一种全新的学习方式与体验。

在系统开发过程中，我们采用了分词、词形还原、词性标注、语料匹配等技术来自动分析语料，生成填空模板，使用DHP模型并训练数据生成模型文件供系统使用安排单词出现的频率。这些技术的应用克服了人工生成语言学习内容难以实现大规模与个性化的问题。我们使用的TMX翻译记忆库作为知识源，其包含的丰富语料使得生成内容更加自然。这也印证了语料库在自然语言处理与人工智能中的应用价值。

为评估该系统，我们采用了记忆曲线分析、主观评价等手段。这些测试结果显示，系统具有较高的分类准确性，生成内容也较好地满足了学习者的需求。但是，系统的性能还有待提高，同时也需要增加更加丰富的填空形式来实现对学习者全面且深入的支持。

未来，我们将继续扩充语料库，优化分类算法与特征选择方法来增强判断填空模板的准确性。我们也将实现更加丰富的填空形式，如名词、形容词、副词填空，并拓展到其他语言与学习内容的生成。此外，我们将不断完善评测机制与手段，包括增加更全面的数据测试集、深入的用户体验评价等，从而使系统的性能与体验不断提高。

总之，机器学习与语料库技术为自动生成高质量语言学习内容提供了重要手段。这一系统的建立有助于实现教育内容的智能化生产与学习方式的改变。我们相信，随着相关技术与评测手段的不断进步，这一系统也必将实现更强的智能化，最终成为学习者高效而极具个性化的英语学习助手。这必将为语言学习与教育产业带来深远的变革与影响。

致谢

感谢国家！

感谢党！

感谢社会主义新中国！

# 参考文献

* 1. 周游,周炜.对比教学法在轮机英语教学中的应用[J].青岛远洋船员职业学院学报,2022,43(04):57-60.
  2. Mitchell T M. Machine learning[M]. New York: McGraw-hill, 2007.
  3. 姜丰辉,刘祥鹏,邵巍,陈春平,于龙振.影响因子操纵期刊识别与分类方法构建与应用[J].中国科技期刊研究,2023,34(02):136-143.
  4. 张曌曌.社区教育“游学”:面向全民终身学习的新型教育模式[J].山东开放大学学报,2023(01):18-21.
  5. Yip F W M, Kwan A C M. Online vocabulary games as a tool for teaching and learning English vocabulary[J]. Educational media international, 2006, 43(3): 233-249.
  6. 王延桃. 基于微信小程序的词汇学习平台设计与实现[D].云南师范大学,2020.
  7. 李成洁. 深度学习理论下初中英语词汇教学策略的实验研究[D].石河子大学,2022.000915.
  8. 姚歆蕾. 基于艾宾浩斯遗忘曲线的英语词汇学习微信小程序的设计与实现[D].云南师范大学,2021.001596.
  9. 李振兴. 基于Android平台移动生活信息化系统的设计与实现[D].西安电子科技大学,2013.
  10. 王甜甜. 高中英语学困生词汇教学的实践研究[C]//广东省教师继续教育学会.广东省教师继续教育学会教师发展论坛学术研讨会论文集(二).[出版者不详],2023:268-271.
  11. Thornton P, Houser C. Using mobile phones in English education in Japan[J]. Journal of computer assisted learning, 2005, 21(3): 217-228.
  12. 吴倩寅.平板电脑协助下的初中英语阅读教学活动设计研究以Unit 6 Sunshine for all为例[J].英语教师,2021,21(13):117-123.
  13. 张思茹,张朝霞.“口语100”辅助初中英语教学的有效性研究[J].海外英语,2023(05):186-188+194.
  14. 赵陶.英语单词记忆移动平台的实现[J].微型电脑应用,2019,35(08):130-132.
  15. 李文慧,王俊丽,冯小玲.单词记忆软件中自主学习策略应用的探究[J].海外英语,2022(07):86-87+104.
  16. 滕培培,张海峰,苏志祁,李文倩,张博文.基于ArUco码的轨道运输设备定位系统的研究及应用[J].冶金自动化,2022,46(05):36-42.
  17. Kara L. Design Principles[J]. 2010.
  18. Chen P P S. The entity-relationship model—toward a unified view of data[J]. ACM transactions on database systems (TODS), 1976, 1(1): 9-36.
  19. 骆科林,刘湘乡,胡金有,谢朝,邹炼.PyRERT:面向放射治疗研究的Python环境[J].中国医疗器械杂志,2023,47(02):135-139.
  20. 王翯,刘诗雨,王慧,李伟,吕蒙.高光谱智能分析技术在中药材领域的研究与应用[J/OL].中国中药杂志:1-10[2023-05-24].
  21. 和占全,李晓明,冯汝结,靳清平.基于信创的互联网新闻数据信息安全管理系统的设计与实现[J].网络安全技术与应用,2023(03):94-96.
  22. 刘柯言,党佳伟,郑蕊蕊,周瑜,贺建军.满文档案图像智能处理系统设计与实现[J].大连民族大学学报,2022,24(01):2022.01.009.
  23. 杨瑞.智能运销系统建设方案[J].煤炭加工与综合利用,2023(03):5-8+14.