

# 2020 年春季《大学计算机基础》(理科)

## 大作业题目要求

### 一、要求和说明

从下面给出的 5 道题目中，选择 1 道题目，采用 Python 语言实现程序并撰写实验报告。要求独立完成。

(1) 在实验报告中，需要介绍作业完成功能、实现方案、设计思路(给出**关键代码**)和涉及的知识点、创新点或增加功能、遇到的难点及解决办法(给出**关键代码**)、GUI 初始界面、程序运行结果截图等。

(2) 进行课程学习总结：课程收获、课程难点分析、教师授课评价、助教评价、课程进一步改进建议。

### 二、关于申优答辩

大作业设有申优答辩环节，**必须参加申优答辩大作业才可能获得优秀**(90 分及以上)。如果你认为自己的这次作业完成得很好，可以申请优秀，则必须参加线上视频答辩。答辩现场表现占大作业成绩的 40%，另 60%根据源程序完成情况、实验报告质量评定。大作业成绩占总成绩的 15%。

**说明**：你也可以不参加申优答辩，则大作业成绩最多只能是良(89 分)，主要根据源程序完成情况、实验报告质量评定。

申优答辩时间定在第 16 教学周(具体时间由教师、助教及全体同学以小班为基本单位商议决定)，答辩顺序由任课教师提前排定。每位同学有 4 分钟的展示时间以及 1 分钟的回答提问时间。

在任课教师规定的提交截止时间之前，将最终版源代码、实验报告(包括课

程学习总结)、相关材料(如程序运行必须的文件、图片,或者程序运行所必须的其他模块及其安装说明)、答辩用幻灯片(参加申优答辩者)放入以“学号+姓名”命名的文件夹中,压缩后提交至课程中心指定栏目。

**大作业不允许抄袭!我们将对作业进行查重。一旦发现抄袭,则抄袭者和被抄袭者均为 0 分!**

# 题目 1：本地简单数据分析平台

## 一、题目描述

当今社会中统计学、大数据分析等已经成为无论文理的各个学科中必要的研究途径以及展示数据最好的形式，其便捷性及强大的功能也是各个软件所追求的目标之一。

选择本题，你需要完成一个程序，包含一个 GUI 的界面，能够清晰地引导用户导入数据文件，并且给出至少两种数据分析及展示方法，最后能够将结果导出为用户选择的格式。

## 二、基本实验要求

本题你必须给出一个 GUI 界面，需要清晰直观地给出导入和导出文件的按钮，在导入文件之后能够给出数种可以进行一键分析的按钮。

根据用户选择的分析方法，需要自动检测导入的数据是否符合要求（例如单变量数据不能够进行相关分析），并给出相应的提醒或者是报错信息。

展示方法中可以由用户自定义横纵坐标刻度、标题以及图例等相关参数，支持双变量做图等功能，导出前可以由弹窗等形式进行预览。

## 三、评分细则及加分项

### 1. 必做部分

你的程序必须有一个 GUI 页面，要求简洁美观，能够一键完成用户所需的操作：

（1）导入数据文件。基础的单变量数据文件规定为一个  $n+1$  行两列的 csv 文件。

其中包含  $n$  个数据，例如下表：

年份	车辆总计（万辆）
2008	930.61
2009	1087.35
2010	1133.32
2011	1263.75

要求  $n$  不小于 100，数据的来源可以根据自己的专业喜好来选择，可以但不限于从以下网站获取：

a) 国家统计局：<http://data.stats.gov.cn/index.htm>

b) Science Data Bank：<http://www.sciencedb.cn/index>

导入的方式可以是直接从当前目录读取特定名称的 csv 文件，也可以给出一个文本框，由用户输入文件名。

（2）数据分析方法，必须包括以下方法：

a) 描述统计（平均数，中位数，众数等最基本的统计量）

b) 回归分析

在导入数据后，应给出相应的数据分析按钮，用户点击后应能够显示出分析结果。

（3）对于描述统计结果可视化，必须包括以下方式：

a) 柱状图

b) 折线图

(4) 将 Matplotlib 可视化结果可以用弹窗进行预览，并且导出保存，图片格式为 png 格式。保存的位置任意。

## 2. 选做部分

以下内容非必做项，每完成一条选做项，都可以给大作业带来额外的加分，你**同样可以用其他额外的内容来丰富你的大作业**，以获得加分。以下为提供的一些思路和方向。

(1) 在导入数据方面，支持更多的文件格式和数据文件内容，比如 json、xls 文件，多变量数据等。

(2) 在数据分析方面，添加更多的方法，可以包括但不限于：

a) 广义线性模型

b) 非参数检验

(3) 在做图方面，支持更多的做图类型，例如饼图等，支持用户自己定义图例，图线颜色、标题等。对于多变量的数据，可以进行多线图等多变量的图形绘制。

(4) 在 GUI 方面，能够将导入的数据使用表格进行展示。

(5) 在用户选择了分析方法后，可以自动检测导入的数据是否符合要求（例如单变量数据不能够进行相关分析），并给出相应的提醒或者是报错信息。

(6) 导出时，你可以利用 Markdown 文本，将数据与分析结果进行整合，最后生成一份详细的数据报告。

(7) 以上选做部分不要求全部完成，也不必局限于给出的这些内容。

## 四、实验指导

### 1、Pandas: 强大的分析结构化数据的工具集

对于一般的二进制文件当然可以用 open() 函数直接实现，但是对于 json、csv 等需要二次处理的文件，或者是 xls 等非二进制文件，可以使用 pandas 进行处理，同样 pandas 也可以写入这些文本。在数据处理时 pandas 的 dataframe 也能使整个工作流更简化。

官方中文网站: <https://www.py pandas.cn/>

### 2、NumPy: 更强大的科学计算的基础软件包

在处理更大量的数据和多维数组时，利用 NumPy 能更高效和简便地处理数据。

官方中文网站: <https://www.numpy.org.cn/>

### 3、Matplotlib: 基础的绘图工具

这个不做过多的说明，课程内的内容。

官方中文网站: <https://www.matplotlib.org.cn/>

### 4、PyQt5 | Tkinter: GUI 开发库

使用 Tkinter 或者 PyQt5 完成大作业均可。非官方中文教程:

PyQt5: [https://maicss.gitbooks.io/pyqt5/content/hello\\_world.html](https://maicss.gitbooks.io/pyqt5/content/hello_world.html)

Tkinter: <https://www.runoob.com/python/python-gui-tkinter.html>

## 题目 2：本地简单密码存储查询系统

### 一、题目描述

你是否被以下问题所困扰：录入志愿时长码的时候，忘记了自己的志愿北京账号和密码？不得不找回密码，却意识到每录入一次时长码就要重置一次密码，非常的不方便。拿纸笔记下又怕弄丢了，而且还会被别人看到，有泄露信息的风险。但是你可以用 Python 完成密码的记录，加密以及对应的解密和查询的功能，来记录一些很少用到却很重要，或是密码十分繁杂的账户和密码了。

选择本题，你需要完成一个程序，包含一个 GUI 界面，能够让用户存入自己的一些账户密码（为了和该程序的登录密码区分，以下会称用户希望存储的密码为密码数据），清楚地列出已经存入的账户，并且在用户需要使用密码时能够查询到对应的密码数据。除此之外还应保证安全性，不应将密码数据内的密码展示给其他人。

### 二、基本实验要求

本题你必须给出一个 GUI 界面，需要清晰直观地给出添加数据和查询数据的按钮，并且直观地显示出目前已经存储的所有账户。

### 三、评分细则及加分项

#### 1. 必做部分

你的程序必须有一个 GUI 页面，要求简洁美观，能够通过按钮、输入框完成用户所需的操作：

- （1）拥有一个用于登陆的窗口，用户可以进行登录（账号密码可以事先给定）。
- （2）用户可以添加想要存储的密码数据。每一条密码数据包含账户名、用户名及密码，需要支持所有 ASCII 编码的字符，例如：

账户名：BUAA-ss0

用户名：Marvolo

密码：123456

同时要求添加密码数据在独立的窗口中完成。

- （3）程序的主窗口需要直观地列出用户的所有已存储的账户，方便用户知晓已经存储了哪些账户；

- （4）用户可以通过输入一个账户名来查询有关该账户的密码数据；
- （5）密码可以在显示明文与显示“\*”之间（或是任意表示遮盖的其他字符）切换，默认应不显示密码明文（包括登录界面的密码）；

- （6）用户添加的密码数据应当保存在本地文件中，以便在下一次打开时仍然会保留之前添加的密码数据。在保存密码时需要对密码进行加密，加密方式任意，简单或复杂的均可，不做具体要求。

#### 2. 选做部分

- （1）每条密码数据可以包含更多项目，例如哪些网站或者 APP 使用了这条密码。也可以让用户自定义项目名称；

- （2）可以增加修改和删除某一条密码数据的功能，其中修改密码数据应当在独立的窗

口中完成；

(3) 可以支持用户在主窗口中直接选中某一账户名后单击某个按钮一键完成查询、修改或删除的功能，而不用在搜索框中输入账户名；

(4) 可以给列出账户名的控件添加滚动条；

(5) 可以支持 UTF-8 编码的密码数据；

(6) 可以采用更强的加密手段来确保用户存储的密码数据的安全性，比如可以对于不同用户采用不同的加密手段；

(7) 可以选择使用二进制文件存储数据以便云备份。

(9) 以上选做部分不要求全部完成，也不必局限于给出的这些内容。

## 四、实验指导

### 1. PyQt5 | Tkinter: GUI 开发库

非官方中文教程：

PyQt5: [https://maicss.gitbooks.io/pyqt5/content/hello\\_world.html](https://maicss.gitbooks.io/pyqt5/content/hello_world.html)

Tkinter: <https://www.runoob.com/python/python-gui-tkinter.html>

另外，教材上对于 Tkinter 的讲解已经比较详细了。如果不使用 ttk 控件的话，只看教材也可以做出不错的 GUI。

### 2. bytes 类型：

类似于字符串，但存储的是由字节所组成的“字节串”。

下面提到的摘要以及加密库均需要对 bytes 类型进行操作。

使用 `str.encode(self, 'utf-8')` 方法，可以返回将字符转使用 UTF-8 编码转换为“字节串”后的结果；

使用 `bytes.decode(self, 'utf-8')` 方法，可以返回将“字节串”使用 UTF-8 解码为字符串后的结果。

### 3. hashlib: 内置的摘要算法库

提供了常用摘要算法如 MD5 和 SHA1 等，无需额外安装即可直接使用。摘要算法用于“单向加密”某个数据（带引号是因为这个词好像并不正规，正规说法就是摘要），相同的明文可以得到相同密文，但是无法直接通过密文反推明文，可以用于登陆时的身份验证。

非官方中文教程：

<https://www.liaoxuefeng.com/wiki/1016959663602400/1017686752491744>

### 4. Crypto: 加密算法库

该库为第三方库，使用之前需要先安装该库。使用 `pip install` 时，Win 系统需执行指令 `pip install pycryptodome`，MacOS 需执行 `pip install pycrypto`（待验证），下载速度过慢可以在指令最后加上 `-i https://pypi.tuna.tsinghua.edu.cn/simple` 通过清华镜像下载第三方库。

该库支持 AES、DES、RSA 等各种常见加密算法，不过很不幸，暂无官方中文使用说明，且非官方使用教程比较分散。想要使用该库的同学可以尝试搜索“使用 python 完成 AES 加密”等关键字。

## 题目 3：简易 B 站弹幕分析工具

### 一、题目描述

网络爬虫是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本。通过爬虫技术，你甚至可以获取互联网上任何你想要的信息。

哔哩哔哩现为中国年轻世代高度聚集的文化社区和视频平台，而弹幕也是 B 站的一大特色。本题意在使用爬虫技术对 B 站弹幕进行抓取，并利用相关工具对爬取的弹幕进行分析。

### 二、基本实验要求

本题希望同学们充分发挥自己的自学能力和资料查找能力，利用 Python3 及其强大的第三方库设计实现能够自动爬取指定视频的弹幕，并给出相关分析报告的 B 站弹幕分析软件。

### 三、评分细则及加分项

#### 1. 必做部分

- (1) 设计一个美观简洁的 GUI，使用户可以通过 GUI 操作软件。
- (2) 用户输入视频 BV 号，你的程序需要根据 BV 号来爬取相应视频的弹幕，并在主窗口显示至少前 100 条弹幕的内容（如果弹幕数量足够多）。显然一个屏幕的长度不足以显示 100 条弹幕的内容，你的程序可以通过加入滚动条或其他方式让用户浏览全部内容。
- (3) 为了让用户对此时视频的弹幕有更直观的认识，你的程序需要进行相应的统计分析，包括但不限于：高能进度条（具体会在下文讲解）、弹幕数量 top10 的柱状图、弹幕类型的统计图、弹幕颜色的统计图等。你将数据刻画得越充分，你的分数也会有相应的提高。
- (4) 显示中文的位置禁止出现乱码。
- (5) 操作过程中禁止弹出第三程序，禁止依赖命令行完成操作（即使用 os 库中的 system 函数）。

#### 2. 选做部分

- (1) 网络异常或其他原因导致爬取失败时应给出提示，而不是报错或闪退，对于 B 站没有版权、已失效或分 P 的视频，你的程序应该给出提示或有相应处理方式。
- (2) 按日期爬取弹幕并进行上述分析。
- (3) 分析用到的统计图可以做得非常美观，至少不是简单的使用默认参数生成。
- (4) 根据一定规则进行分词，并生成一个词云（你可能需要一份停用词表）。
- (5) 爬取的弹幕可以导出成 Excel 文件，生成的词云可以导出成 png 文件。
- (6) 弹幕筛选功能：输入一个时间段，筛选出相应时间段出现的弹幕；或输入一串字符，筛选出包含这一串字符的弹幕；或输入一个数字，筛选出长度不大于这个数字的弹幕。增加任意一种筛选功能均可。
- (7) 弹幕筛选功能支持简单的逻辑筛选和正则表达式筛选。
- (8) 常见语言禁止显示乱码，包括但不限于：简中、繁中、英语、日文、韩文。

(9) 以上选做部分不要求全部完成，也不必局限于给出的这些内容。

## 四、实验指导

### 1. 高能进度条

将一个视频分成长度相等的许多小段，通常以视频长度的 1% 作为每一小段的长度，统计每一小段时间出现的弹幕数量，并做成条形统计图，形式上可以尽量接近 b 站原生的样式，当然这一点不做要求。



### 2. API

你可以利用哔哩哔哩提供的 API 轻松获取视频的弹幕信息。由于每个视频有独一无二的 cid，因此你需要先通过另一个 API 获取到对应 BV 号的 cid。

视频 cid API: <http://api.bilibili.com/x/web-interface/view?bvid={BV 号}>

或: <http://api.bilibili.com/x/web-interface/view?avid={AV 号}>

使用指南:

<https://github.com/SocialSisterYi/bilibili-API-collect/blob/master/video/info.md>

data 对象中的 cid 的内容即为对应视频的 cid

弹幕 API: <http://api.bilibili.com/x/v1/dm/list.so?oid={cid}>

使用指南:

<https://github.com/SocialSisterYi/bilibili-API-collect/blob/master/danmaku/danmaku.md>

### 3. 停用词表，如果你找不到停用词表，可以看看下面几个链接:

<https://blog.csdn.net/shijiebei2009/article/details/39696571>

<https://github.com/goto456/stopwords>

当然你也可以使用自己的停用词表。

### 4. PyQt5 | Tkinter: GUI 开发库

Tkinter 是课程内的内容。PyQt5 是一个被广泛使用的 GUI 库，功能比 Tkinter 更强大。

非官方中文教程:

PyQt5: [https://maicss.gitbooks.io/pyqt5/content/hello\\_world.html](https://maicss.gitbooks.io/pyqt5/content/hello_world.html)

Tkinter: <https://www.runoob.com/python/python-gui-tkinter.html>

NumPy | Pandas: 强大的分析结构化数据的工具集

对于一般的二进制文件当然可以用 open() 函数直接实现，但是对于 json、csv 等需要二次处理的文件，或者是 xls 等非二进制文件，建议使用 Pandas 进行处理，同样 pandas 也可以写入这些文本，在数据处理时 Pandas 的 dataframe 也能使整个 workflow 更简化。在处理更大量的数据和多维数组时，利用 NumPy 能更高效和简便地处理数据。

Pandas 官方中文网站: <https://www.py pandas.cn/>

NumPy 官方中文网站: <https://www.numpy.org.cn/>

Matplotlib: 基础的绘图工具停用词表

这个不做过多的说明，课程内的内容。

官方中文网站: <https://www.matplotlib.org.cn/>

其他常用的绘图库还有: ggplot, pycharts



BeautifulSoup4|requests|urllib: 配合使用的强大爬虫工具

通常情况下爬取静态网站, 使用 requests 库中 `get()` 函数可以轻松获取网页源代码, 之后使用 BeautifulSoup4 中的 `a = BeautifulSoup (r.content,"lxml")` 语句将网页源代码进行读取, 并通过 `find()` 方法找到有用的信息。另外, 如果网站上有音频、视频等非文本信息, 可以使用 urllib 库进行爬取。

非官方中文教程:

BeautifulSoup4: [https://beautifulsoup.readthedocs.io/zh\\_CN/v4.4.0/](https://beautifulsoup.readthedocs.io/zh_CN/v4.4.0/)

Requests: <https://www.liaoxuefeng.com/wiki/1016959663602400/1183249464292448>

Urllib: <https://www.liaoxuefeng.com/wiki/1016959663602400/1019223241745024>

## 题目 4：同化棋

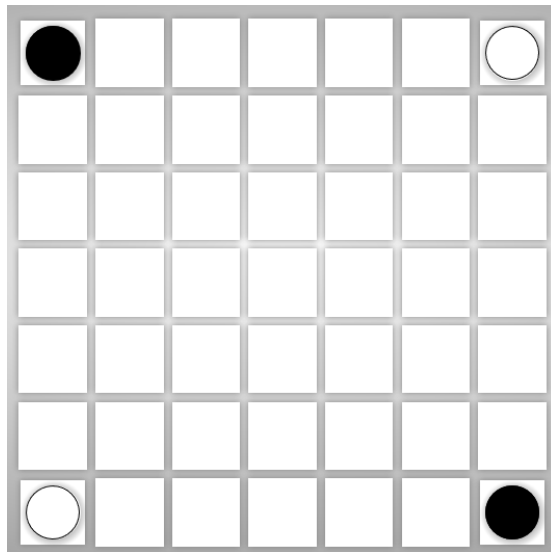
### 一、题目描述

同化棋，英文名 Ataxx，是 Dave Crummack 和 Craig Galley 在 1988 年发明的一种双人棋类，和黑白棋比较类似。1990 年随着电视游戏的出品而风靡世界。

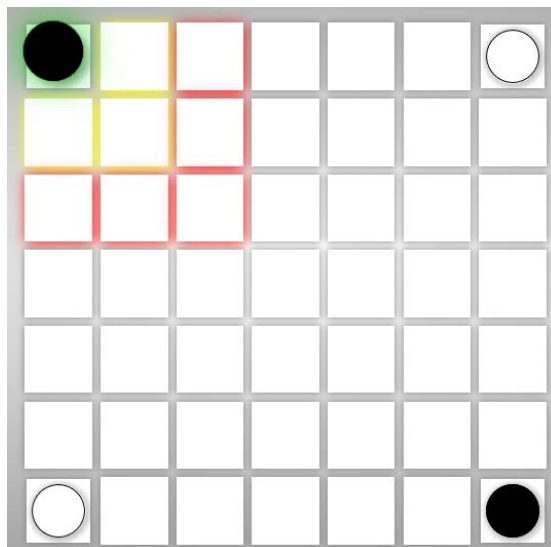
### 二、基本实验要求

你的任务是实现一个能够进行双人同化棋游戏的 GUI 页面。要求包含基本的 7\*7 棋盘，黑白棋子，能够通过鼠标控制的下棋命令等。

注：同化棋的规则如下。一开始的棋盘状态如下图所示：



黑白双方轮流移动棋子，黑方先行。移动时，选取现在在棋盘上己方的一枚棋子，再选择一个落子位置。要求落子处为空，既没有对方棋子也没有己方的棋子。移动有两种方式，如下图所示：

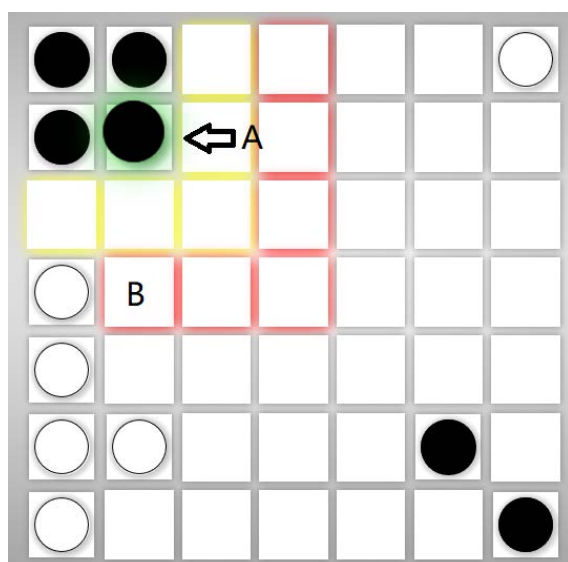


1. 落子位置在以选取的棋子为中心的 3x3 的范围内。此时选取的棋子会复制自身到落子位置。一共有 8 个位置可以选择。

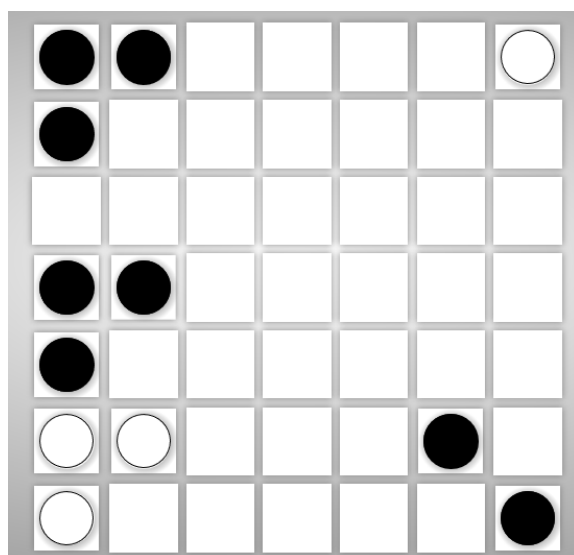
2. 落子位置在以选取的棋子为中心的 3x3 的范围外、5x5 的范围内，此时选取的棋子会移动自身到落子位置，一共有 16 个位置可以选择。

总的来说，第一种移动是“复制粘贴”，第二种移动是“剪切粘贴”。

同化规则：当一个棋子完成移动后，会同化其落点附近的 8 个格子中的棋子，也就是说这 8 个格子内的棋子颜色会变成和刚刚移动的棋子同一种颜色。例如下面的移动过程：



黑子 A 选择移动到正下方的红色格子 B 中，移动后的效果如下：



有两枚白棋被同化。且这次移动是“剪切粘贴”。

若有一方无法移动棋子，则游戏结束，将所有无子的地方算作有子可下一方的棋子，然后数出双方棋子数，棋多者胜。任意时刻对方棋子数为 0 时，己方胜利。如果棋子数相同，视为平局。

### 三、评分细则及加分项

## 1.必做部分

(1) 棋盘棋子：要求 GUI 页面中要有一个 7\*7 的棋盘，上面可以显示黑白两色的棋子。棋盘格式随意，但要有明确的划分线和边界线。如果该项未完成，则大作业直接记为 0 分。

(2) 落子：下棋的过程要求全部使用鼠标操作。不可以通过类似键盘输入坐标的方式完成棋子的移动。落子的过程包括选中一枚棋子，再选中一个空白位置，之后自动完成棋子的移动以及同化过程。应能够判断一些非法行为并给出适当的提示，例如不允许选中对方的棋子，不允许往有棋子的地方移动。该部分分数根据落子的流畅情况以及识别鼠标点击位置的准确性给出。

(3) 胜负判断：在一方无法移动时，应能够及时地终止程序并给出胜负的判断，并给出双方的棋子数之比，例如“黑 30：白 19，黑胜”这样的形式。

## 2.选做部分

(1) 悔棋：可以在棋盘边增加一个悔棋按钮。按下这个按钮后可以实现悔棋操作。比如轮到黑方移动，当其选择点击悔棋按钮后，自动撤销上一次白棋的移动以及上上次黑棋的移动，也就是回到黑方上一次移动时的状态。能够多次悔棋可以再获得一些额外加分。

(2) 移动提示功能：如上面的样例图所示，选中一个棋子后，可以给出移动位置的提示，可以选择用不同颜色区分，也可以用其他方式区分。

(3) 导入&导出棋谱：在一个棋局完成后，可以提供一个导出棋谱按钮，点击后要求能够生成一个棋谱文件，文件后缀任意，棋谱的描述格式自定义，要求在实验报告中说明。在对局开始时，可以提供一个导入棋谱按钮，点击后应弹出一个文件选择页面，选中一个棋谱后，要求棋盘上能够自动按照棋谱行棋。两次移动间应留有足够的时间间隔。

(4) 人机对战功能：在对局开始时，可以额外提供一个模式：人机对战。人机对战模式要求实现一个同化棋 AI，而且可以由玩家自由选择黑方还是白方。对 AI 的实力无过多的要求，但最起码要有一定的智商，例如不能写一个随机移动棋子的程序就当做 AI 提交。发现类似滥竽充数的行为会直接将该部分分数扣光。该部分分数视 AI 实现情况给出。AI 的实现思路见下方实验指导。

(6) 以上选做部分不要求全部完成，也不必局限于给出的这些内容。

## 四、实验指导

1. 实验中可能会用到一些第三方库，可以在 Anaconda 的页面中完成库的安装，也可以按照网上的教程用别的方式完成安装。这方面的教程有很多，就不再赘述。推荐大家使用 Pygame 这个库来实现大作业。Pygame 中封装好了很多现成的函数以及功能，例如鼠标位置的捕捉，游戏背景图片的导入。使用这个库会大大减少工作量，当然也可以选择一些别的库，这里不做强制要求。

2. 有一个 AI 在线对战网站：<https://botzone.org.cn/>。里面有含同化棋在内的大量支持 AI 对战的游戏。注册账号后可以在对应游戏的页面中点击 Bot 排行榜，然后任意选择一个 bot，点击人机对战，就可以和别人写的同化棋 AI 进行在线对战。当然也可以上传自己的同化棋 AI。上传后系统会自动安排你的 AI 和别人的 AI 进行对战，并计算你的天梯分数。如果

在这个网站上上传了 AI，可以在实验报告中注明，最后会根据 AI 的天梯分数给一些额外的加分。

3. 关于 AI 的策略，比较直观一些的是贪心法，例如可以每次都走同化最多对方棋子的下法。但显然这种策略并不是最优的。推荐大家自行学习一点博弈论的相关知识（一点就可以了），使用搜索算法去寻找一定步数内的最优解。比如可以暴力枚举所有可能的走法，模拟双方一定步数的移动，同时给出一个估价函数来分析局面，最后选择局面最好的情况来走棋；也可以使用 **Alpha-Beta** 剪枝来对搜索过程进行优化。AI 的思考时间建议不要设置得太长，也就是说搜索的深度最好不要设置得太深，以优化人机对战的体验。AI 的设计思路有很多，不一定拘泥于上面给出的这些，欢迎大家自由发挥。

## 题目 5：基本文字识别

### 一、问题描述

近几年机器视觉已成为万众瞩目的行业，本题意在鼓励同学们自行钻研这个方向。

选择本题，你需要完成一个程序，能够识别“一二三四五六七八”八个汉字的印刷体和书写体。

程序输入要求至少为  $160 \times 160$  的 01 矩阵，这可以表示一个  $160 \times 160$  的黑白图片，而“至少”的具体含义将在加分项中解释。你的程序需要通过计算来得到这个黑白图片中的文字是“一二三四五六七八”中的哪一个。

### 二、基本实验要求

本题中你可以**不设计** GUI 相关内容。

本题的得分极大程度取决于你的程序在 **2s** 内对所要求的“ $160 \times 160$  黑白两色文字图片”的识别准确率。


我们将给出有关数据的详细信息：

1. 测试集 1：内含 40 张  $160 \times 160$  黑白两色图片，平均包含八个文字的**印刷体**。
2. 测试集 2：内含 40 张  $160 \times 160$  黑白两色图片，平均包含八个文字的**书写体**。
3. 数据集：内含 400 张  $160 \times 160$  的黑白两色图片，其中每个文字有 50 张，且已标注。

注：测试集 2 和数据集共 440 张来源于 55 组手写体文字，每组手写体文字中最多只有一张文字图片在测试集中，其余七个文字**乱序**分在数据集中。每张图片中只有一个汉字。

以上三种数据只有数据集公开，测试集图片不会公开，但给同学们提供测试。同学们需自行在北航云盘自行下载数据集：

<https://bhpan.buaa.edu.cn:443/link/89B767156874BF8F3A384D84059BE92B>

数据集例图：

具体图片可云盘下载查看。所有测试集与数据集归课程组所有，除完成大作业外，请勿作他用。

此外，我们**不推荐**直接使用机器学习理论以及框架，但这些均在加分项中；我们推荐同学们从自己的想法入手，通过巧妙的平面图计算，研究文字的特征，通过  $160 \times 160$  的矩阵，来完成任务，具体方式在实验指导中有说明。

在编程过程中，你可以随时在给定测试集上测试代码，但不能获得测试集。测试方法为：在 [luogu.org](https://www.luogu.com.cn/team/26730) 上注册账号并申请加入 <https://www.luogu.com.cn/team/26730> 团队，加入团队后可看到测试比赛，比赛邀请码为“gcmf”（不含引号），比赛中有两道题目，分别对应两个测试集。注：**luogu.org 的 py 环境中**有 NumPy 库，但没有 pytorch。

题目的输入为  $160 \times 160$  的 01 块（1 代表白色），输出为数字 1 到 8 中的一个，例如：

11111

10001

11111

是一个  $5 \times 3$  的 01 块（并且大部分人眼中看起来比较像“一”）。每个测试点的时限为

2s, 内存 128MB。在提交后你将很快得到你的代码在两种测试集上的准确率。显然提交次数没有限制。测试集会在结课后公开。

### 三、评分细则及加分项

1. 程序的准确率是指：代码在给定两个测试集上的正确率平均数。得到的方式是直接将代码提交到洛谷 oj 上，由洛谷返回准确率。

2. 在不使用现有深度学习框架的情况下，你的大作业总得分分为报告与答辩得分和准确率得分加权，其中：

(1) 报告与答辩得分占 20%。主要介绍你所使用的手段，通过什么样的特征判断文字，讲解实验中学到的知识。

(2) 准确率得分占 80%。在准确率的 100 分中，20%-60%的准确率线性映射至 0-80 分。60%-80%的准确率线性映射至 80-100 分，80%以上则准确率得分为满分。（举例来说，某同学代码识别准确率是 60%，报告与答辩得分是满分，则大作业得分  $100 \times 0.2 + 80 \times 0.8 = 84$  分。

(3) 由于实验中只有八种输出，实际上提交 `print("1")` 能获得 12.5%的准确率，所以我们认为 20%准确率以下的程序没有参考价值。

(4) 请避免使用随机方法来计算。在最终测试中我们会用每个测试集测试你的程序三次，取最低的准确率（举例来说，你的程序在第一次测试时运行两个测试集的准确率分别为 100%和 0%，在第二次测试时为 0%和 100%，那么你的程序准确率为 0%）。

3. 加分项：

(1) 在程序运行**自己本地的测试**的过程中，能够连续展示、可视化当前所识别的图片，在连续的弹窗中显示识别的**正确**结果至少五张。可加分 0-10 分。

(2) 程序可以兼容其他格式或其他颜色的矩阵输入（但请注意你的准确度得分仍然取决于对于给定测试集的精确识别）。识别 300000 像素点及以上的图片可加分 0-5 分。识别 rgb 三色输入的**彩色**图片（如：艺术字）可加分 0-5 分。识别更多文字(需超过 10 个笔画不少于 3 个的汉字，且准确率不低于 50%)可加分 0-10 分。

(3) 在程序设计过程中，使用机器学习相关方法（要求程序在训练所使用的数据集中达到**99%准确率**，且程序中至少含有一个**超参数**并且报告中说明该参数在代码中的实际作用，以及训练该参数所使用的方式）。可加分 0-10 分。

注：“（3）”的要求是因为：数据集上不能达到 99%的模型一般被认为不够收敛，不收敛的原因需要大家自己寻找并解决，这是机器学习的常见问题之一。另外“超参数”应是用来计算结果的重要参数（可能是阈值，比如高度<20 的认为是一横，>=20 认为这个笔画不是横，其中的 20 应当是训练或测试中得到的），不能是无关紧要的某项数值，

(4) 在程序设计过程中，使用深度学习框架(包括但不限于 Tensorflow,pytorch 等)，在此条件下，你的评分方式**可以**改为使用现有深度学习框架的评分方式。（也可以选择使用上面“2”的评分方式）但你的实验报告需叙述完整应用各个接口的细节，描述完整你的网络结构、训练日志。

(5) 加分后实验总分不超过 100 分。

4. 使用深度学习框架的**可选**评分方式：难度系数分、答辩报告分、准确度得分加权，

其中：

(1) 难度系数分可占 0-80%，在计算过程中确实使用了已有框架表示学习网络，并使用学习网络来计算所要求的图片，该项获得满分。

(2) 答辩报告分可占 0-10%，主要讲述代码中用到的主要框架，各层类型，网络结构，以及训练手法。

(3) 准确度得分可占 0-20%，由 20%-60%的准确度线性折合得到，**意味着使用成熟框架完成且具有 60%以上准确度则可以不答辩(需要提交报告)而获得大作业满分。**

(4) 注意，使用深度学习框架的写法仍然可以选择使用“不用框架的评分方式”。

(5) 注意，使用该方法代码将不能提交洛谷，具体评测方法如下：若现场答辩，将使用 U 盘拷贝数据文件，格式为一个文件夹“test”下 40 个 png 图片（图片的名字分别为 1.png,2.png,...,40.png），请使用该方法的同学自行准备接口，测试后将删除测试图片；如在线答辩，将统一发放数据文件，所有同学在 10 分钟内返回一个 txt 文档（由程序输出），一共 80 行，每一行两个数字 ab 用空格隔开，a 表示图片编号(1 到 40)，b 表示识别结果（1 到 8）。对于该种评测方式，由于评测数据格式已给出，大家可以自行先测试程序能否正常运行。若测试过程中出现“程序运行出错”或“文件权限出错”等问题，导致无法得到规范的识别结果，或无法在给定时间内返回结果，视为准确度 0%。

(6) 注意，使用此方法我们没有限制数据集，尤其是训练集，给出的 400 张图片供参考。

## 四、实验指导

1. 本题推荐使用平面处理来完成。例如：横纵向跨度（比如汉字“一、二、三”的纵轴跨度可能较小，而“七、八”等字的纵轴跨度可能较大）等平面图特征，其他关键判断需同学们自行思考。

2. 在图像处理时可使用 opencv 包，一般来说可使用“pip install opencv-python”来安装。其具体指令较简单，包括 imread 读入图片、imwrite 输出绘图等简单语句，使用这些语句可以在运行中间步骤输出当前图片的样子，以方便参考，思考接下来如何判断文字。但洛谷没有这个包，提交时需删除展示图片相关代码。

3. 在读入输出文件时可能涉及目录操作，在 Linux 中可能还会涉及权限操作，这些操作并不复杂，一般在学习过程中即用即查即可，例如百度“windows 中打开 D 盘目录的指令”，逐渐熟练使用这些语句也会使后续其他课程或项目中接触这些内容时更轻松。

4. 对于想要借此学习机器学习的同学，原理入门视频：<https://b23.tv/QDWFGFR>。由于本题要求十分简单，不建议使用较强的结构，但对于希望深入学习这方面的同学，可以通过学习经典结构“AlexNet, ResNet”等(连续几年内出现的五六个左右)入手，通过读论文，了解当前网络结构大体构造，以及目前人们对各种网络层的作用理解。还有一些推荐的网课及资料：

莫凡 Pytorch: <https://www.bilibili.com/video/BV1Vx41lj7kT>

李宏毅机器学习: <https://www.bilibili.com/video/BV13x41lv7US>

斯坦福大学机器学习中文笔记: <http://www.ai-start.com/ml2014/>

Python 机器学习的应用: <http://www.ai-start.com/ml2014/>