

## 第 5 章：MATLAB 文本数据处理入门篇课后习题

（本文档节选自数学建模清风老师的 MATLAB 入门课程）

b 站观看地址：<https://www.bilibili.com/video/BV1dN4y1Q7Kt>

### 本章小节

本章全面且深入地探索了 MATLAB 中文本数据处理的核心概念和技术，以下是各节概述：

- ASCII和Unicode编码：本节详细介绍了字符在计算机中的存储方法，包括ASCII和Unicode编码。这部分内容为理解MATLAB中文本数据的处理奠定了重要的基础。
- 字符数组：本节深入探讨了字符数组的创建和操作技巧。通过丰富的示例，本节展示了如何有效地创建和操纵字符向量和字符矩阵，突出了字符数组在 MATLAB 中的灵活性和重要性。
- 元胞数组的应用：本节讨论了元胞数组的使用方法和应用场景。我们详细介绍了元胞数组在存储和管理不同大小、不同类型数据的优势，然后具体讲解了如何利用字符向量元胞数组和基础的文本处理函数进行文本分析。
- 字符串数组的操作：本节全面介绍了字符串数组的操作和应用方法。内容涵盖了 MATLAB 中各种文本处理函数的使用，例如字符串的搜索、替换、拼接、拆分等。这些函数的实际应用展示了 MATLAB 在文本数据处理方面的卓越能力和灵活性。

另外，本章 5.3.3 节和 5.4.5 节提供了一系列实用的文本处理案例。这些案例不仅展示了如何运用所学的知识点高效处理和分析文本数据，也进一步加深了大家对上述知识点的理解。

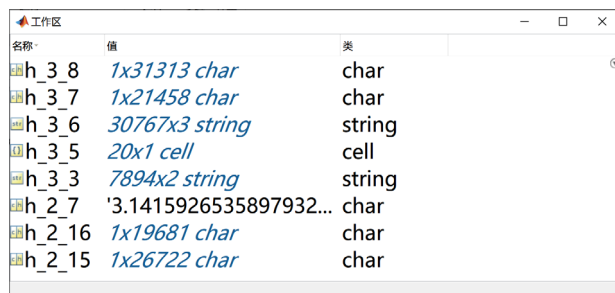
### 课后习题

由于本章没有介绍如何导入和导出文本数据（下一章中会讲解），因此我提前准备了一些数据，大家做课后习题之前请先使用下面的代码导入对应的数据：

*load homework5.mat*

（导入数据前，确保 MATLAB 的当前文件夹下存在 homework5.mat 这个文件。  
不会导入的同学可以参考本章 5.3.3 节或者看讲解视频）

导入成功后你将在 MATLAB 的工作区看到一些变量：



名称	值	类
h_3_8	1x31313 char	char
h_3_7	1x21458 char	char
h_3_6	30767x3 string	string
h_3_5	20x1 cell	cell
h_3_3	7894x2 string	string
h_2_7	'3.1415926535897932...	char
h_2_16	1x19681 char	char
h_2_15	1x26722 char	char

变量的命名规则如下所示： $h_{i_j}$ 。h 是 homework 的缩写、i 表示题目类别（i=1,2,3 分别对应基础篇、提高篇和挑战篇）、j 表示题号，例如  $h_{2_7}$  表示提高篇第七题的数据。

作业参考答案可在 b 站观看：《MATLAB 课程第 5 章课后习题讲解--数学建模清风老师》，  
播放地址：<https://www.bilibili.com/video/BV1Dg4y1S7u7>

## 基础篇

### Q1: 填空题

1. 本章介绍了两种对字符数据进行编码的方式，分别是\_\_\_\_\_编码和\_\_\_\_\_编码。
2. 创建空的元胞数组可以使用\_\_\_\_\_函数。
3. 引用元胞数组中的数据可以使用\_\_\_\_\_进行索引。
4. 在处理元胞数组时，我们经常需要对数组中保存的每个数据应用相同的函数进行计算，MATLAB 提供的\_\_\_\_\_函数可以帮我们快速实现这个目的。当我们在该函数中使用的函数句柄返回的不是标量值时，应该设置\_\_\_\_\_参数。
5. 要创建一个空的字符串数组，可以使用\_\_\_\_\_函数。
6. 使用\_\_\_\_\_函数能够比较两个字符向量是否完全相同；如果不区分大小写则可以使用\_\_\_\_\_函数。
7. 要获取字符串中字符的数量，可以使用\_\_\_\_\_函数。
8. 连接两个字符串标量可以使用\_\_\_\_\_运算符。
9. 要将字符串数组中的元素按特定分隔符进行连接，可以使用\_\_\_\_\_函数。
10. 使用\_\_\_\_\_函数可以删除字符串数组中的子文本。
11. 使用\_\_\_\_\_函数能够将数组转换为元胞数组，转换后的元胞数组中的数据大小相同。
12. 将文本中的大写字母转换为小写字母的函数是\_\_\_\_\_；将文本中的小写字母转换为大写字母的函数是\_\_\_\_\_。
13. 为字符串添加前导或尾随字符，以达到特定长度的函数是\_\_\_\_\_；使用\_\_\_\_\_函数可以调整字符串数组中文本的对齐方式。
14. 函数\_\_\_\_\_只会删除文本末尾的空白字符，不会删除开头的空白字符。
15. 创建一个换行符的命令是\_\_\_\_\_。
16. 判断字符串中是否包含特定模式的函数是\_\_\_\_\_。
17. 判断两个元胞数组是否等效可以使用\_\_\_\_\_函数。
18. 反转字符串中的字符顺序的函数是\_\_\_\_\_。
19. 已知变量 `s="123"`，那么命令 `s{1}` 返回的结果是\_\_\_\_\_。
20. 已知 `cc = {[1 2 3;4 5 6],'abc'}`，那么 `cc{2}(3)` 返回的结果是\_\_\_\_\_；这种索引方式称为\_\_\_\_\_；使用这种索引方式进行索引时，需要注意\_\_\_\_\_，如果不遵守的话会报错。
21. 检查字符串是否以指定的文本开始可以使用函数\_\_\_\_\_。
22. 有时候一行代码很长，为了便于阅读和理解，我们可以使用\_\_\_\_\_将这行代码分割到多行。
23. 转义字符用于在文本中表示特定的字符，例如\_\_\_\_\_表示换行符。
24. 使用\_\_\_\_\_函数可以显示工作区各变量的详细信息，包括变量的名称、大小、占用的内存大小和数据的属性。
25. 统计字符串中特定模式出现的次数的函数是\_\_\_\_\_。
26. 根据指定分隔符拆分字符串数组的函数是\_\_\_\_\_。
27. 使用\_\_\_\_\_函数可以将包含数值的文本数据类型转换回数值数组，它支持字符数组、字符向量元胞数组和字符串数组三种类型。
28. 在指定的起点和终点之间替换子字符串可以使用\_\_\_\_\_函数；如果是提取起点和终点之间的子字符串可以使用\_\_\_\_\_函数。
29. 将旧文本替换成新的文本有两个不同的函数，分别是\_\_\_\_\_和\_\_\_\_\_。
30. 将输入变量分发给输出变量的函数是\_\_\_\_\_。

## Q2: 代码练习题

下面给出了一个任务表，表中每个编号对应一个特定的数据处理任务。你需要根据所给的输入数据，编写相应的代码来生成对应的输出数据。例如：任务 1 的答案为：`cc = upper(c)`

编号	任务说明	输入的数据	输出的数据
1	将 <code>c</code> 中的小写字母变成大写	<code>c = 'abcDEF123'</code>	<code>cc = 'ABCDEF123'</code>
2	将 <code>c</code> 转换成数值向量 <code>d</code>	<code>c = '1.5 4.3 6.5'</code>	<code>d = [1.5 4.3 6.5]</code>
3	删除 <code>s</code> 中由 <code>&lt;&gt;</code> 构成的 <code>html</code> 标签	<code>s = "&lt;a&gt;&lt;div&gt;abcd&lt;/div&gt;&lt;/a&gt;"</code>	<code>ss = "abcd"</code>
4	将 <code>cc</code> 转换成字符串数组	<code>cc = {'ab','cde'}</code>	<code>ss = [ "ab", "cde"]</code>
5	判断 <code>c</code> 中是否存在小写英文字母	<code>c = 'AB123456CD'</code>	<code>flag = logical 0</code>
		<code>c = 'AB123456cd'</code>	<code>flag = logical 1</code>
6	删除 <code>s</code> 中的所有数字	<code>s = "a325ds0d4s5a4s7"</code>	<code>ss = "adsdsas"</code>
7	删除 <code>s</code> 中的空字符串	<code>s = ["", "ab", "c", "", "d"]</code>	<code>s = ["ab", "c", "d"]</code>
8	将 <code>x</code> 分割为四个子块，并将结果保存到元胞数组 <code>c</code> 中	<code>x = [1 2; 3 4; 5 6]</code>	<code>c = 2×2 cell 数组 {[1]} {[2]} {[3;5]} {[4;6]}</code>
9	统计 <code>s</code> 中大写字母出现的次数	<code>s = "aAdE55G6F"</code>	<code>num = 4</code>
		<code>s = "abcdefg"</code>	<code>num = 0</code>
10	在空格处拆分字符串	<code>s = "ab de f g"</code>	<code>ss = ["ab", "de", "f", "g"]</code>
11	删除 <code>s</code> 中各元素开头的空白字符	<code>s = [" abc ", "def", " g"]</code>	<code>ss = ["abc ", "def", "g"]</code>
12	计算 <code>s</code> 中各元素的频数表	<code>s = ["22", "21", "21", "22", "11"]</code>	<code>c = 3×3 cell 数组 {[22]} {[2]} {[40]} {[21]} {[2]} {[40]} {[11]} {[1]} {[20]}</code>
13	确定 <code>c</code> 中的哪些字符属于数字	<code>c = 'ad25me0'</code>	<code>ind = 1×7 logical 数组 0 0 1 1 0 0 1</code>
14	判断 <code>s</code> 中的元素是否以数字开头	<code>s = ["1abc", "5ac", "cc12"; "css", "9df88", "43"]</code>	<code>ind = 2×3 logical 数组 1 1 0 0 1 1</code>
15	删除元胞数组 <code>C</code> 的第二行元素	<code>C = {'apple', 'banana'; 'pear', 'cherry'}</code>	<code>C = {'apple', 'banana'}</code>
16	将 <code>C</code> 的第一列数据换成 <code>'xyz'</code>	<code>C = {'apple', 'banana'; 'pear', 'cherry'}</code>	<code>C = {'xyz', 'banana'; 'xyz', 'cherry'}</code>
17	计算 <code>C</code> 中每个向量的最大值	<code>C = {[1 5], [6 7 1], [4 0 9 2]}</code>	<code>mc = [5 7 9]</code>
18	对 <code>C</code> 中的每个向量分别升序排列	<code>C = {[6 7 1], [4 0 9 2]}</code>	<code>sc = {[1 6 7], [0 2 4 9]}</code>
19	将 <code>C</code> 中的数据拼接成一个向量	<code>C = {[1 5], [6 7 1], [4 0 9 2]}</code>	<code>d = [1 5 6 7 1 4 0 9 2]</code>
20	删除 <code>C</code> 中元素和小于 10 的向量	<code>C = {[1 5], [6 8], [2 3], [4 6]};</code>	<code>C = { [6 8], [4 6]}</code>

## 提高篇

## Q1: DNA 序列分析

A, T, C, G(腺嘌呤、胸腺嘧啶、鸟嘌呤和胞嘧啶)是生物 DNA 中的四种碱基，它们是构成 DNA 的四种基本单位，通过不同的排列组合编码着生物的遗传信息。请完成以下问题：

- (1) 随机生成长度为 9000 的一段 DNA 序列，例如'TCGGTTTCAG...'，保存为变量 `dna`（注意，随机生成 DNA 序列在生物学中并不科学，本题仅供练习 MATLAB 的语法）。
- (2) 计算每种碱基（A、T、C、G）在 `dna` 中出现的次数，将结果保存为长度为 4 的向量 `P` 中。
- (3) 假设序列中的 'T' 碱基可以被 'U' 替代（U 为尿嘧啶），返回转换后的结果 `rna`。
- (4) DNA 序列中每 3 个连续的碱基表示一个密码子，因此变量 `dna` 中应存在 3000 个密码子，请将这 3000 个密码子保存到字符串向量 `S` 中，并统计有多少种不同的密码子。
- (5) 假设 `S` 中有 `k` 种不同的密码子，计算一个大小为  $k \times 3$  的元胞数组 `C_DNA`，`C_DNA` 的第一列表示这 `k` 种不同的密码子，第 2 列表示它们出现的次数，第三列表示它们出现的频率（你能不使用 `tabulate` 函数得到 `C_DNA` 吗？）。

## Q2: 优化 5.3.3 节案例 2（计算共有的兴趣爱好数量）的代码

课堂上给出的代码中，每次比较两名同学的兴趣爱好时，都会重复执行 `strsplit` 函数来获取兴趣列表。你能否优化代码来提高程序的运行效率？比较代码优化前后的运行时间。

## Q3: 统计文本中各字符出现的频次

`s` 是一个字符串标量，请统计 `s` 中各字符出现的频次，并生成一段描述结果的文本 `T`。`T` 是一个带有换行符的字符串标量，具体格式请见下表（按照字符出现的频次从高到低排序）：

示例 <code>s</code>	期望得到的 <code>T</code>
"aaccdcc"	"字符 'c' 出现了 4 次 字符 'a' 出现了 2 次 字符 'd' 出现了 1 次"
"3.141592653589793238462643383279502884197" % 右侧结果中有一些字符的频数相同，如果频数相同则按照字符的 Unicode 编码从小到大进行排序	"字符 '3' 出现了 7 次 字符 '2' 出现了 5 次 字符 '8' 出现了 5 次 字符 '9' 出现了 5 次 字符 '4' 出现了 4 次 字符 '5' 出现了 4 次 字符 '1' 出现了 3 次 字符 '6' 出现了 3 次 字符 '7' 出现了 3 次 字符 '.' 出现了 1 次 字符 '0' 出现了 1 次"
"人要是行，干一行行一行。"	"字符 '行' 出现了 4 次 字符 '一' 出现了 2 次 字符 '。' 出现了 1 次 字符 '人' 出现了 1 次 字符 '干' 出现了 1 次 字符 '是' 出现了 1 次 字符 '要' 出现了 1 次 字符 ',' 出现了 1 次"

**Q4: 将十进制正整数转换为十六进制数**

本章 5.2.2 节中，我们有一道将十进制正整数转换为二进制数的例题。请仿照这个例题，写一段代码将十进制正整数转换为对应的十六进制数。提示：十六进制数由 0 1 2 3 4 5 6 7 8 9 A B C D E F 组成。它与十进制的对应关系是：0-9 对应 0-9、A-F 对应 10-15。

例如：十进制数 666 对应的十六进制数为'29A' ( $666 = 2 \times 16^2 + 9 \times 16 + 10$ )、十进制数 7788 对应的十六进制数为'1E6C' ( $7788 = 1 \times 16^3 + 14 \times 16^2 + 6 \times 16 + 12$ )。

**Q5: 找出所有能被 `deblank` 函数去除的空白字符**

`deblank` 函数能够去除文本末尾的空白字符，请你写一段程序，找出它能识别的所有空白字符对应的 Unicode 编码，并将结果保存到一个向量中。（在 Unicode 字符集中，十进制 0 到  $2^{16}-1$ （十六进制 0000-FFFF）涵盖了绝大多数常用字符，包括各种语言的文字、符号以及特殊字符，因此识别的字符范围可以限定在这个区间内）

**Q6: 使用字符构建一颗圣诞树**

如下表所示，表格左侧有一个 14 行 19 列的字符数组 `cc`，它仅由字符 '\*' 和 '|' 构成，为了美观我使用了绿色表示里面的字符元素，这样看起来有点像一颗圣诞树。其中，'\*' 构成了树叶（有 10 行），'|' 构成了下面的树干（有 4 行）。

类似的，表格右侧有一个更小的圣诞树，它的树叶有 6 行，树干有 3 行。

现在给定树叶的行数  $n$  和树干的行数  $m$  ( $n$  和  $m$  均为正整数)，请你构造一个表示圣诞树的字符数组 `cc`，并使用 `disp(cc)` 输出结果。（你也可以使用字符串数组类型表示这颗圣诞树）

<pre>cc = 14×19 char 数组</pre> <pre>       *      ***     *****    *********   ***********  ***** ***** ***** ***** ***** ***** ***** ***** ***** </pre>	<pre>cc = 9×11 char 数组</pre> <pre>       *      ***     *****    *****   *****  *****             </pre>
---	--

**Q7: 有趣的圆周率  $\pi$ （本题数据为 `h_2_7`）**

请按照本小节开头的提示导入好作业的数据，本题用到的数据为 `h_2_7`，它是一个字符向量，保存着小数点后 10000 位的圆周率  $\pi$  ('3.141592653589793238...')。

- 验证  $\pi$  的小数点后 144 位数字相加的和 ( $1+4+1+5+9+2+\dots$ ) 等于 666。
- 统计  $\pi$  的小数点后 10000 位中各个数字出现的频数和频率。
- 假设你的生日为 4 月 6 日，将其转换为字符向量为 '0406'，验证能否在  $\pi$  的小数点后 10000 位中找到这个子文本。将 '0406' 换成你自己的真实生日，能找到吗？
- 构造所有可能的生日（366 种可能），将其保存在一个字符串数组中。判断哪些生日能在  $\pi$  的小数点后 10000 位中找到，若能找到返回其第一次出现的小数点位数。



**Q8: 判断字符向量能否作为 MATLAB 中的变量名**

在第二章中，我们介绍过 MATLAB 中变量的命名规则：

- 变量名必须以字母开头，之后可以是任意的字母、数字或下划线\_。
- 变量名不超过 63 个字符
- 不能定义与 MATLAB 关键字同名的变量（例如 if 或 end）。要获取关键字的完整列表，可以使用 iskeyword 函数。

给你一个字符向量 c，判断它能否作为 MATLAB 中的变量名。

**Q9: 探索无限非循环小数中特定数字序列的位置**

已知小数 0.1234567891011121314... 是一个由连续自然数拼接组成的无限不循环小数，问小数点后面出现的第一个 2019 中的 2 是小数点后面的第几位？（本题选自知乎，答案是 6572）

**Q10: 循环移位字符向量生成字符串数组**

给定一个字符向量 c，请编写代码生成一个字符串数组 s，其中 s 的每个元素都是 c 的循环移位。下面举两个例子帮助大家理解：

c = 'abcde'	s = 5×1 string 数组 "abcde" "eabcd" "deabc" "cdeab" "bcdea"
c = 'x0x1'	s = 4×1 string 数组 "x0x1" "1x0x" "x1x0" "0x1x"

**Q11: 字符组合的全排列生成**

给定 n 个两两互不相同的字符，返回所有可能的排列，将结果保存到一个字符串向量 s 中，并进行排序。例如 'a'、'b'、'c' 这三个字符的全排列有六种情况，对应的 s 为：

```
s =
6×1 string 数组
"abc"
"acb"
"bac"
"bca"
"cab"
"cba"
```

（提示：你可能需要用到第三章课后习题挑战篇 Q19 中介绍的一个函数）

**Q12: 模拟生成高考数学单选题的随机答案**

在高考数学考试中，有八道单选题，每题的标准答案分别为 DCCAABBC。一名考生对考试内容一无所知，因此不得不随机猜测每道题的答案。这名考生有一种倾向：他更可能选择 C 选项。具体来说，他有 40% 的几率选择 C，而选择 ABD 的几率各为 20%。

（1）请根据这个概率分布帮助这名考生随机生成一组八道题的答案，并计算出他正确答对的题目数量。

（2）进行十万次模拟，计算出在这种答题策略下，他平均能正确答对多少道题。

**Q13: 数字之和与整除条件的数字分析任务**

编写程序依次完成以下三个任务：

**(1) 计算数字之和满足特定条件的整数：**

对于 1 至 10000 范围内的每个整数，计算它们每一位数字之和（例如 135 这个整数的数字之和为  $8 = 1+3+5$ ）。

创建一个  $10 \times 1$  的元胞数组  $x$ 。对于每个  $k$ （1 到 10 之间的整数）， $x\{k\}$  应包含那些其数字之和能被  $k$  整除的整数列表。例如：

$x\{10\}$  包含的整数是每位数字之和能被 10 整除的，比如 19（ $1+9=10$ ）或 9993（ $9+9+9+3=30$ ）。

**(2) 统计元胞内各数据的元素数量：**

统计  $x\{k\}$  中包含的元素数量，并将这些计数结果保存在一个长度为 10 的向量  $y$  中。

例如， $y$  的第一个元素应为 10000，因为 1 到 10000 中的每个数的数字之和都能被 1 整除。

**(3) 统计数字出现次数和频率：**

统计 1 至 10000 中每个数在  $x$  中出现的次数和频率，并按次数降序排列结果。

**Q14: 数独辅助解题器：提示每个位置可填入的数字**

本章 5.3.1.8 节介绍 `mat2cell` 函数时，我们讲过一个数独的例题。本题需要大家编写一段程序来辅助解决数独谜题。

数独是一个  $9 \times 9$  的网格，其中部分单元格已填入数字，剩余的单元格需要根据数独的规则来填写。数独的规则要求每行、每列以及每个  $3 \times 3$  的小网格（宫）中的数字 1 到 9 各出现一次。

	<b>% 用 0 表示要填充的空单元格</b> <code>sd = [0 0 6 7 0 2 3 0 0;  0 0 0 4 0 3 0 0 0;  3 0 0 0 1 0 0 0 6;  9 8 0 6 0 1 0 5 4;  0 0 5 0 0 0 2 0 0;  7 4 0 3 0 9 0 1 8;  4 0 0 0 3 0 0 0 2;  0 0 0 1 0 7 0 0 0;  0 0 8 2 0 6 5 0 0];</code>								
--	---	--	--	--	--	--	--	--	--

给定上面这个部分填写的数独盘面，将每个空单元格可以填入的数字列表保存在一个  $9 \times 9$  的元胞数组  $C$  中。如果某个单元格已有确定数字，则该单元格在  $C$  中对应的位置保持原始数字。例如  $C$  中第一行第一列的数据为  $[1,5,8]$ ，代表该位置只可能填写 1、5 和 8。

完整的答案如下所示，供大家参考：

`C = 9x9 cell`

	1	2	3	4	5	6	7	8	9
1	[1,5,8]	[1,5,9]	6	7	[5,8,9]	2	3	[4,8,9]	[1,5,9]
2	[1,2,5,8]	[1,2,5,7,9]	[1,2,7,9]	4	[5,6,8,9]	3	[1,7,8,9]	[2,7,8,9]	[1,5,7,9]
3	3	[2,5,7,9]	[2,4,7,9]	[5,8,9]	1	[5,8]	[4,7,8,9]	[2,4,7,8,9]	6
4	9	8	[2,3]	6	[2,7]	1	7	5	4
5	[1,6]	[1,3,6]	5	8	[4,7,8]	[4,8]	2	[3,6,7,9]	[3,7,9]
6	7	4	2	3	[2,5]	9	6	1	8
7	4	[1,5,6,7,9]	[1,7,9]	[5,8,9]	3	[5,8]	[1,6,7,8,9]	[6,7,8,9]	2
8	[2,5,6]	[2,3,5,6,9]	[2,3,9]	1	[4,5,8,9]	7	[4,6,8,9]	[3,4,6,8,9]	[3,9]
9	1	[1,3,7,9]	8	2	[4,9]	6	5	[3,4,7,9]	[1,3,7,9]

接下来，请将变量  $C$  保存在名为“我的第一个数据  $C.mat$ ”的 MATLAB 数据文件中，然后重启 MATLAB 软件并重新导入这个数据。

## Q15: 提取古诗名称和诗人 (本题数据为 h\_2\_15)

请按照本小节开头的提示导入好作业的数据, 本题用到的数据为 h\_2\_15, 它是一个字符向量, 来自某个古诗词网站, h\_2\_15 的开头如下所示:

```
'<div class="sons">
<div class="typecont">
<div class="bookMI">五言绝句</div>
<span><a href="/shiwenv_45c396367f59.aspx" target="_blank">行宫</a>(元稹)</span>
<span><a href="/shiwenv_c90ff9ea5a71.aspx" target="_blank">登鹳雀楼</a>(王之涣)</span>
<span><a href="/shiwenv_5917bc6dca91.aspx" target="_blank">新嫁娘词</a>(王建)</span>
<span><a href="/shiwenv_f324eea45183.aspx" target="_blank">相思</a>(王维)</span>
<span><a href="/shiwenv_8d889937d1fe.aspx" target="_blank">杂诗</a>(王维)</span>
<span><a href="/shiwenv_e9b1a8b4def0.aspx" target="_blank">鹿柴</a>(王维)</span>
<span><a href="/shiwenv_4809b5e7a16a.aspx" target="_blank">竹里馆</a>(王维)</span>
<span><a href="/shiwenv_6368d3d62fcd.aspx" target="_blank">山中送别</a>(王维)</span>
<span><a href="/shiwenv_d09fef17613b.aspx" target="_blank">问刘十九</a>(白居易)</span>
<span><a href="/shiwenv_94eb5d41fec6.aspx" target="_blank">哥舒歌</a>(西鄙人)</span>
<span><a href="/shiwenv_c35a60c1a8e2.aspx" target="_blank">静夜思</a>(李白)</span>
<span><a href="/shiwenv_68fe1f940020.aspx" target="_blank">怨情</a>(李白)</span>
<span><a href="/shiwenv_ee9af27de9f2.aspx" target="_blank">登乐游原</a>(李商隐)</span>
<span><a href="/shiwenv_ed8b644fd298.aspx" target="_blank">听箏</a>(李端)</span>
```

(1) 请从 h\_2\_15 中提取出所有的古诗名称和对应的诗人, 并将结果保存在一个两列的字符串数组 S 中。S 中应有 320 行, 结果供大家参考:

```
S = 320x2 string
    "行宫"      "元稹"
    "登鹳雀楼"  "王之涣"
    "新嫁娘词"  "王建"
    "相思"      "王维"
    "杂诗"      "王维"
    "鹿柴"      "王维"
    "竹里馆"    "王维"
    "山中送别"  "王维"
    "问刘十九"  "白居易"
    "哥舒歌"    "西鄙人"
    ...
    ...
    ...
```

(2) 请统计 S 中有多少个不同的诗人, 计算这些诗人古诗的数量和频率, 你可以将结果保存到元胞数组 C 中, 并按照古诗的数量降序排列。C 中应有 75 行, 结果供大家参考:

C = 75x3 cell

	1	2	3
1	'杜甫'	39	12.2642
2	'李白'	34	10.6918
3	'王维'	29	9.1195
4	'李商隐'	24	7.5472
5	'孟浩然'	15	4.7170
6	'韦应物'	12	3.7736
7	'刘长卿'	11	3.4591
8	'杜牧'	10	3.1447
9	'王昌龄'	7	2.2013



### Q16: 提取王者荣耀游戏中英雄的名称（本题数据为 h\_2\_16）

请按照本小节开头的提示导入好作业的数据，本题用到的数据为 h\_2\_16，它是一个字符向量，来自王者荣耀官网，h\_2\_16 的开头如下所示：

```
海诺</a></li><li><a href="herodetail/duoliya.shtml" target="_blank">朵莉亚</a></li><li><a href="herodetail/yalian.shtml" target="_blank">亚连</a></li><li><a href="herodetail/jixiaoman.shtml" target="_blank">姬小满</a></li><li><a href="herodetail/laixiao.shtml" target="_blank">莱西奥</a></li><li><a href="herodetail/zhaohuaizhen.shtml" target="_blank">赵怀真</a></li><li><a href="herodetail/haiyue.shtml" target="_blank">海月</a></li><li><a href="herodetail/geya.shtml" target="_blank">
```

(1) 请从 h\_2\_16 中提取出所有英雄的名称（参考上图箭头指示的位置），并将结果保存到字符串向量 S 中。S 中应有 117 行，结果供大家参考：

S = 117×1 string

```
"海诺"
"朵莉亚"
"亚连"
"姬小满"
"莱西奥"
"赵怀真"
"海月"
"戈娅"
"桑启"
"暃"
⋮
```

(2) 统计每个英雄名称的长度，并将相同长度的英雄名称归为一组，每组之间用中文的顿号（、）分隔。你可以将结果保存到一个两列的元胞数组 C 中，第一列是英雄名称的长度，第二列是对应长度的英雄名称列表。C 中应有 4 行，结果供大家参考：

C = 4×2 cell

	1	2
1	1	"暃、澜、镜、曜、瑶、铠"
2	2	"海诺、亚连、海月、戈娅、桑启、金蝉、云缨、艾琳、蒙恬、蒙犽、西施、马超"
3	3	"朵莉亚、姬小满、莱西奥、赵怀真、司空震、夏洛特、阿古朵、云中君、猪八戒"
4	4	"鲁班大师、上官婉儿、百里玄策、百里守约、干将莫邪、东皇太一、太乙真人、"

挑战篇

Q1: 密码安全性评估与得分系统设计

为了保障账户的安全，我们需要设计一个密码验证系统来确保用户设置的密码符合特定的规则。一个合格的密码必须满足以下两点规则：

规则一：密码长度不低于 8 位，最多 16 位；

规则二：密码中只能包含英文字母（大小写的英文字母都可以）、数字或者以下标点符号：.,!,:?#%&:<>+-\*/，不能包含其他字符。

请解决以下两个问题：

- (1) 任意给定一个字符串标量，代表用户设置的密码，请判断该密码是否符合上述两点规则，你的输出结果应为逻辑值 1 或逻辑值 0。
- (2) 假设用户设置的密码符合上述两点规则，现在要对该密码的强度进行评分，评分标准和细则如下表所示：

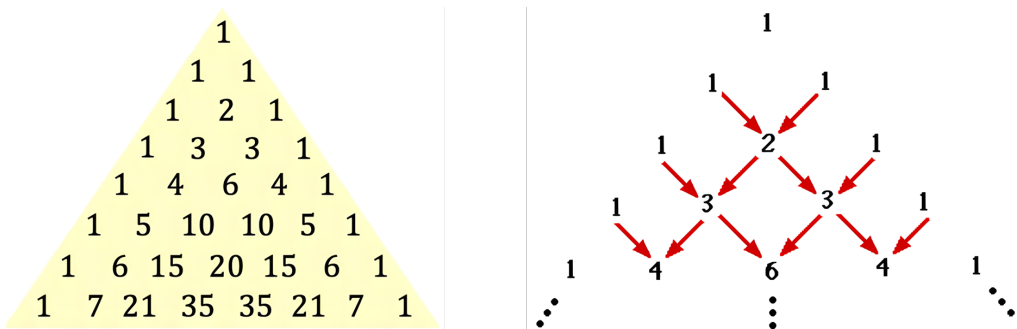
标准	细则	得分
密码长度	8 个字符	5
	9 到 12 个字符	10
	大于等于 13 个字符	25
字母的使用情况	没有字母	0
	只有小写字母或只有大写字母	10
	既有大写又有小写字母	20
数字的使用	没有数字	0
	1 个数字	10
	大于 1 个数字	20
标点符号的使用	没有标点符号	0
	1 个标点符号	10
	大于 1 个标点符号	25
额外奖励	既有字母又有数字，但是没有标点符号	2
	数字、标点符号和字母都有，但字母全为大写或全为小写	5
	大写字母、小写字母、数字和标点符号都有	10

请根据上表计算用户设置的密码的得分。

例如密码 "abcd123456" 的得分为  $10+10+20+0+2 = 42$ ；密码 "Abc528963.." 的得分为  $10+20+20+25+10 = 85$ 。

Q2: 创建并显示杨辉三角

杨辉三角是组合数学中的经典结构，它与二项式系数紧密相关。其独特的形状和规律性的数字排列使其在数学、物理、计算机科学和其他领域中都有广泛的应用。



(1) 生成一个  $n$  行杨辉三角矩阵

接下来，你需要根据杨辉三角的规则填充杨辉三角矩阵 `yh`。规则如下：

- 例如：当 n 等于 9 时，对应的 yh 矩阵为：

```

yh = 9x9
1 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 0
1 2 1 0 0 0 0 0 0
1 3 3 1 0 0 0 0 0
1 4 6 4 1 0 0 0 0
1 5 10 10 5 1 0 0 0
1 6 15 20 15 6 1 0 0
1 7 21 35 35 21 7 1 0
1 8 28 56 70 56 28 8 1

```

## （2）格式化显示杨辉三角

例如下面是一个 9 行 38 列的字符数组 cc，它表示一个 9 行的杨辉三角：

```
cc = 9x38 char 数组
```

### Q3: 汉字拼音转换器 (本题数据为 h 3 3)

在中文文本处理中，将汉字转换为对应的拼音是一项常见的任务，尤其是在处理语音合成、文本到语音转换等应用中。请按照本小节开头的提示导入好作业的数据，本题用到的数据为 `h33`，它是一个 7894 行 2 列的字符串数组，例如它的前 12 行如下所示：

12×2 **string** 数组

"伉"	"zhòu"
"伲"	"yì"
"叻"	"fǔ"
"𪔐"	"hǎn"
"𪔑"	"duō"
"𪔒"	"yāo"
"𪔓"	"xié"
"𪔔"	"jié"
"𪔕"	"tǒng"
"𪔖"	"sī"
"𪔗"	"chù   cù   zhòu"
"𪔘"	"sǒng"



### Q5: 从网页源码中提取成语（本题数据为 h\_3\_5）

请按照本小节开头的提示导入好作业的数据，本题用到的数据为 h\_3\_5，它是一个 20 行 1 列的字符向量元胞数组，这 20 个字符向量中保存着 20 个网页的源代码。这些网页源代码来自于某个介绍成语故事的网站，每个网页中都有 100 个成语。

例如 h\_3\_5 中保存的第一个网页源代码（即 h\_3\_5{1}）的开头如下所示：

```

<!DOCTYPE HTML>
<html>
<head>
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge,chrome=1" />
<meta http-equiv="Cache-Control" content="no-siteapp">
<meta name="viewport" content="width=device-width,initial-scale=1.0,minimum-scale=1.0,maximum-scale=1.0" />
<meta name="applicable-device" content="pc,mobile" />
<link rel="dns-prefetch" href="f.bmcx.com" />
<link rel="canonical" href="https://chengyu.bmcx.com/e1zdh_1_chengyulist/" />
<link rel="apple-touch-icon-precomposed" sizes="57x57" href="//f.bmcx.com/file/chengyu/i_c_o_57x57.png" />
<link rel="apple-touch-icon-precomposed" sizes="72x72" href="//f.bmcx.com/file/chengyu/i_c_o_72x72.png" />
<link rel="apple-touch-icon-precomposed" sizes="114x114" href="//f.bmcx.com/file/chengyu/i_c_o_114x114.png" />
<meta name="format-detection" content="telephone=no" />

```

这个网页的源代码非常长，大家可以先将这个字符向量中的内容复制到记事本中，然后往下翻，在下方区域中可以看到我们要提取的成语：

```

<div align="center"><span class="all_an2_0">成语大全</span><span class="all_an2_1"><a href="https://chengyujielong.bmcx.com/">成语接龙
</a></span><span class="all_an2_1"><a href="https://cytk.bmcx.com/">成语填空</a></span></div>
<ul class="list">
<li><a href="/qiancicuoyi_qng_chengyuchaxun/" target="_blank">遣词措意</a></li><li><a href="/qianciliyi_chengyuchaxun/" target="_blank">遣词立意
</a></li><li><a href="/qiancizaoyi_chengyuchaxun/" target="_blank">遣词造意</a></li><li><a href="/qiancicuoyi_chengyuchaxun/" target="_blank">遣辞
措意</a></li><li><a href="/qianjiangdiaobing_chengyuchaxun/" target="_blank">遣将调兵</a></li><li><a href="/qianjiangzhengbing_chengyuchaxun/"
target="_blank">遣将征兵</a></li><li><a href="/qianxingtaoqing_chengyuchaxun/" target="_blank">遣兴陶情</a></li><li><a href="/qiaozuerdai_s34_chengyuchaxun/" target="_blank">趑趄而待</a></li><li><a href="/qianlijieyan_chengyuchaxun/" target="_blank">千里结言
</a></li><li><a href="/qianlijiechou_chengyuchaxun/" target="_blank">千里借筹</a></li><li><a href="/qianlijungu_chengyuchaxun/" target="_blank">千
里骏骨</a></li><li><a href="/qianli_liang_shiyoujise_chengyuchaxun/" target="_blank">千里饒粮，士有饥色</a></li><li><a href="/qianlimingjia_chengyuchaxun/" target="_blank">千里命驾</a></li><li><a href="/qianlishenjiao_chengyuchaxun/" target="_blank">千里神交
</a></li><li><a href="/qianlisongemao_chengyuchaxun/" target="_blank">千里送鹅毛</a></li><li><a href="/qianlitiaotiao_chengyuchaxun/"
target="_blank">千里迢迢</a></li><li><a href="/qianlitiaoyao_chengyuchaxun/" target="_blank">千里迢遥</a></li><li><a href="/qianlitongfeng_chengyuchaxun/" target="_blank">千里同风</a></li><li><a href="/qianliwuyan_chengyuchaxun/" target="_blank">千里无烟
</a></li><li><a href="/qianliyiqu_chengyuchaxun/" target="_blank">千里一曲</a></li><li><a href="/qianliyixi_chengyuchaxun/" target="_blank">千里移檄
</a></li><li><a href="/quzhiruwu_chengyuchaxun/" target="_blank">趋之如鹜</a></li><li><a href="/quzhiruwu_chengyuchaxun/" target="_blank">趋之

```

请从这 20 个网页的源码中提取出所有的成语，并将结果保存到一个长度为 2000 的字符串向量 cy 中。

### Q6: 成语世界探秘和接龙游戏（本题数据为 h\_3\_6）

请按照本小节开头的提示导入好作业的数据，本题用到的数据为 h\_3\_6，它是一个 30767 行 3 列的字符串数组，里面保存着某网站提供的 30767 个成语和拼音：

```

h_3_6 = 30767x3 string
    "一丁不识"    "yī dīng bù shí"    "yī dīng bu shi"
    "一丁点儿"    "yī dīng diǎn ér"    "yī dīng dian er"
    "一不作，..."    "yī bù zuò , èr bù xiū"    "yī bu zuo , er bu xiu"
    "一不做，..."    "yī bǔ zuò , èr bù xiū"    "yī bu zuo , er bu xiu"
    "一不压众..."    "yī bù yā zhòng , bǎi bù su..."    "yī bu ya zhong , bai bu su..."
    "一不扭众"    "yī bù niǔ zhòng"    "yī bu niu zhong"
    "一世之雄"    "yī shì zhī xióng"    "yī shi zhi xiong"
    "一丘一壑"    "yī qiū yī hè"    "yī qiu yi he"
    "一丘之貉"    "yī qiū zhī hé"    "yī qiu zhi he"
    "一丝一毫"    "yī sī yī háo"    "yī si yi hao"
    :
    :
    :

```



根据上面的数据完成以下四个任务（各任务是独立的，没有先后顺序）：

### （1）查找指定结构形式的成语

在博大精深的中文成语世界里，不同的字排列组合形成了丰富多样的表达形式。其中，有一些成语的结构形式特别规整，如 AABB 式、AABC 式和 ABAB 式等，它们以其独特的形式展现了成语的韵律之美。

以 AABB 式为例，这种形式的成语要求第一个字与第二个字相同，第三个字与第四个字相同，同时第一个字与第三个字不同。例如，“世世代代”、“严严实实”和“家家户户”都是典型的 AABB 式成语。

现在，请你从给定的数据中，探寻这些具有特定结构形式的成语。你的任务是分别提取出满足 AABB 式、AABC 式和 ABAB 式的四字成语，并将它们分类保存到三个不同的字符串向量中。

### （2）提取生肖成语

中国的传统文化中，十二生肖（鼠、牛、虎、兔、龙、蛇、马、羊、猴、鸡、狗、猪）是非常重要的部分，每个生肖都有其独特的象征意义和特点。这些生肖也经常出现在我们的成语中，丰富了成语的表达方式和文化内涵。

请从给定的数据中，提取出包含十二生肖的成语（不需要考虑生肖的别称，例如狗的成语中不需要包括“犬牙交错”），并将结果保存到一个 12 行 1 列的字符向量元胞数组中，元胞数组中每个位置的数据保存一个生肖的成语，同一生肖内不同的成语之间使用换行符隔开。

### （3）探索高频汉字与拼音

在成语中，某些汉字和拼音的出现频率高于其他。这些高频的汉字和拼音往往揭示了我们在日常使用成语时的偏好和习惯。

请你对给定的成语数据进行深入探索：首先，提取所有成语中的汉字，并找出出现频率最高的前 10 个汉字。将这 10 个汉字合并为一个字符串标量，中间用中文顿号“、”隔开；其次，提取所有成语的无声调拼音（h\_3\_6 的第三列），并找出出现频率最高的前 10 个拼音。同样地，将这 10 个拼音合并为一个字符串标量，中间用中文顿号“、”隔开。

参考答案：“不、之、一、无、人、心、天、风、大、如”以及“yi、bu、zhi、shi、wu、ji、yu、qi、li、yan”。

### （4）成语接龙游戏

在本任务中，你需要设计一个成语接龙游戏，具体的游戏规则如下：

**开始游戏：**游戏开始时，你可以任意输入一个成语，或输入“提示”来获取随机的成语。

**接龙规则：**每个成语的最后一个字必须是下一个成语的第一个字。

**提示功能：**在你的回合中，你可以输入“提示”来获取可以接的成语建议（如果可以接的成语较多的话，只需要给出最多五个成语进行提示）。

**退出游戏：**你可以输入“退出”来结束游戏。

**游戏结束条件：**如果你或者电脑找不到可以接的成语时，游戏结束。

**游戏结束后：**输出这一轮游戏中出现的所有成语。

**注意事项：**游戏中出现的所有成语都必须包含在 h\_3\_6 中，且已经使用过的成语不能再次使用。如果用户输入了一个不在 h\_3\_6 中的成语，或者输入了一个已经被使用过的成语，程序中应给出相应的提示。

下面是供大家参考的两次游戏过程，程序中通过 input 函数获得用户的输入内容：

请输入一个成语开始游戏(退出游戏请输入退出、如需提示请输入提示): 三心二意  
你的成语是三心二意, 我接的是: 意合情投

-----分割线-----

你需要输入以“投”开头的成语

请输入你的答案(退出游戏请输入退出、如需提示请输入提示): 提示

提示: 你可以接的成语有: 投井下石、投其所好、投卵击石、投怀送抱、投机倒把

-----分割线-----

你需要输入以“投”开头的成语

请输入你的答案(退出游戏请输入退出、如需提示请输入提示): 投怀送抱

你的成语是投怀送抱, 我接的是: 抱瓮灌园

-----分割线-----

你没有可以接的成语了, 游戏结束!

-----分割线-----

本局游戏的成语如下:

"三心二意"

"意合情投"

"投怀送抱"

"抱瓮灌园"

请输入一个成语开始游戏(退出游戏请输入退出、如需提示请输入提示): 高高兴兴  
你的成语是高高兴兴, 我接的是: 兴妖作怪

-----分割线-----

你需要输入以“怪”开头的成语

请输入你的答案(退出游戏请输入退出、如需提示请输入提示): 怪我太傻

你输入的可能不是成语, 请重新输入!

-----分割线-----

你需要输入以“怪”开头的成语

请输入你的答案(退出游戏请输入退出、如需提示请输入提示): 提示

提示: 你可以接的成语有: 怪事咄咄、怪声怪气、怪形怪状、怪模怪样、怪腔怪调

-----分割线-----

你需要输入以“怪”开头的成语

请输入你的答案(退出游戏请输入退出、如需提示请输入提示): 怪声怪气

你的成语是怪声怪气, 我接的是: 气喘吁吁

-----分割线-----

你没有可以接的成语了, 游戏结束!

-----分割线-----

本局游戏的成语如下:

"高高兴兴"

"兴妖作怪"

"怪声怪气"

"气喘吁吁"

## Q7: 整理王者荣耀英雄数据（本题数据为 h\_3\_7）

请按照本小节开头的提示导入好作业的数据，本题用到的数据为 h\_3\_7，它是一个长度为 21458 的字符向量，其中保存了王者荣耀中 117 名英雄角色的相关数据，数据开头如下：

```
{
    "ename": 105,
    "cname": "廉颇",
    "id_name": "lianpo",
    "title": "正义爆轰",
    "pay_type": 10,
    "new_type": 0,
    "hero_type": 3,
    "skin_name": "正义爆轰|地狱岩魂",
    "moss_id": 3627
}, {
    "ename": 106,
    "cname": "小乔",
    "id_name": "xiaoqiao",
    "title": "恋之微风",
    "new_type": 0,
    "hero_type": 2,
    "moss_id": 3644
}, {
    "ename": 107,
    "cname": "赵云",
    "id_name": "zhaoyun",
    "title": "苍天翔龙",
    "new_type": 0,
    "hero_type": 1,
    "hero_type2": 4,
    "skin_name": "苍天翔龙|忍●炎影|未来纪元|皇家上将|嘻哈天王|白执事|引擎之心",
    "moss_id": 3661
}, {
```

你的任务是提取每位英雄的 cname（英雄名称）、title（英雄雅称）和 skin\_name（皮肤名称）。需要注意的是，某些英雄可能没有皮肤数据，对于这些英雄，请使用空字符串来表示他们的皮肤名称。

最终，你需要将结果保存到一个 117 行 3 列的字符串数组 D 中，其中每一行对应一个英雄，三列分别对应英雄的名称、雅称和皮肤名称。

参考答案如下：

```
D = 117×3 string
    "廉颇"      "正义爆轰"      "正义爆轰|地狱岩魂"
    "小乔"      "恋之微风"      ""
    "赵云"      "苍天翔龙"      "苍天翔龙|忍●炎影|未来纪元|皇家上将|嘻哈天王|白执事|引擎之心"
    "墨子"      "和平守望"      ""
    "妲己"      "魅力之狐"      "魅惑之狐|女仆咖啡|魅力维加斯|仙境爱丽..."
    "嬴政"      "王者独尊"      "王者独尊|摇滚巨星|暗夜贵公子|优雅恋人|..."
    "孙尚香"      "千金重弩"      "千金重弩|火炮千金|水果甜心|蔷薇恋人|杀..."
    "鲁班七号"      "机关造物"      "机关造物|木偶奇遇记|福禄兄弟|电玩小子|..."
    "庄周"      "逍遥梦幻"      "逍遥梦幻|鲤鱼之梦|蜃楼王|云端筑梦师"
    "刘禅"      "暴走机关"      "暴走机关|英喵野望|绅士熊猫|天才门将"
    ...
```

## Q8: 整理王者荣耀装备数据（本题数据为 h\_3\_8）

请按照本小节开头的提示导入好作业的数据，本题用到的数据为 h\_3\_8，它是一个长度为 31313 的字符向量，其中保存了王者荣耀中 136 件装备的相关数据，部分数据如下所示：

```
{, {
    "item_id": 1121,
    "item_name": "风暴巨剑",
    "item_type": 1,
    "price": 546,
    "total_price": 910,
    "des1": "<p>+80物理攻击</p> "
}, {
    "item_id": 1122,
    "item_name": "日冕",
    "item_type": 1,
    "price": 426,
    "total_price": 710,
    "des1": "<p>+20物理攻击<br>+5%冷却缩减<br>+400最大生命</p>",
    "des2": "<p>唯一被动-残废：技能对首个命中的敌方英雄造成10%减速，持续3秒，该效果有8秒冷却时间</p>"
}, {
    "item_id": 1123,
    "item_name": "狂暴双刃",
    "item_type": 1,
    "price": 534,
    "total_price": 890,
    "des1": "<p>+15%攻击速度<br>+10%暴击率<br>+5%移速</p>"
}, {
```

你需要从数据中提取以下信息：装备的名称 (item\_name)、装备的价格 (price)、装备的总价(total\_price)以及装备的描述 (des1，如果存在 des2 也需要一并提取)。

注意：装备描述中可能包含 HTML 标签（如 <p> 和 <br>），在提取描述时，请确保去除这些标签，只保留纯文本信息。另外，装备名称和描述信息请使用字符串向量保存，价格和总价请使用数值向量保存。

参考答案如下：

price = 136×1	total_price = 136×1	item_name = 136×1 string
150	250	"铁剑"
174	290	"匕首"
192	320	"搏击拳套"
174	290	"吸血之镰"
270	450	"雷鸣刃"
330	550	"冲能拳套"
546	910	"风暴巨剑"
426	710	"日冕"
534	890	"狂暴双刃"
558	930	"陨星"
⋮	⋮	⋮
des = 136×1 string		
"+20物理攻击"		
"+10%攻击速度"		
"+8%暴击率"		
"+8%物理吸血"		
"+40物理攻击"		
"+15%暴击率"		
"+80物理攻击"		
"+20物理攻击+5%冷却缩减+400最大生命+唯一被动-残废：技能对首个命中的敌方英...		
"+15%攻击速度+10%暴击率+5%移速"		
"+55物理攻击+唯一被动-切割：+60物理穿透"		
⋮		