

[转]写给程序员的 Unicode 入门介绍

faremax (/u/88ca7ceed50e) [+ 关注](#)

2018.09.02 16:46* 字数 5933 阅读 0 评论 0 喜欢 0

(/u/88ca7ceed50e)



本文转自 微信公众号 jobbole

程序员世界对这个名字发自内心的恐惧和敬畏。我们都知道在我们的软件中应该“支持 Unicode”（无论是什么意思——对所有的字符串使用 `wchar_t`，是吗？）。但 Unicode 很深奥，它有千页的 Unicode 标准，还有几十页的补充附录、报告和注解，简直太吓人了。即使 Unicode 诞生 30 多年后，程序员们还觉得它很神秘。

几个月前，我开始对 Unicode 着迷，决定花些时间仔细了解一番。在本文，我来从程序员的视角对其做介绍。

我主要关注字符集，与字符串处理和 Unicode 文本相关的东西。因此，这里我不会过细地聊字体、文本布局、形状、渲染，或本地化，那些是另外的议题，超出了我的能力（知识）范围。

多样性和内在复杂性

当你开始学习 Unicode，有一件事情很明显，就是它和你熟悉的字符集（比如 ASCII）相比，Unicode 复杂性要高了一大截。这不仅仅是指 Unicode 包含了很多的字符，虽然这是一个方面。Unicode 还有很多内部结构，特性和特殊情况，使其不只是人们所认为的纯粹的“字符集”。本文后续会介绍一些相关内容。

当面对所有的复杂性时，尤其是作为工程师，很难不问自己，“为什么我们需要这么多？真的有必要吗？可以简化吗？”

然而，Unicode 的目标是准确地表示全世界的书写系统（writing systems）。Unicode 协会的目标是“让全世界的人们不论什么语言都可以使用电脑”，所以你可想见，书面语言的多样性是巨大的！迄今为止，Unicode 支持 135 种不同的书写系统，包含约 1100 种语言，但目前还有超过 100 种书写系统没有支持，包括现代的和已成为历史的，Unicode 协会还在努力将其加进来。

鉴于分支的多样性，要表示它们必然是一个复杂的项目。Unicode 接受了它的多样，接受了任务（包含所有人类的书写系统）中的内在复杂性，它没有在名字简化上做太多取舍，但是它对需要完善任务的地方的规则，做了异常处理。

此外，Unicode 承诺不仅支持单一语言的文本，还支持多种语言共存于一个文本中——引进了更多的复杂性。

大多数编程语言都有处理底层文本操作的库，但是作为程序员，你仍然需要知道一些 Unicode 特性，知道何时怎样去应用它。要了解这些东西可能得花些时间动动脑筋，但别灰心——想想有数以亿计的人，如果你的软件支持他们的语言，那他们也可以使用你的软件啦。所以，拥抱复杂吧！

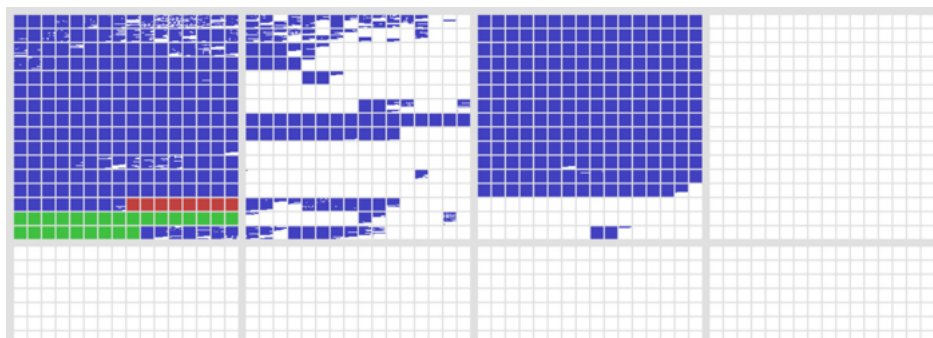
Unicode 编码空间

我们先从几个大的方向入手。Unicode 的基本元素——它的“字符”，虽然这种叫法不是太贴切——被称作编码点（Code Point）。编码点通过数字来区分，通常写成 16 进制的形式再加前缀“U+”，例如 U+0041 表示拉丁字母“A”、U+03B8 表示希腊字母“θ”。每个编码点都有一个简称，还有一些其他属性，Unicode 字符数据库对此有详细说明。

所有编码点组成的集合被称作编码空间（Code Space）。Unicode 编码空间包含 1,114,112 个编码点。然而，其中只有 128,237 个编码点——编码空间的 12% 被赋值，目前。还有很多空间用来增长！Unicode 还保留了另外 137,468 字符作为“自用”空间，这些字符没有标准的含义，可以被个人应用所使用。

空间分配

为了对编码空间的布局有个了解，把它可视化会比较直观。下面是整个编码空间的布局，一个像素代表一个编码点。使用小方块来表示以保证视觉的一致性；每个小方块是 $16 \times 16 = 256$ 个编码点，每个大方块是一个面有 65536 个编码点。总共加起来有 17 个面板。



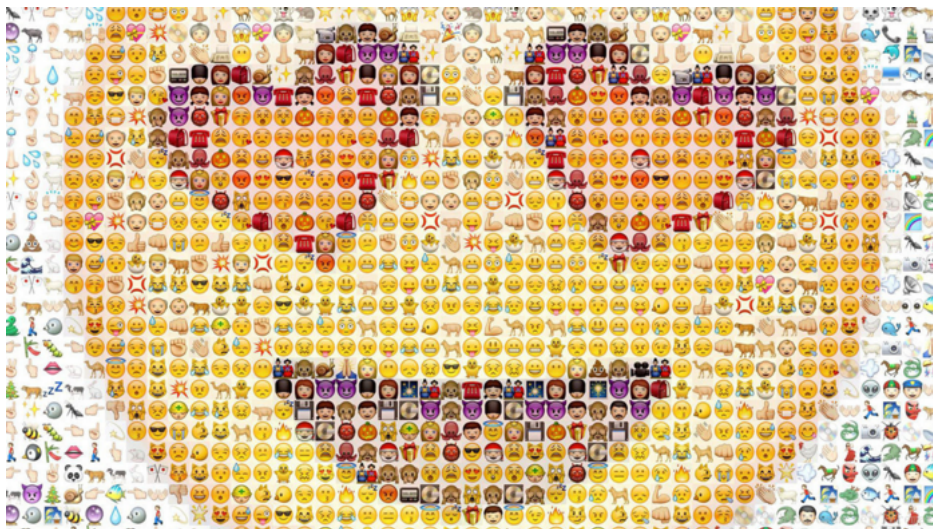
- 白色表示未用空间
- 蓝色表示已用空间
- 绿色表示自用区域
- 小的红色区域是代理区（surrogates，后面会讲）

如你所见，被使用的区域分布有点稀疏，但都集中在前三个面里。

0 号面板也被称作“基本多语言面板（Basic Multilingual Plane，简称 BMP）”。BMP 包含现代文本所需的基本所有字符，包括拉丁文、斯拉夫文、希腊文、汉字（中国），日文、朝鲜文、阿拉伯文、希伯来文、梵文（印度）等等。

（过去，编码空间只有 BMP 而已——Unicode 最初设想是一个 16 Bit 的编码，只包含 65536 个字符。在 1996 年扩充到现在的规模。然而，绝大多数现代字符属于 BMP。）

1 号面板包含历史上的文字，比如苏美尔楔形文字和埃及象形文字，还有 emoji 和其他各种符号。2 号面板包含一大块不常用的和历史上的汉字字符。剩下的面是空的，除了 14 号面板中有一小部分被用作格式化字符；15-16 号面板全部保留自用。



书写系统

让我们放大前三个面板，因为这是最重要的部分：

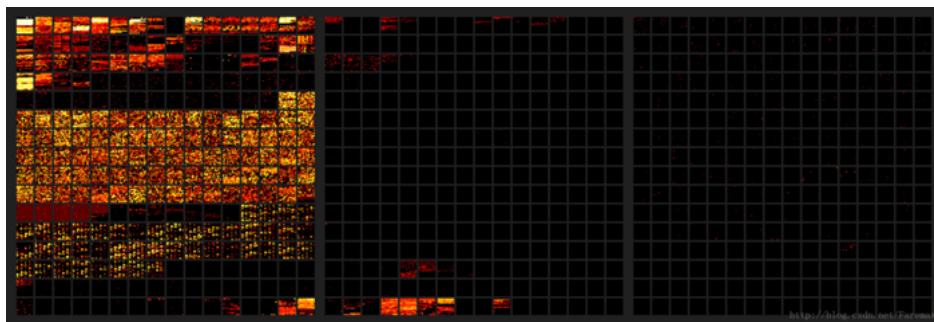


这张图用颜色表示了 Unicode 中135 种不同的书写系统。你可以看到汉字（蓝色）和朝鲜语（棕色）占了 BMP 很大一部分（右边的大方块）。与之相对，此图中所有的欧洲，中东，南亚语言加起来刚好占了 BMP 的第一行。

编码空间的很多区域都和更早的编码兼容或相同。例如，Unicode 的前 128 个字符就是 ASCII 的拷贝。显然是对兼容性很有好处——很容易无损的从小编码转向 Unicode（反过来也一样，只要没有使用小编码之外的字符）。

使用频率

可视化编码空间还有一个有趣的方法，就是看使用频率的分布——换句话说，就是每个编码点在真实世界中使用的频率。0-2 号面的热力图是基于来自维基百科和 推特（所有语言）的大量文本所得。频率增长的方向是黑（没出现）、红、黄、白。



你可以看到，绝大多数书样本文本都分布在 BMP 中，有些零散的使用来自1-2 号面。最大的异常是 emoji，它点亮了 1 号面最底下那的几个小方块。

编码

我们知道 Unicode 编码点，通过它们在编码空间中的下标来定义，范围从 U+0000 到 U+10FFFF。但是在内存或文件中编码点如何用字节表示呢？

对计算机友好的最省事方式是用 32 位整数来存储编码点下标。这样做是可行，但是每个字符用 4 个字节有点浪费。当你处理大量文本的时候，使用 32 位整数存储 Unicode 会占用大量额外存储、内存、带宽等。

于是，Unicode 有了几个紧凑的编码。32 位整数编码被称作 UTF-32（UTF=“Unicode Transformation Format”），但是很少被用来存储。顶多作为临时内部表示出现，用来检查或操作字符串中的编码点。

最常见的是，你会看到 Unicode 文本被编码为 UTF-8 或 UTF-16。这些都是可变长度编码，分别由 8-bit 或 16-bit 为一个单元组成。这些方案中，下标值较小的编码点占用的字节数也少，会节省不少内存。这样做的代价是处理 UTF-8/16 需要以编程的方式来处理，会慢一些。

UTF-8

在 UTF-8 中，每个编码点依据下标值，被存储为 1 到 4 个字节。

UTF-8 使用二进制前缀系统，在此系统中每个字符的最高位的几个比特表明它是否是单个字节，多字节序列的开始，或中间字节；剩余的比特连接起来表示编码点的下标。下面的表格展示了 UTF-8 是如何编码的：

UTF-8 (二进制)	编码点 (二进制)	范围
0xxxxxxx	xxxxxxx	U+0000–U+007F
110xxxx 10yyyyyy	xxxxxyyyyyy	U+0080–U+07FF
1110xxxx 10yyyyyy 10zzzzzz	xxxxyyyyyyzzzzzz	U+0800–U+FFFF
11110xxx 10yyyyyy 10zzzzzz 10wwwwww	xxxyyyyyyyzzzzzzwwwwww	U+10000–U+10FFFF

UTF-8 有一个方便的属性，即最开始128 个字符（ASCII 字符）被编码为单个字节，所有的非 ASCII 字符被编码为 128-255。这产生了两个好处。首先，任何已经是 ASCII 编码的字符串和文件无需转换就可以被 UTF-8 识别。其次，大量的广泛使用的编程惯例——比如 NULL 结尾，分隔符（n,t,',',）等——在 UTF-8 中也是可用的。ASCII 字节不会出现在非 ASCII 编码点中，所以搜索以 NULL 结尾或分隔符结尾的字符串是可以的。

多亏了这个便利，使扩展遗留 ASCII 程序和 API 来处理 UTF-8 字符变得简单。UTF-8 被广泛运用在 Unix、Linux 和网络世界中，还有许多程序员主张 UTF-8 应该作为任何地方的默认编码。

然而，UTF-8 还不能全面替代 ASCII。例如，遍历字符串中的“字符”的代码需要解码 UTF-8 并遍历编码点（或字位簇（grapheme cluster）——后面会讲到），而不是字节。当你测量字符串“长度”时，你得考虑是要字节长度，还是编码点长度，还是文本渲染的宽度为单位的长度还是其它长度。

UTF-16

你可能遇到的另一个编码是 UTF-16。它使用 16-bit 字，每个字符被存储为 1 个或 2 个字。

和 UTF-8 一样，我们可以用二进制前缀的形式表示 UTF-16 的编码规则：

UTF-16 (二进制)	编码点 (二进制)	范围
xxxxxxxxxxxxxxxx	xxxxxxxxxxxxxxxx	U+0000–U+FFFF
110110xxxxxxxx 110111yyyyyyyy	xxxxxxxxxyyyyyyyyy	0x10000U+10000–U+10FFFF

但是，通常人们谈到 UTF-16 是因为它涉及到了一个在编码点术语中被称作“代理（surrogate）”的东西。所有在范围 U+D800-U+DFFF（或在其他范围）中的编码点，这些和上表中二进制前缀 110110 和 110111 匹配的编码点——是 UTF-16 中的保留区域，它们自身不表示任何有效的字符。它们仅用于上面 2 个字的编码模式中，被称作“代理对（surrogate pair）”，代理编码点在任何其他情况下都是非法的！它们不能出现在 UTF-8 和 UTF-32 中。

在过去，UTF-16 是 1996 年之前的 Unicode 版本的派生物，那时只有 65536 个编码点。初衷是不应有不同的编码，Unicode 应该是简单的 16-bit 字符集。后来，编码空间被扩充用来表示不常用的（仍然重要）的汉字字符，这是 Unicode 设计者之前没计划的。代理区在那时被引进，直说了吧，作为拼凑，允许 16-bit 编码访问新的编码点。

如今，Javascript 使用 UTF-16 作为其标准的字符串表示：如果你问一个字符串的长度，或遍历它等，结果都以 16-bit 的字为单位，同时任何 BMP 之外的编码点都用代理对表示。UTF-16 也被微软 WIN32 API 使用；尽管 Win32 同时支持 8-bit 和 16-bit 字

字符串，但是 8-bit 版本仍然莫名其妙地不支持 UTF-8——只支持使用旧编码的代码，像 ANSI。这使得 UTF-16 成为在 Windows 上获得 Unicode 支持的唯一方法。

顺便说一下，UTF-16 字符可以大端存储，也可以小端存储。Unicode 在这个问题上没有说明，虽然它确实鼓励一个惯例，即把 U+FEFF 零宽无间断间隔这个字符放到 UTF-16 文件开头作为字节序标识，来消除字节序问题。（如果文件和系统的字节序不同，BOM（ByteOrderMark）会被解码为 U-FFFE，这不是一个有效的编码点。）

组合标记

目前为止，我们一直在讨论编码点。但是 Unicode 中，字符比单独的编码点更复杂！

Unicode 包含一个系统，可以合并多个编码点，动态组合字符。此系统用各种方式增加灵活性，而不引起编码点的巨大组合膨胀。

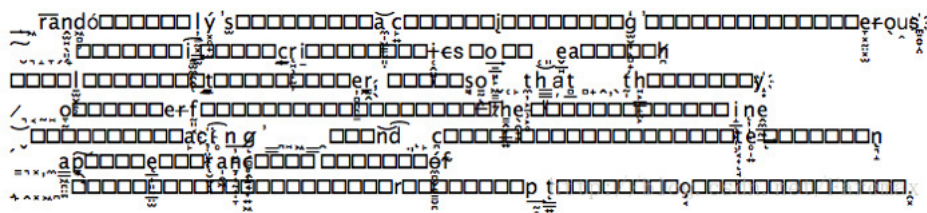
例如，在欧洲语言中，组合标记出现在变音符和字母的使用中。Unicode 支持各种各样的变音符，包括尖音符的和重音符、元音变音符、变音符等等。所有这些变音符可以被使用在任何字母表的字母中。事实上，多个变音符可以被使用在一个字母上。

如果 Unicode 试图为每个字母组合或变音符组合分配一个独立的编码点，事情会变得无法控制。相反，动态组合系统可以让你构造你想要的任何字符，通过以一个基础编码点（字母）开始然后附加额外的编码点，被称作“组合标识”，来指定变音符。当一个文字渲染器看到字符串中有这样的序列时，它会自动堆叠变音符到基础字母的上面或下面来造出一个组合字符。

例如，带重音的字符“Á”会被表示成由两个编码点组成的字符串：U+0041 “A” 拉丁大写字母 a 加上 U+0301 “◌̇” 组合尖音符。这个字符串自动被渲染成单个字符：“Á”。

如今，Unicode 还包含许多“预设的”编码点，每个表示一个被使用过的组合，例如 U+00C1 “Á” 带锐音符的拉丁大写字母 A 或 U+1EC7 “ê” 带扬抑符和下点的小写拉丁字母 e。我怀疑这些大多继承自融入 Unicode 的旧编码，来保证兼容性。实际上，对于欧洲语言中的大多数常见的带变音符的字母都有预设，所以文本中动态组合用的不多。

可是，组合标志系统确实允许任意数量的变音符被叠加到任何基础字符上。使用归谬法的 Zalgo 文本，它通过随机叠加任意数量的变音符在每个字母上，让它溢出行距，产生混乱现象。（如下图）



Unicode 中出现动态组合字符的其他地区：

- 天城体（梵文），这种文字被用在印度北部，梵文和其他南亚语言中，用组合标记标识特定元音的附加到辅音字母上。例如，“ह” + “ि” = “हि”（“h” + “i” = “hi”）
- 表示音节的朝鲜字符，但是它被称作 Jamo，用来表示音节中的元音和辅音。当然也有为朝鲜文预制的编码点，同时也可以动态组合它们的 jamo。例如，“ㅎ” + “ㅏ” + “ㄴ” = “한”（“h” + “a” + “n” = “han”）
- 阿拉伯文和希伯来文中的元音标记。这些语言中，单词通常由元音拼写。它们有

变音符号标记元音（用在字典，语言教学材料，儿童教材，等地方）。这些变音符号用组合标记表示。

| 说明 | 样例 |

| --- | --- |

| 希伯来文(带变音符号) | אֶת דְּלִתִּי הִזִּיז הַנִּיעַ, קָטַב לַשְּׂכֵנִי יִשׂוּד |

| 正常文本(不带变音符号) | את דלתי הוזיז הניע, קטב לשכתי ישוד |

规范等价性

Unicode 中，预设字符和动态组合系统并存。后果就是有多种方法表示同一个字符串——不同编码点序列产生相同用户可感知的字符。例如，我们之前看到的，表示字符“Á”，我们可以用一个编码点 U+00C1，也可以用两个编码点 U+0041 和 U+0301。

另一个歧义来源是一个字符中的多个变音符号。当两个变音符号作用在同意个基本字符上面时，变音符号的顺序很重要，例如，都在上面：“ā”（点然后长音符）和 “â”（长音符然后点）是不一样的。然而，当音节运用在不同边时，例如，一个在上边一个在下边，编码点的顺序不会影响渲染。此外，一个有多个音节的字符，它可能会由一个预制的编码点再加其余的编码点来表示。

例如，越南字母“ê”可以用以下五种方式表示：

- 完全预设：U+1EC7 “ê”
- 部分预设：U+1EB9 “e” + U+0302 “◌̂”
- 部分预设：U+00EA “ê” + U+0323 “◌̣”
- 完全分解：U+0065 “e” + U+0323 “◌̣” + U+0302 “◌̂”
- 完全分解：U+0065 “e” + U+0302 “◌̂” + U+0323 “◌̣”

Unicode 把这样的字符串集合称作“规范等价”字符。在搜索、排序、渲染、文本选择等操作中，规范等价字符应该被同等对待。这影响到了你如何实现文本的操作。例如，假设你的程序有“查找”操作，用户搜索“ê”，理论上应当找到如上所有出现的所有版本的“ê”！

形式正规化

要解决如何处理等值字符串的问题，Unicode 定义了几种正规形式：是几种把字符串转化成规范形式的方法，这样它们就可以被逐点比较（或按字节比较）。

“NFD”正规化方法，完全分解每个字符到基本部件和组合标记，去掉字符串中任何预制的编码点。还会按渲染位置排列每个组合标记，举个例子，在字母底下的变音符号要比在上边的靠前。（不会重排有相同渲染位置的变音符号，因为它们的位置关系是可视的，前面提到过。）

“NFC”正规化方法，反过来，尽可能的把编码点替换成预制编码点。如果使用了不常用的变音符号组合，可能不会有任何预制的编码点，这种情况下 NFC 仍然替换它可以替换的，然后留下组合标志（和 NFD 一样，还是会按渲染顺序重新排序）。

还有一些方法被称作 NFKD 和 NFKC。这里的“K”指的是兼容性分解，它包含了某种程度上“相似”但是视觉上不同的字符。但我不打算讲这些。

字位簇

如上所见，Unicode 包含多种情况，用户认为的一个“字符”事实上底下可能由多个编码点组成。Unicode 使用「字位簇」的概念来表示这种情况。一个由一个或多个编码点组成的字符串构成一个“用户感知的字符”。

UAX #29 为字位簇定义了精确的规则。它大约是“一个基本的编码点接着任意数量的组合标记”，但是真实的定义有点复杂；它包含了朝鲜语字母，和 emoji ZWJ 序列。

字位簇主要被用在文本编辑：它们对光标和文本选择来说是最明显的单元。使用字位簇，确保在复制和粘贴文本时不会突然丢掉一些符号，同时左右方向键也总是以一个可见字符的距离移动，等等。

另一个用到字位簇的地方是，执行字符串长度限制——比如在数据库域中。其实，底层的限制可能是类似 UTF-8 中的字节长度之类的东西，你不能简单的通过截断字节的方式来限制长度。至少，你得“舍去”最近的编码点；但更好的是，舍去最近的字位簇。除此以外，你可以通过舍弃它的一个注音符号破坏一个字符，中断一个 jamo 序列或 ZWJ 序列。


更多...

从程序员的角度来看，关于 Unicode 还有很多东西可以讲！我还没有深入一些有趣的主题，比如映射、排序、兼容性分解和容易混淆的词，Unicode 正则表达式，和双向文本。还有个我没谈到的是实现主题——如何有效存储和查找分布稀疏的编码点数据，或者如何优化 UTF-8 解码、字符串比较和 NFC 标准化。也许我会在未来的文章中讲到这些。

Unicode 是个令人着迷的复杂系统。在字节和编码点之前有多对一的映射，除此之外编码点和“字符”之间也有（某些情况下多对多）多对一的映射关系。在每个角落都有古怪的特例。没人声称表示全部书写系统很容易，但很明显我们不会回到使用不兼容编码来拼凑的艰难岁月了。

小礼物走一走，来简书关注我

赞赏支持

 CodingChat (/nb/29106875)

[举报文章](#) © 著作权归作者所有



faremax (/u/88ca7ceed50e)

写了 137566 字，被 17 人关注，获得了 30 个喜欢

(/u/88ca7ceed50e)

+ 关注

小编初入职场，由于精力有限，博客产量有点低，请同学们见谅

喜欢



更多分享



(/apps/redirect?utm_source=note-bottom-click)

(/apps/redirect?utm_source=side-banner-click) ×



登录 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-comment-form)

评论

智慧如你，不想发表一点想法 (/sign_in?utm_source=desktop&utm_medium=not-signed-in-nocomments-text)
咩~

推荐阅读

更多精彩内容 > (/)

写给2025年沈彦的一封信 (/p/779c2151d2a6?utm_campaign=maleskine&utm_content=note&utm_source=recommendation)
2025年的沈彦：你好！当你看到这封信的时候，一定会很惊喜吧？我是2018年的沈彦，我是过去的你，你是未来的我，你和我同一个人，又并非同一个人

颖睿飞沈彦 (/u/6ffe1f5f1f8e?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

蟋蟀追逐的梦想 (/p/182ce9ec92a8?utm_campaign=maleskine&utm_content=note&utm_source=recommendation)
潇潇秋风，吹动白云，送来了阵阵的寒意，使人不禁想起秋月下的独酌，往往物是人非，没有人生的凄凉，却有岁月的沧桑，任谁也挡不住。蟋蟀在堂，岁聿其逝。这是两千多年前《诗经》里的诗，可见那时

小喵静莉 (/u/3bef051f55a1?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

你不努力，没人能给你想要的生活 (/p/7d72ee39c02d?utm_campaign=maleskine&utm_content=note&utm_source=recommendation)
文|十三夜 1 加班结束，一个人撑着雨伞选择步行回家，已是晚上10点的时间，湿漉漉的街道边，一个八旬乞讨的老奶奶跪坐在地上，俯着头，一动一动的，她前方生了锈的小铁盆里，零零散散的躺着几块零钱。

十三夜 (/u/bfe4c3547845?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

我们来简书，究竟应该怎样去写作 | 我是道长是名思维 (/p/0226b8b45eb1?utm_campaign=maleskine&utm_content=note&utm_source=recommendation)
01 不知道，你是因为什么来简书？每个人都有自己的答案，每个人在别人的眼里也有不同答案。但我认为，我们一定有一个共同的目的——或多或少都是

道长是名思维贩子 (/u/92eb338437ee?utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

一张白纸 (/p/cf61df5d4a00?utm_campaign=maleskine&utm_content=note&utm_source=recommendation)


文署雨林季风 曾经的一张白纸 中国首创 纯洁表白 文明复写 而后的一张白纸 狼藉一片不平等 黄河怒吼 罄竹难书 如今的一张白纸 和平与你共享 长江为你开卷 崛起中国

雨林季风 (/u/1417ef767446?

utm_campaign=maleskine&utm_content=user&utm_medium=pc_all_hots&utm_source=recommendation)

字符、编码和Java中的编码 (/p/1b00ca07b003?utm_campaign=malesk...


字符是用户可以读写的最小单位。计算机所能支持的字符组成的集合，就叫做字符集。字符集通常以二维表的形式存在。二维表的内容和大小是由使用者的语言而定，是英语、是汉语、还是阿拉伯语。人类阅读

 刘惜有 (/u/4671bca15f69?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

[转] UTF-8编码的详细讲解 (/p/a592c5057f43?utm_campaign=maleski...

UTF-8 编码提供了一种简便而向后兼容的方法,使得那种完全围绕 ASCII 设计的操作系统,比如 Unix,也可以使用 Unicode. UTF-8 就是 Unix, Linux 已经类似的系统使用 Unicode 的方式. 现在是你了解它的时候了.

 谢大见 (/u/17065b4870ba?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)


(/p/64ec0f6b6245?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

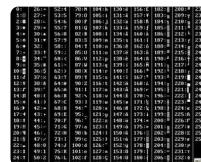
从Emoji的限制到Unicode编码 (/p/64ec0f6b6245?utm_campaign=mal...

某一天, leader找到我说, felix啊, 这里有个小需求, 给我们的实名认证中的地址加入字数限制, 一天时间绰绰有余了吧。我一听, 小事啊, 赶紧拍拍胸脯告诉leader, 一天都不用, 以我的效率1个小时就够了。领

 felix9 (/u/d11c662ab925?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)


(/p/375bb0bebe0d?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

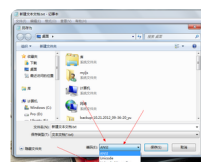
中文编码杂谈 (/p/375bb0bebe0d?utm_campaign=maleskine&utm_con...

编码问题的例子 在windows自带的notepad (记事本) 程序中输入“联通”两个字, 保存后再次打开, 会发现“联通”不见了, 代之以“◆◆”的乱码。这是windows平台上典型的中文编码问题。即文件保存的时候是

 天天向上1234567 (/u/22dad893e5a4?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)


(/p/f974cf4200a2?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)

【语言编码】ANSI、ASCII、Unicode、UTF-8 (/p/f974cf4200a2?utm...

一. 前情 在之前在notepad++里用markdown插入并预览本地图片时遇到了这样一个问题: 在记事本或notepad++中都能正常显示的几个汉字, 在markdown预览界面却成了乱码。 几经周折后, 发现记事本

 胡同口的蛙 (/u/e5a78443a033?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/0297814aae5b?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
焕然一新 (/p/0297814aae5b?utm_campaign=maleskine&utm_content=...

今晚把旅行箱仔细擦洗了一下，把买回时商家就告诉使用前最好把薄膜撕掉的事情做了一下，果然焕然一新，干净漂亮起来了。什么事情只要花上时间功夫，没有做不好的，一勤天下无难事，确实不假。默默

 劲汶 (/u/eb51179069b6?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

2018.3.29 四 多云 (/p/318159d14006?utm_campaign=maleskine&utm...


今天我看见马浩轩和罗一摔跤，让我想起了一部电影《摔跤吧，爸爸》。我对他们说：“你们是装模作样吗？”他们没有回答我，我就走开了。

 格格之歌 (/u/dfb84720e001?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

《盒饭财经今日说》 (/p/77f16375f649?utm_campaign=maleskine&ut...


财经史上的2月17日：1996年2月17日，卡斯帕罗夫与超级电脑深蓝的的国际象棋比赛结束，卡斯帕罗夫以4:2的战绩获胜。第二年，卡斯帕罗夫与经过改进的“深蓝”对垒，结果以2.5:3.5的比分败北。不过20年

 温婉姐姐 (/u/5ce44976c7c0?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

晚安 (/p/86a270bb1411?utm_campaign=maleskine&utm_content=not...

睡前按下遥控器的开关，老空调制造的最后一丝冷气缓慢的沉落下来，像一声悠长的叹息。

 禾阿乃 (/u/QBMXsY?


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)

(/p/a5aa4f0263b7?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommendation)
家 (/p/a5aa4f0263b7?utm_campaign=maleskine&utm_content=note&...

这个字会让你联想起哪些词 温暖、依靠、快乐、踏实、私密、安全、幸福、温馨、自在。我渴望这样的家

 熊小雯 (/u/d387619fbf96?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommendation)