

字符编码 UTF-8 学习笔记

原创

kfepiza

已于 2022-06-05 01:38:03 修改

146

收藏

版权

分类专栏：

编码解码进制转换规约协议

JAVA

文章标签：

学习

java



编码解码进制转换规... 同时被 2 个专栏收录 ▾

0 订阅 4 篇文章

订阅专栏

UTF-8

- UTF-8 是 Unicode标准中的一种 , 还有 UTF-16 , UTF-32
- UTF-8 的长度可变, 例如字母站一字节,汉字占三字节
- UTF-8 分为带BOM开头的 , 和不带BOM开头的, 两种
BOM占用开头三字节, 分别是 : 0xef , 0xbb , 0xbf
- UTF-8 不需要BOM就能识别, 它有明显的特点

UTF-8 编码的特点, 可以用来识别文件格式是否为 UTF-8

百度百科说UTF-8最大支持4字节,

1. 一个US-ASCII字符只需1字节编码 (Unicode范围由U+0000~U+007F) 。
2. 带有变音符号的拉丁文、希腊文、西里尔字母、亚美尼亚语、希伯来文、阿拉伯文、叙利亚文等字母则需要2字节编码 (Unicode范围由U+0080~U+07FF) 。
3. 其他语言的字符 (包括 中 日韩文字、东南亚文字、中东文字等) 包含了大部分常用字, 使用3字节编码。
4. 其他极少使用的语言字符使用4字节编码。

但看其特点, 支持6字节也是可以的

1. 一字节 0xxxxxxx
2. 二字节 110xxxxx 10xxxxxx
3. 三字节 1110xxxx 10xxxxxx 10xxxxxx
4. 四字节 11110xxx 10xxxxxx 10xxxxxx 10xxxxxx
5. 五字节 111110xx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx
6. 六字节 1111110x 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx 10xxxxxx

一字节时和ascii , iso8859-1 , gbk 是一样的, 不能用来判断

二字节开始的规律

一个字节二进制有几个1开头, 后面就有几减一个10开头的字节

五字节和六字节不是Unicode编码范围, 2003年RFC3629规定Utf-8只能使用原Unicode规定的区域(小于等于四字节)

JAVA 的 UTF-8 相关

Java 的 sun.nio.cs.UTF_8 类中有一段注释 , 也可以参考

```
1  /* Legal UTF-8 Byte Sequences
2  *
3  * #      Code Points      Bits      Bit/Byte pattern
4  * 1              7          0xxxxxxx
5  *      U+0000..U+007F          00..7F
6  *
7  * 2              11         110xxxxx  10xxxxxx
8  *      U+0080..U+07FF          C2..DF  80..BF
9  *
10 * 3              16         1110xxxx  10xxxxxx  10xxxxxx
11 *      U+0800..U+0FFF          E0          A0..BF  80..BF
12 *      U+1000..U+FFFF          E1..EF  80..BF  80..BF
13 *
14 * 4              21         11110xxx  10xxxxxx  10xxxxxx  10xxxxxx
15 *      U+10000..U+3FFFF          F0          90..BF  80..BF  80..BF
16 *      U+40000..U+FFFFF          F1..F3  80..BF  80..BF  80..BF
17 *      U+100000..U10FFFF          F4          80..8F  80..BF  80..BF
18 *
19 */
20
21 public final class UTF_8 extends Unicode {
22
```

🔗 文章知识点与官方知识档案匹配, 可进一步学习相关知识

中文字符 UTF-8 编码查询表	10-06
为大家提供Python的UTF-8编码查询表，大家可以对照左列的编码查询右列的汉字。 例：\u4e00对应汉字“一”	
JAVA 的 UTF-8 相关	kfepiza的博客 451
记录一些Java 与 utf-8 相关的知识	
编码知识学习笔记之一_rfc3629_littletigerat的博客	7-29
UTF:传输这些文字 九.UTF有哪些常见规范(也就是编码格式) 1.UTF-7 2.UTF-8 3.UTF-16 十.UTF的描述UTF-8和UTF-...	
UTF-8 编码格式(python)_python utf-8_小胖_@的博客	8-12
utf-8简介 UTF-8(8-bit Unicode Transformation Format)是一种针对Unicode的可变长度字符编码,由Ken Thompson于1...	
什么是UTF-8编码	格物致知的专栏 [音视频编解码 网络协议 计算机视... 2万+
UTF-8（8-bit Unicode Transformation Format）是一种针对Unicode的可变长度字符编码，也是一种前缀码。它可以用...	
java.lang.StackOverflowError at sun.nio.cs.UTF_8\$Encoder.encodeLoop(UTF_... iteye_2413的博客	623
INTERNAL_SERVER_ERROR Caused by: java.lang.StackOverflowError at sun.nio.cs.UTF_8\$Encoder.encodeLoop...	
UTF8常识_孤独斗士的博客	8-7
RFC 3629UTF-8 November 200314. Informative References1215. URI's	
奇文共欣赏~UTF-8设计与源码实现_musl支持unicode吗_桔子code的博客-CSDN...	7-28
RFC3629规定的Unicode字符集和UTF-8编码的对应关系是下图这样的: 1、设计原则 在Unicode官网上V1.1.0的附录部...	
UTF-8的编码规则	xuechanba的博客 5639
UTF-8的编码规则: 1、对于单字节的字符，字节的第一位设为0，后面七位为这个字符的Unicode码。 因此对于英文字...	
字符编码的概念（UTF-8、UTF-16、UTF-32都是什么鬼）	顾小暖的博客 6万+
字符集为每个字符分配了一个唯一的编号，通过这个编号就能找到对应的字符。在编程过程中我们经常会使用字符， ...	
网页设计中 utf-8和gb2312编码_weixin_34290352的博客	7-24
1、utf-8 UTF-8(8-bitUnicodeTransformation Format)是一种针对Unicode的可变长度字符编码,又称万国码。由Ken Tho...	
字符编码--UTF-8(1992年创建)_weixin_34337265的博客	8-7
UTF-8是UNICODE的一种变长度的编码表达方式《一般UNICODE为双位元组(指UCS2)》,它由肯·汤普逊(Ken Thomp...	
js将字符转换为UTF-8字符的工具	10-29
在下面的文本框中输入中文文字，按“转化”，即可将其转化为UTF-8字符。	
UTF-8字符集汉字对照表.txt	11-09
此文本文档是UTF-8字符集中汉字编码对照表，可以用于查看某个汉字在UTF-8编码集中的位置。此编码集对照表非官...	
utf8编解码详解_utf8编码解码_twwk120120的博客	8-9
RFC3629定义了utf8的编码定义:1-4字符编码 RFC2279中在unicode2的时候做了变更:将字符范围限制在0000-10FFFF...	
ISO8859-1、utf-8、gb2312_iso88591是什么编码_Dawn_Bells的博客-CSDN博...	8-11
ISO8859-1,通常叫做Latin-1。Latin-1包括了书写所有西方欧洲语言不可缺少的附加字符。gb2312是标准中文字符集。...	
Java 所有字符串转UTF-8 万能工具类-GetEncode.java	04-04
不需要关心接受的字符串编码是UTF_8还是GBK，还是ios-8859-1，自动转换为utf-8编码格式，无需判断字符串原有...	
字符编码笔记 ASCII，Unicode和UTF-8	10-27
下面就是我的笔记，主要用来整理自己的思路。但是，我尽量试图写得通俗易懂，希望能对其他朋友有用。毕竟，字...	
RFC3629_lyclowlevel的博客	7-25
A direct consequence is that a plain ASCII string is also a valid UTF-8 string. Yergeau Standards Track [Page 2] RFC...	
使用UTF8编码的意义	qq_52530620的博客 739
先来看一个例子： 以"I am Chinese"为例 用ANSI储存：12 Bytes 用Unicode/UCS2储存：24 Bytes + 2 Bytes(header) ...	
Unicode编码详解(三)：UTF-8编码	黄邦勇帅的博客 5223
Unicode编码详解(三)：UTF-8编码 若觉得本文写得还可以，请多多关注本人所作书籍《C++语法详解》电子工业出版...	
python文件开头声明UTF-8编码的几种常用形式	weixin_42468029的博客 5万+
Python默认ASCII编码，如包含中文，为防止乱码，往往需要在编码开头重新声明编码类型 常用的形式有以下几种， ...	
字符编码（ASCII、GBK、UTF-8、ANSI）详解	devilzcl的博客 4089
目录一、ASCII 码二、GB2312、GBK、GB18030、Big5三、Unicode四、ANSI 一、ASCII 码 ASCII (American Stand...	
UTF-8 与 UTF-16编码详解	小鲁蛋的博客 7159
UTF-16是Unicode字符编码五层次模型的第三层：字符编码表（Character Encoding Form，也称为 "storage format"...	
Python的编码注释# -*- coding:utf-8 -*- 热门推荐	arbel的专栏 29万+
如果要在python2的py文件里面写中文，则必须要添加一行声明文件编码的注释，否则python2会默认使用ASCII编码。...	
UTF-8编码规则（转）	qingfeng_博客 4733
UTF-8是Unicode的一种实现方式，也就是它的字节结构有特殊要求，所以我们说一个汉字的范围是0X4E00到0x9FA5...	
Python中关于coding=utf-8以及中文字符前加u的解释	敲代码的quant的博客 5万+
写了很久的Python了，每次写之前都要在开头加上coding=utf-8，只知道是设置编码格式，但并没有太在意，今天在写...	
字符转为utf-8 c++ 最新发布	05-01
UTF-8是一种通用的字符编码方式，它可以将不同编码方式的字符进行转换和存储，以确保它们在不同的平台上都可以...	

“相关推荐”对你有帮助么？

 非常没帮助

 没帮助

 一般

 有帮助

 非常有帮助



kfepiza

码龄6年

 暂无认证

362

3万+

6886

29万+



原创

周排名

总排名

访问

等级

3802

59

120

41

569

积分

粉丝

获赞

评论

收藏




















私信

关注

搜博主文章



热门文章

JAVA8之 日期时间时区之
Zoneld[ZoneOffset, ZoneRegion] 笔记 

17147

Java byte转换为int 

9836

Java8(291)之后，禁用了TLS1.1，使JDBC
无法用SSL连接SqlServer2008怎么办,以下是解决办法 

8695

vi vim 快速跳到文件末尾 GA 在最后一行下
方新增一行 (光标换行,文字不换行) GO 

7090

Linux之 如何查看文件是`硬链接`还是`软链
接` 

7063

分类专栏

 English named defined

1篇

 依赖管理 Maven Gradle 等

3篇

 文本代码编辑器 IDE

1篇

 文本,正则RegExp,text,stri...

12篇

 安装软件 开发环境搭建

14篇

 说明书 手册 帮助文档 指南...

3篇



最新评论

Ubuntu22.04Desktop桌面版设置静态Ip

wangtong2012:   ，看了楼主的文
章，终于搞清楚了Ubuntu现在的网络配 

Linux之 如何查看文件是`硬链接`还是`软...

奔跑的小码农: 创建硬连接 ln -s 源文件 硬
链接名 或 link -s 源文件 硬链接名 写错了...

Java实例化 new和newInstance反射 性能...

kfepiza: 所言极是,有所启发,非常感谢,又做
了放外面的测试,确实快很多,不过我本意...

Java实例化 new和newInstance反射 性能...

weixin_40158409: 看代码里面把反射耗时
也算进去了，getDeclaredConstructor这 

Received message too long Ensure the r...

无数_mirage: 服了，原来是有echo

您愿意向朋友推荐“博客详情页”吗？











强烈不推荐

不推荐

一般般

推荐

强烈推荐

最新文章

CSS font-family 等宽字体

Echarts dataZoom x轴横坐标缩放

git pull origin master 时, 遇到 fatal: refusing
to merge unrelated histories 230626

2023

08月

1篇

07月

1篇

06月

50篇

05月

12篇

04月
4篇

03月
8篇

02月
20篇

2022年 191篇

2021年 78篇

目录

UTF-8

UTF-8 编码的特点, 可以用来识别文...

JAVA 的 UTF-8 相关