文件的基础

黄天羽

北京理工大学

python 语言程序设计

文件

- 存储在外部介质上的数据或信息的集合 。
  - 程序中的源程序
  - 数据中保存着数据
  - 图像中的像素数据
  - …
- 有序的数据序列。

编码

- 信息从一种形式转换为另一种形式的过程
- ASCII码
- Unicode
- UTF-8
- …

常用的编码

python 语言程序设计

常用的编码

- Unicode
  - 跨语言、跨平台进行文本转换和处理
  - 对每种语言中字符设定统一且唯一的二进制编码
  - 每个字符两个字节长
  - 65536 个字符的编码空间
  - “严” ：Unicode的十六进制数为4E25

常用的编码

- UTF-8编码
  - 可变长度的Unicode的实现方式
  - “严”：十六进制数为E4B8A5

Unicode与UTF8编码字节范围对应关系

Unicode符号范围 (十六进制)	UTF-8编码方式（二进制）
0000 0000-0000 007F	0xxxxxxx
0000 0080-0000 07FF	110xxxxx 10xxxxxx
0000 0800-0000 FFFF	1110xxxx 10xxxxxx 10xxxxxx
0001 0000-0010 FFFF	11110xxx 10xxxxxx 10xxxxxx 10xxxxxx

- GBK编码
  - 双字节编码
  - 举例

文件数据

- 文本文件
  - 以ASCII码方式存储的文件…
- 二进制文件

python 语言程序设计

多行文本

- \n表示换行
- 举例：

```
Hello
World

Goodbye 32

Hello \nWorld \n \nGoodbye 32 \n
```

- 存储在文件中，得到字符序列：

- 二进制文件ASCII码
  - 照片、音乐、视频、计算机程序等
  - 优点：
    - 更加节省空间
    - 采用二进制无格式存储
    - 表示更为精确

- 注意：
  - 文本文件是基于字符定长的ASCII；
  - 二进制文件编码是变长的，灵活利用率要高；
  - 不同的二进制文件解码方式是不同的。