cjkspace 使用说明

gosman

http://gosman.blogbus.com

2007 年 8 月 26 日

目 录

1 介绍

cjkspace 在中英文之间插入波纹号(~)，也可以是其他符号如空白或问号等。中英文之间空白的讨论可以参考李果正先生的博文：中英文字间空白。虽然这里说的是中英文，但程序中汉字的定义使用的是Unicode的 CJK 字符集，因此应该也支持日文和韩文。程序对中英文的定义都可以重新定义或扩展，如对德文或法文的支持等。cjkspace 使用的是 Python 的内置编码器，支持大部分的文字编码，如 gbk、gb18030、big5 等，不指明的话，使用 utf8，可以使用“cjkspace -l”查询已知编码，更多信息参考 Python 的帮助文档：标准编码。

程序在插入间隔符前，将删除中英文之间的所有空格。在多文件输入时，输出文件名由程序自动在原文件名后加.out 后缀，若此时使用‘-o’指定输出文件名，程序将予以警告，不会出错。

2 安装

cjkspace 是个 Python 脚本，因此下载后使其可执行(chmod u+x ./cjkspace)就可运行。程序在 Debian lenny/sid Python 2.4 下开发测试，推荐在 Python 2.4 以上的版本运行。

下载及更新请关注我的博客：http://gosman.blogbus.com。

3 选项说明

程序各选项及说明如下:

-d,–delimeter token 指定间隔符，不指定时默认为‘~’。

-c,–cover 不指定输出文件，直接覆盖输入文件，默认备份输入文件。

-n,–nobackup 在覆盖源输入文件时，不备份输入文件。

-o,–output filename 指定输出文件名。

-v,–verbatim verbatim 添加新的 verbatim 环境，请检查已有环境，最好不要重复。

-x,–noverbatim verbatim 取消指定的 verbatim 环境，若无定义将警告。

-a,–all 忽略注释、verbatim 环境，在中英文之间插入间隔符，此时通常是空格，仍然会删除多余空格。

-e,–encoding encoding 指定文件编码为 encoding，默认为 utf8，程序不具有自动检测编码的功能，请一定要指定正确的编码，否则程序出错或输出错误的结果。

-l,–list 列出已知编码。

-h,–help 帮助信息。

-V,–version 程序版本信息。

4 举例

4.1 用法

1. 使用 utf8 编码，输出到标准输出，即屏幕。

```
cjkspace utf.tex
```

2. 使用 utf8 编码，输出覆盖输入文件，在例子中即 utf.tex，自动备份输入文件为 utf.tex.bak。使用‘-n’或‘–nobackup’选项可以取消自动备份。

```
cjkspace -c utf.tex
```

3. 指定输出文件。

```
cjkspace -o utf2.tex utf.tex
```

或

```
cjkspace utf.tex > utf2.tex
```

4. 指定间隔符，默认使用的是‘~’，以空格为例，也可以其他任何字符。

```
cjkspace -d ' ' utf.tex
```

5. 添加 verbatim 环境，默认有 verbatim、Verbatim、comment；请不要重复定义，无法保证程序不出错，虽然现在好像没问题。添加 code、example 环境。

```
cjkspace -v 'code example' utf.tex
```

6. 去除 verbatim 环境，若去除未包含的 verbatim 环境，程序将予以警告，但不会出错。

```
cjkspace -x 'Verbatim comment' utf.tex
```

7. 使用其他编码文件，以 gb18030 为例。

```
cjkspace -e gb18030 gbk.tex
```

8. 多输入文件。

```
cjkspace utf.tex utf2.tex
```

或

```
cjkspace utf*
```

4.2 输出

原文件为：

```
通过Python的 re      模块\cmd命令\cmd    命令english usa python 正则表式的\%汉english个入门
\begin{verbatim}
这是中文Chinese,, 中文English
\end{verbatim}
这是中文~Chinese,, 中%文English\%english和

\begin{enumerate}
\item 这是中文~Chinese,, 中文~English
\end{enumerate}
这是中文~Chinese,, 中%文English
```

经 cjkspace 处理后，输出为:

```
通过~Python~的~re~模块\cmd命令\cmd   命令~english usa python~正则表式的\%汉~english~个入门
\begin{verbatim}
这是中文Chinese,，中文English
\end{verbatim}
这是中文~Chinese,，中%文English\%english和

\begin{enumerate}
\item 这是中文~Chinese,，中文~English
\end{enumerate}
这是中文~Chinese,，中%文English
```

4.3 技巧

1. 在注释、verbatim 环境里间隔符使用空格，正文使用波纹号。

   ```
   cjkspace -a -d ’ ’ test.txt > test2.txt
   cjkspace test2.txt > test3.txt
   ```

2. 由于 verbatim 环境变数太多，请合理使用‘-v’和‘-x’选项，以达到预期效果。最好不要使用覆盖原文件的方式，虽然这样比较省事，最好每次都指定输出文件。

3. 本文档使用以下方式处理：

   ```
   cjksapce -v code cjksapce.tex > cjkspace2.tex
   ```

5 定制

要修改 CJK 字符集定义、英文字符定义及 verbatim 环境，可以直接更改源程序 cjkspace 的“Edit Are”区域。

```
#================================== Edit Area ==================================
# If you want to change or add the Range or Verbatim keyword,you can change in here.
# More information see the help file cjkspace.pdf
#
# Reference http://blog.oasisfeng.com/2006/10/19/full-cjk-unicode-range/
# http://www.unicode.org
CJK_Range = [[u'\u3400',u'\u4db5'],
[u'\u4e00',u'\u9fa5'],
[u'\u9fa6',u'\u9fbb'],
[u'\uf900',u'\ufa2d'],
[u'\ufa30',u'\ufa6a'],
[u'\ufa70',u'\ufad9'],
[u'\u20000',u'\u2a6d6'],
[u'\u2f800',u'\u2fa1d']]
Letter_Range = [[u'A',u'Z'],
[u'a',u'z'],
[u'0',u'9']]
Verbatim_Str = ['verbatim','Verbatim','comment']
#==============================================================================
```

5.1 CJK 字符集

CJK 字符集的信息可以参考完整的 CJK Unicode 范围(5.0 版)和http://www.unicode.org。CJK 比划的 Unicode 范围为 31C0-31EF，要添加对 CJK 比划的支持，只需添加[u'\u31c0',u'\u31ef']到 CJK_Range 类表。

```
CJK_Range = [[u'\u3400',u'\u4db5'],
[u'\u4e00',u'\u9fa5'],
[u'\u9fa6',u'\u9fbb'],
[u'\uf900',u'\ufa2d'],
[u'\ufa30',u'\ufa6a'],
[u'\ufa70',u'\ufad9'],
[u'\u20000',u'\u2a6d6'],
[u'\u31c0',u'\u31ef'],
[u'\u2f800',u'\u2fa1d']]
```

5.2 英文字符

英文字符的定义和 CJK 字符是一样的，比如添加范围为 03B1 到 03C9 的希腊字符，只需修改 Letter_Range 列表。

```
Letter_Range = [[u'A',u'Z'],
[u'a',u'z'],
[u'\u03b1',u'\u03c9'],
[u'0',u'9']]
```

5.3 verbatim 环境

要添加 verbatim 环境，修改 Verbatim_Str 列表即可，比如添加 code、example。

```
Verbatim_Str = ['verbatim','Verbatim','comment','code','example']
```