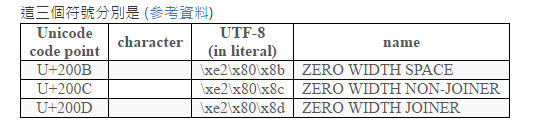
1. **过滤<U+200B><U+200C><U+200D>等不可见字符。**

文本中出现上述不可见字符：



这些字符其实就是排版过程中产生的，而排版使用的规范是Unicode编码标准。



通过如下方法过滤：

sed 's/\xe2\x80\x8b//g' inputfile

sed 's/\xe2\x80\x8c//g' inputfile

sed 's/\xe2\x80\x8d//g' inputfile

<http://www.ruanyifeng.com/blog/2007/10/ascii_unicode_and_utf-8.html>

(**字符编码笔记：ASCII，Unicode 和 UTF-8**)